

Capacity Planning and Design by Derivation

Load-carrying capability, sizing and redundancy

Draft · 343 pages

Ricardo Marquez, Ph.D.

ricardo-marquez.com

Capacity Planning and Design by Derivation

Volume I

From the planner's questions to probabilistic inference and decisions

Ricardo Marquez

October 31, 2026 (draft for review)

How to read this book

This book is written for the capacity planner, the mechanical design engineer and the facilities planner of a data center. Its questions are the ones a design review asks: whether the plant will carry the load, how much to build, where the bottleneck is, what the uncertainty in the design inputs costs, how the design holds when a unit fails, and what the design's risk is worth. A design sheet answers them by summing nameplate capacities, setting a design-day load against them and applying a margin, and returns one number with no band. It carries handbook values as if they were exact, reads one constraint at a time, and keeps availability and capacity in separate tools. Each question in this book is instead a probability statement over the unknowns of one model of the plant, and the book derives and computes each one.

The plant grows with the book. Part I works on one equipment hall cooled by a single unit, a state of two and then three entries, so that every object of the method can be seen whole. Part II takes one data hall with three cooling units and five racks, calibrated on a night of its own telemetry, and asks what load it can carry. From Part III the plant is the whole site: two halls on one chilled-water plant with its chillers, pump, shared supply pipe and power feed, an economizer that is drawn but not yet built, a year of monthly operating records, and the design alternatives the planner must choose between. On each plant one fitted model serves every question: a what-if is that fit read at a load, an ambient or a unit state not yet seen, and a build decision is an enumeration of such what-ifs with a cost attached. The chapters say which fit they read.

Every chapter that answers a question follows the same progression, and the progression is the method.

Step 0: the plant graph. The plant as it is, before any question is asked: its nodes, paths and sources; which quantities are pinned and which are free; the physical relations that must hold among them. The free quantities form the state vector $\mathbf{z}$.

Step 1: the question. The planner's question in the planner's words, then as a probability or an expectation of a function of the state, with the unknown on the left of the conditioning bar and everything treated as known on the right.

Step 2: the graph extended for the question. Scenario nodes for a load, an ambient, a unit out of service or a design alternative not yet built, and the record attached as rows.

Step 3: the posterior of the state. Derived and computed once, before any question is answered: the MAP state, the information, the covariance, and the exact posterior where one exists.

Step 4: the answers. Each question read from that posterior down to its number: the load at which a limit binds, a joint tail over the constrained entries, an availability from an enumeration of states, or an expected cost.

Step 5: evaluation and checks. The gaps between the reported values and the exact values or draws collected, the approximations tested and, because the record is constructed from a stated truth, the answers tested against that truth.

Step 6: the report. The headroom, the probability, the date or the design the planner signs, with its band and the exact value beside it where one exists.

Step 7: the next measurement. If the band is too wide to decide the build, the tests and surveys on the existing plant ranked by the width each would take off the quantity that decides it.

Five kinds of question cover what the book asks, distinguished by the mathematical object that answers them rather than by the word the design review uses: *estimation*, *prediction under a threshold*, *hypothesis comparison*, *diagnosis* and *design*. Beneath them sit five computations: the *posterior* of the state, the *trend* of an entry across windows, the *evidence* of a drawing of the plant, the *preposterior* after a measurement not yet taken, and the *expected consequence* of a decision. Capacity planning is mostly the posterior read at a scenario and the expected consequence of a design, with the preposterior deciding what to measure before committing. The notation chapter that follows fixes the symbols, and the appendix collects the planner's questions in an atlas that maps each to its mathematical form, its computation and the section that answers it.

The framework the book teaches is physics-constrained probabilistic modeling: a plant written as a graph, its unknowns as a state vector, every reading as a row, and every question as a probability over that state. TERRA is the author's own implementation of this framework; the book names the framework and the computations, not the software, and every number in it is reproducible from the derivations as written. The method and its notation are those of the companion books *Reliability Engineering by Derivation* and *Operations and Control by Derivation*, which ask the reliability engineer's and the operator's questions of the same plant.

Each chapter closes with a statement of what to retain, a table of its terms, and exercises at three levels, formulation, derivation and application, each followed by its answer.

Contents

How to read this book **ii**

Notation and vocabulary **vii**

I Thinking in Probabilities **1**

1 Planned plants as graphs **2**
1.1 One system and several questions . . . 2
1.2 The posterior of the state (step 3) . . . 15
1.3 The answers (step 4) . . . 20
1.4 The checks (step 5) . . . 25
1.5 The report (step 6) . . . 28
1.6 The next measurement (step 7) . . . 29
1.7 What to retain . . . 33
1.8 Exercises . . . 34

2 The language of capacity questions **37**
2.1 Planning questions and their mathematical form . . . 37
2.2 The grammar of probability statements . . . 40
2.3 The general forms and the five computations . . . 42
2.4 Estimation, prediction and joint events . . . 54
2.5 Model comparison, diagnosis and decisions . . . 66
2.6 Measurements, classification and workflow . . . 70
2.7 What to retain . . . 76
2.8 Exercises . . . 79

3 Uncertainty and conditional probability **82**
3.1 Uncertainty and evidence in design questions . . . 82
3.2 Conditional probability and Bayes' theorem . . . 87
3.3 Bayesian diagnosis of a capacity shortfall . . . 90
3.4 Marginalization and dependence . . . 93
3.5 Conditioning in a physical model . . . 97
3.6 Predictive variance and the posterior covariance . . . 106
3.7 What to retain . . . 112
3.8 Exercises . . . 115

4 The anatomy of probabilistic inference **117**
4.1 From a plant graph to a design probability . . . 117
4.2 The posterior of the state (step 3) . . . 125
4.3 The answers (step 4) . . . 138

4.4 The checks (step 5) . . . 145
4.5 The report (step 6) . . . 150
4.6 The next measurement (step 7) . . . 152
4.7 What to retain . . . 155
4.8 Exercises . . . 159

II Load-Carrying Capability 161

5 Thermal headroom 162
5.1 How much more load hall B can take . . . 162
5.2 The posterior of the state (step 3) . . . 169
5.3 The answers (step 4) . . . 172
5.4 The checks (step 5) . . . 179
5.5 The report (step 6) . . . 180
5.6 The next measurement (step 7) . . . 182
5.7 What to retain . . . 185
5.8 Exercises . . . 186

6 Load growth scenarios 188
6.1 When hall B leaves its envelope . . . 188
6.2 The posterior of the state (step 3) . . . 191
6.3 The answers (step 4) . . . 191
6.4 The checks (step 5) . . . 198
6.5 The report (step 6) . . . 199
6.6 The next measurement (step 7) . . . 200
6.7 What to retain . . . 202
6.8 Exercises . . . 203

7 Design-day assessment 205
7.1 Hall B on the design day . . . 205
7.2 The posterior of the state (step 3) . . . 209
7.3 The answers (step 4) . . . 210
7.4 The checks (step 5) . . . 216
7.5 The report (step 6) . . . 217
7.6 The next measurement (step 7) . . . 219
7.7 What to retain . . . 220
7.8 Exercises . . . 221

8 Load diversity and coincident peaks 223
8.1 The coincident peak of hall B at its design population . . . 223
8.2 The posterior of the two states (step 3) . . . 228
8.3 The answers (step 4) . . . 231
8.4 The checks (step 5) . . . 237
8.5 The report (step 6) . . . 238
8.6 The next measurement (step 7) . . . 239
8.7 What to retain . . . 242
8.8 Exercises . . . 243

III Sizing and Redundancy **244**

9 Redundancy level **245**
9.1 N+1 or N+2 for the site's chillers . . . 245
9.2 The posterior of the state (step 3) . . . 251
9.3 The answers (step 4) . . . 254
9.4 The checks (step 5) . . . 260
9.5 The report (step 6) . . . 262
9.6 The next measurement (step 7) . . . 263
9.7 What to retain . . . 265
9.8 Exercises . . . 266

10 Unit sizing **268**
10.1 Two quotes for one chiller plant . . . 268
10.2 The posterior of the state (step 3) . . . 274
10.3 The answers (step 4) . . . 275
10.4 The checks (step 5) . . . 282
10.5 The report (step 6) . . . 283
10.6 The next measurement (step 7) . . . 284
10.7 What to retain . . . 286
10.8 Exercises . . . 287

11 Expansion timing **289**
11.1 When the fourth chiller has to be ordered . . . 289
11.2 The posterior of the state (step 3) . . . 294
11.3 The answers (step 4) . . . 294
11.4 The checks (step 5) . . . 301
11.5 The report (step 6) . . . 302
11.6 The next measurement (step 7) . . . 304
11.7 What to retain . . . 305
11.8 Exercises . . . 306

12 Oversizing and stranded capacity **308**
12.1 The build-out drawing at the ten-year load . . . 308
12.2 The posterior of the state (step 3) . . . 313
12.3 The answers (step 4) . . . 315
12.4 The checks (step 5) . . . 321
12.5 The report (step 6) . . . 322
12.6 The next measurement (step 7) . . . 324
12.7 What to retain . . . 326
12.8 Exercises . . . 327

A Atlas of planning questions **329**

Notation and vocabulary

Every symbol below is used with the same meaning in every chapter. The unknown always stands on the left of a conditioning bar and the evidence on the right. Vectors and matrices are bold. Hall subscripts are $h \in \{A, B\}$, cooling-unit subscripts $u \in \{A1, A2, A3, B1, B2, B3\}$, rack subscripts $k \in \{A1, \ldots, A5, B1, \ldots, B5\}$ and chiller subscripts $i \in \{C1, C2, C3\}$; the hall of a unit or a rack is the letter of its name.

Symbol	Unit	Meaning
The plant graph and its state		
$\mathbf{z}$, z_i	varies	State vector of the plant graph: every free quantity at once; its i-th entry.
$\mathbf{z}_t$, t	varies, –	The state of window t and the window index: the five-minute windows of the night record in Part II, the monthly windows of the site's operating year from Part III.
$\mathbf{y}$, y	varies	The evidence: all readings and prior information treated as known; a single reading, subscripted by what it reads (y_{T_s}, $y_{P_{\mathrm{ch}}}$).
$r_k(\mathbf{z})$	varies	Row k of the model: a physical relation that must hold ($r_k = 0$), a reading compared with the state, or a prior.
w_k, σ_k	varies	Weight (precision) of row k and the standard deviation it corresponds to, $w_k = 1/\sigma_k^2$. A row may equivalently be written normalized, r_k/σ_k, with unit weight.
σ_p	varies	Tolerance of a physics row; an exact constraint is the limit $\sigma_p \to 0$.
The equipment hall of Part I		
T, T_c	°C	Temperature of the hall air; the supply air held by the cooling unit's discharge controller, pinned.
Q, G	kW, kW/K	Heat load of the hall (30 kW now, 60 kW the design load); effective conductance of its air path to the supply air.
g, G_{ref}	–, kW/K	Log-conductance $\ln(G/G_{\mathrm{ref}})$, the entry of the state that carries the conductance; its reference value.
y_T, σ_T	°C, K	The hall sensor's reading and the standard deviation of its error.
The halls of Part II and the site		
T_s	°C	Chilled-water supply temperature, 14°C at the set point.
Q_h, $Q_{o,h}$	kW	IT heat load of hall h; its other heat (fans, lighting), pinned.
G_u, $\bar{G}_u$	kW/K	Effective conductance of the coil of cooling unit u; its rated value, pinned.

Symbol	Unit	Meaning
g_u	–	Log-derate $\ln(G_u/\bar{G}_u)$ of unit u.
C_u, Q_u	kW/K, kW	Air capacity rate of unit u, pinned; the heat it rejects to the chilled-water loop.
$T_{r,h}$, $T_{c,h}$, $T_{d,u}$	°C	Return and cold-aisle air temperature of hall h; discharge air temperature of unit u.
T_k, β_k	°C, –	Inlet temperature of rack k; its recirculation fraction.
The chilled-water plant, from Part III		
Q_{IT}, Q_L	kW	Site IT load; heat to the chilled-water loop, Q_{IT} plus the halls' other heat.
T_{wb}	°C	Outdoor wet-bulb temperature, a known input of each window; 22°C at the design day.
N_i, κ_i	kW, –	Nameplate capacity of chiller i at the design day; its as-built log-ratio, the entry that carries the capacity $N_i e^{\kappa_i}$.
d	1/K	Capacity derate per kelvin of wet bulb above the rating condition.
G_e, g_e	kW/K, –	Conductance of the economizer path, drawn and not yet built; its log-ratio to the handbook value.
COP, c	–, –	Efficiency of the chiller plant; the log of its value at 18°C wet bulb.
V, ΔP	L/s, kPa	Loop flow; pressure drop across the shared supply pipe.
k_p, η	varies, –	Flow coefficient of the shared pipe in L/s per $\mathrm{kPa}^{1/2}$, $\Delta P = (V/k_p)^2$; pump efficiency.
P_{ch}, P_{pump}, P_{feed}	kW	Chiller power, pump power, power through the feed.
Scenarios, capacity and design		
x^*	varies	A starred quantity is read at a scenario node: the plant re-solved at a load, ambient, unit state or design alternative not yet seen, with $\mathbf{\Sigma}$ carried through. A starred limit ($T_{\max}$, f^*) is a requirement level; the context tells the two apart, as in the companion books.
$T_{\max}$	°C	A limit on a temperature: the hall's 27.5°C at the design load in Part I; a rack inlet's 27°C, the recommended envelope, from Part II.
$Q_e^{\dagger}$	kW	Binding load of element e: the site or hall IT load at which its limit binds.
H	kW	Headroom: the lowest binding load less the load today.
m, m_0	–	Capacity margin, capacity over the design load less one; the drawing's margin, 0.20.
A	–	Service availability, the probability of the states in which the load is carried; never equated with a capacity margin.
EUE	kWh/yr	Expected unserved energy: the IT load shed per year, in expectation.
λ, μ	1/yr, 1/h	Failure rate and repair rate of a unit.
γ_L	1/yr	Growth rate of the IT load, a stated input.
Probability		
$\Pr(A)$, $\Pr(A \mid B)$	–	Probability of event A; conditional on event B.

Symbol	Unit	Meaning
$p(\cdot)$	varies	A probability density; $p(\mathbf{z} \mid \mathbf{y})$ is the posterior of the state, $p(\mathbf{y} \mid \mathbf{z})$ the likelihood, $p(\mathbf{z})$ the prior.
$\mathbb{E}$, Var, Cov	varies	Expectation, variance, covariance.
$\mathcal{N}(\mu, \sigma^2)$	–	Normal distribution with mean μ and variance σ^2; Φ its standard cumulative distribution function, φ its density.
χ^2_ν	–	Chi-square distribution with ν degrees of freedom.
Inference objects		
$\mathcal{L}(\mathbf{z})$	–	The objective, minus the log posterior up to a constant: $\mathcal{L} = \frac{1}{2} r(\mathbf{z})^\mathsf{T} r(\mathbf{z})$ for normalized Gaussian rows.
$\hat{\mathbf{z}}$	varies	The maximum a posteriori (MAP) state, the minimizer of $\mathcal{L}$.
$\mathbf{J}$	varies	Jacobian of the normalized rows with respect to the state, evaluated at $\hat{\mathbf{z}}$.
$\mathbf{\Lambda}$	varies	Information matrix (Gauss–Newton Hessian of $\mathcal{L}$), $\mathbf{J}^\mathsf{T}\mathbf{J}$ for normalized rows.
$\mathbf{\Sigma}$, Σ_{ij}	varies	Posterior covariance of the state under the Laplace approximation, $\mathbf{\Lambda}^{-1}$; its entries. The posterior spread of entry i is written $\sqrt{\Sigma_{ii}}$; σ_k always names a row's standard deviation.
$f(\mathbf{z})$, $\mathbf{g}$	varies	A derived quantity of the state and its gradient $\partial f / \partial \mathbf{z}$ at $\hat{\mathbf{z}}$.
f_L, f_U	varies	The ends of the credible interval of a derived quantity.
$\mathbf{J}_a$, $\mathbf{\Lambda}_a$, $\mathbf{\Sigma}_a$	varies	The Jacobian rows of a candidate measurement a, its information $\mathbf{J}_a^\mathsf{T}\mathbf{J}_a$, and the preposterior covariance $(\mathbf{\Sigma}^{-1} + \mathbf{\Lambda}_a)^{-1}$.
M, B_{AB}	–	A drawing (model) of the plant; the Bayes factor $p(\mathbf{y} \mid M_A)/p(\mathbf{y} \mid M_B)$.
$\mathcal{R}(a)$, $L(a, \cdot)$	varies	Expected cost (risk) of a candidate design or action a; its loss.
U, U_a	–	Unavailability, $1 - A$; of drawing a (Part III).

Vocabulary

Plant graph The physical system as nodes (places where a conserved quantity balances), edges (paths along which it flows) and sources, with the relations that must hold among them. Built before any question is asked.

State vector The collection of free quantities the fit estimates at once.

Pinned, free A quantity treated as known (boundary condition) against one carried in the state.

Residual, row A relation written so that it equals zero when satisfied; physics rows constrain the state, reading rows compare it with the record, prior rows carry handbook or commissioning knowledge.

Calibration The fit of the existing plant's parameters to its operating record: the MAP state and its covariance. Every what-if reads it.

Scenario (what-if) The plant re-solved at a scenario node, with a load, an ambient, a unit state or a design alternative pinned to a value not yet seen, and the covariance carried through.

Design day The ambient condition the design is drawn for, taken from the site's climate data; here a wet bulb of 22°C.

Headroom, binding load The load the plant can still take before a limit binds; the load at which an element's limit binds.

Capacity margin Capacity at the design day over the design load, less one. A statement about capacity, never about availability.

Coincidence (diversity) factor The ratio of the coincident peak of a group of loads to the sum of their individual peaks.

N+1, N+2 redundancy Installed units beyond the N needed to carry the design load.

Concurrent maintainability Any one unit can be taken out for service without interrupting the load.

Binding constraint, debottlenecking The element whose limit binds first as the load grows; raising that element's rating.

Importance measure The sensitivity of a system measure to one element: here the gradient of headroom with respect to an element's rating.

Minimal cut set, common-mode failure A smallest set of elements whose joint loss fails the service; a failure that defeats redundant units through an element they share.

Cascading failure A failure whose shifted load drives a second element past its envelope.

Expected unserved energy, loss-of-load probability The IT energy shed per year in expectation; the probability that the load is not carried.

Stranded (oversized) capacity Capacity built and paid for that the load does not use.

Laplace approximation, delta method Replacing the posterior by the Gaussian at the MAP state with the inverse curvature as covariance; linearizing a derived quantity so that its variance is $\mathbf{g}^\top \boldsymbol{\Sigma} \mathbf{g}$.

Variance budget The share of a derived quantity's variance carried by each entry of the state.

Preposterior, value of information The posterior that would result from a measurement not yet taken; the expected improvement in the decision that measurement buys.

The five computations Posterior, trend, evidence, preposterior and expected consequence: the objects every planner question is answered from.

Part I

Thinking in Probabilities

Chapter 1

Planned Plants as Graphs

1.1 One system and several questions

A design review has one data center equipment hall on its agenda. The hall carries 30 kW of servers today, and its design load is 60 kW, twice that; at the design load the hall air must stay below 27.5°C. The record is thin: one hall sensor, with an error of about half a kelvin, read 23.2°C at commissioning, under today's load, and the design handbook gives the conductance of the hall's air path to the supply air, held at 18°C by the cooling unit's discharge controller, as 7.5 kW/K, right to within a factor of about 1.35. The design sheet carries two numbers for the hall at its design load. The first is the nameplate forward solve, which works from the handbook value. Heat leaves the hall in proportion to its excess over the supply air, so at 7.5 kW/K the hall sits at 22°C under 30 kW, comfortably below its present limit of 24°C, and at 26.0°C under 60 kW, well below 27.5°C: the sheet marks the design load "pass" with room to spare.

The second number works from the commissioning reading and tells another story. Taken as exact, 23.2°C at 30 kW means a conductance of 5.77 kW/K, and at that conductance the design load puts the hall at 28.4°C, over the limit: the sheet marks it "fail". The review must sign off one of the two, and the sheet cannot say how sure either is; each treats its own input as exact, and neither is weighed against the other. In the planner's words the question is

Will the hall carry its 60 kW *design load inside the* 27.5°C *limit, and how sure is that?*

Its answer is a probability, that the hall stays below 27.5°C at a load it has never carried, and that probability can only come from holding the handbook value, the reading and the energy balance that ties them together in one description of the hall, each with its own uncertainty. The question of this chapter is therefore

how a planned plant is represented so that the same representation serves every probabilistic question asked of it, and what must be fixed before any inference is performed.

Jargon: plant graph, node, path, source, pinned and free quantities, state vector, residual, row, observation, prior, posterior, state estimation, model discrepancy, scenario node, Jacobian, normal equations, information matrix, covariance, Laplace approximation, factor graph

A *plant graph* represents a physical system as *nodes*, places where a conserved quantity such as energy or mass must balance; *paths* (edges), along which the quantity flows; and *sources*, which supply it. A quantity is *pinned* when it is treated as known, as a boundary condition, and *free* when it is unknown and must be inferred. The free quantities, collected in one vector, form the *state vector* $\mathbf{z}$. A *residual* is a relation written so that it equals zero when it is satisfied; each residual is a *row* of the model. A physical row states a relation that must hold among entries of the state; an *observation* (a reading) is a measured value related to an entry of the state with an error, and its row compares the two. The *prior* expresses what is known about the state before the readings, for example a handbook value of a conductance; the *posterior* $p(\mathbf{z} \mid \mathbf{y})$ expresses what is known after the evidence $\mathbf{y}$; fitting the state to its rows, which yields the posterior's most probable state and its spread, is what the planner calls the *calibration* of the existing plant (state estimation). A *model discrepancy* is the difference between the physical system and the model of it that would remain if every entry of the state were known exactly. A *scenario node* is a node added to the graph for an operating condition that has not occurred, so that a question can be asked about it. The *Jacobian* $\mathbf{J}$ is the matrix of derivatives of every row with respect to every entry of the state, one row per row of the model and one column per entry. The *normal equations* $(\mathbf{J}^\mathsf{T}\mathbf{J})\,\Delta\mathbf{z} = -\mathbf{J}^\mathsf{T}\mathbf{r}$ give the step that best reduces all the rows at once when each is replaced by its tangent. The *information matrix* $\mathbf{J}^\mathsf{T}\mathbf{J}$ states how sharply the rows together fix the state, direction by direction, and its inverse is the posterior *covariance* $\boldsymbol{\Sigma}$, whose diagonal holds the variances of the entries and whose off-diagonal entries hold how they vary together. The *Laplace approximation* replaces the posterior by the normal distribution centered on its most probable state with that covariance. A *factor graph* draws the entries of the state and the rows that act on them, one node each.

Key idea

A graph describes the plant; a probabilistic question states what is to be learned about it. The plant graph, its pinned and free quantities and its physical rows are built once, and each question is a probability over the state of that one graph, with the unknown on the left of the conditioning bar and the evidence on the right. The posterior of the state is carried by the most probable state $\hat{\mathbf{z}}$, the covariance $\boldsymbol{\Sigma}$ and the Jacobian $\mathbf{J}$ of the rows, and every answer is an entry of the state read off them or a function of the state carried through them.

1.1.1 Several questions about one equipment hall

A planner reviewing the hall's design might ask any of the following:

1. What is the hall air temperature at today's load?
2. What is the effective conductance of the air path?
3. Will the hall carry its 60 kW design load without going over its limit?
4. Is the hall sensor reading true, or is it biased?
5. Would a measurement taken before the sign-off narrow the answer?
6. What happens if the cooling unit loses half its airflow?

The questions are different. The first two ask for the value of an unknown quantity; the third asks for the probability that a future temperature stays within a limit; the fourth asks which of two descriptions of the sensor is true; the fifth asks what a measurement not yet taken would be worth; the sixth asks about an operating condition that has not occurred. All six, however,

concern the same physical system: the same hall, the same air path, the same sensor.

Constructing a separate mathematical model for every question would repeat the same physics six times, and the six models could disagree about it. The method of this book begins instead with one common representation of the equipment and its physical relations, the *plant graph*. It is step 0 of the method stated in *How to read this book*: the system as it is, before any question is asked, built once and kept for every question that follows. Section 1.1.6 writes the questions that this chapter answers over the graph.

1.1.2 The system, the question and the evidence

Three things must be fixed before any probabilistic inference is performed, and the plant graph is where they are fixed. The first is the division of the quantities: which are known and pinned (the supply-air temperature, the heat load), which are unknown and free (the hall temperature, the conductance), and which are observed (the sensor reading). The field often separates the unknowns into *states*, such as temperatures and flows, and *parameters*, such as conductances and rates; in this book both kinds are entries of one state vector, and the inference treats them alike. The second is the set of physical relations that must hold among the quantities, such as conservation of energy at the hall. The third is the evidence: the readings, the records and the prior knowledge, attached to the graph as rows of their own.

The question is a separate object. It is stated over the graph after the graph is built, and it decides what is computed, not what the system is. The separation makes the questions of a design review systematic: when a new question arrives, the graph is already in hand, and the work is to write the question over its state.

> **Principle 1.1.** *The physical model is reusable; the probabilistic question determines what is calculated from it. A plant graph is built once from the equipment and its physics, and every question is a probability or an expectation over the state of that graph.*

1.1.3 Larger plant graphs and factor graphs

The equipment hall of this chapter has one thermal path and two unknown quantities, the smallest graph with every kind of element. The same construction serves a plant of many components, and the graph on which an inference operates is a second object, drawn from the plant graph and the evidence attached to it.

A cooling network. Real plants may include hundreds or thousands of components, each with its own equations and uncertainties. Consider a cooling network with two chillers, two pumps, a common electrical supply and several temperature sensors. Its plant graph may contain temperatures, pressures, flows and electrical states at its nodes; physical components represented by paths or by component relations; heat, mass and power conservation equations; telemetry and historical records; the run and trip states of equipment; and shared parameters and shared dependencies.

Figure 1.1 draws a reduced version. The green boxes are free nodes: the two chillers and the two pumps, each carrying its status and capacity or flow, the supply header with its water temperature T_s and flow, and the hall with its temperature T. The gray box on the left is the pinned outdoor air, to which the chillers reject heat; the white box in the middle is the common electrical supply, which feeds all four machines; the white box at the upper right is the heat load of the hall. Solid green arrows are physical paths of heat and water; thin gray arrows are electrical feeds; dashed arrows are observations, two temperature readings and the maintenance record of a chiller. The common supply is a single node on which four components depend: if it trips, all four stop together, and a graph that omitted it would treat their trips as independent.

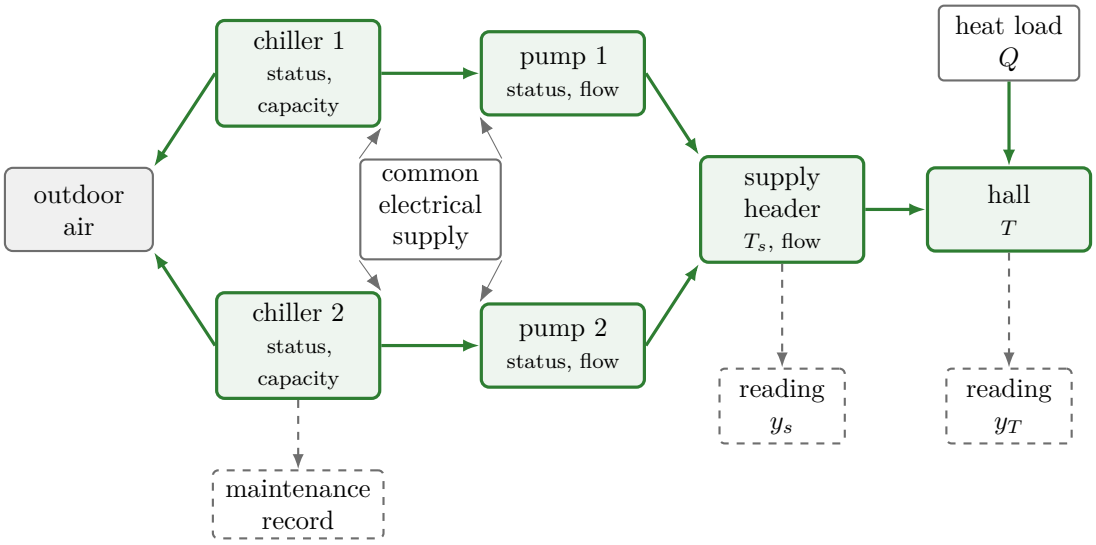


Figure 1.1: Plant graph of a reduced cooling network (illustrative; return piping omitted). Free: the status and capacity of each chiller, the status and flow of each pump, the header water temperature T_s and flow, and the hall temperature T. Pinned: the outdoor air, to which the chillers reject heat, and the heat load Q; the common electrical supply feeds all four machines and is the shared dependency. Observed: T_s and T through two readings, and chiller 2 through its maintenance record.

The mathematical principle is the one the single hall shows below. The network is a collection of physical relations,

$$r_1(\mathbf{z}) = 0, \quad r_2(\mathbf{z}) = 0, \quad \ldots, \quad r_m(\mathbf{z}) = 0,$$

one per balance and per component relation, over a state vector $\mathbf{z}$ that now has many entries. Observations and probability models for the uncertain quantities are attached as rows of their own. What is ultimately sought is the joint posterior of the state,

$$p(\mathbf{z} \mid \mathbf{y}, M),$$

where M represents the assumed system model, including its structure: which components exist, how they are connected and which dependencies are represented. Writing M explicitly matters because the structure itself may be uncertain, and a question about the structure is answered by comparing two such models, not by the posterior inside either.

Physical graphs and factor graphs. A physical graph and a probabilistic factor graph are related, but they are not the same object. The physical graph represents equipment and connections: halls, paths, chillers, pumps, sensors. A *factor graph* represents the variables of the inference and the mathematical relations among them, including the observations, the priors and the dependencies: each entry of the state is one node, each row of the model is one factor node, and each factor is linked to the entries its row involves. It is constructed from the physical graph and the evidence attached to it, and it shows which row acts on which unknown. The distinction becomes important when competing system structures and diagnosis are studied in later chapters. Two physical drawings of the same plant, one with a single air path and one with two, give two factor graphs with different variables and rows, and comparing them is comparing two models M. A sensor bias adds a variable to the factor graph without changing the equipment at all.

1.1.4 The plant graph (step 0)

The simplest plant graph with every kind of element has one free node, one pinned node, one path, one source and one observation, and the equipment hall is that graph. It is built before any question is asked and before any number is fitted.

Nodes, a path, a source and an observation. Consider a simplified equipment hall with one heat-removal path, the hall's air path to the supply air that the cooling unit delivers. Figure 1.2 draws it as a plant graph. The green box on the left is the hall, a free node whose air temperature T is unknown; the gray box on the right is the supply air, a pinned node whose temperature $T_c = 18°\text{C}$ is held by the cooling unit's discharge controller. The solid green arrow between them is the heat-removal path, characterized by its effective thermal conductance G. The box above the hall is the heat load Q supplied by the servers, a source. The dashed box below is the hall sensor, and the dashed arrow from the hall to it is an observation: the sensor reports a value related to T, not T itself. Solid connections are physical relations; the dashed connection is evidence.

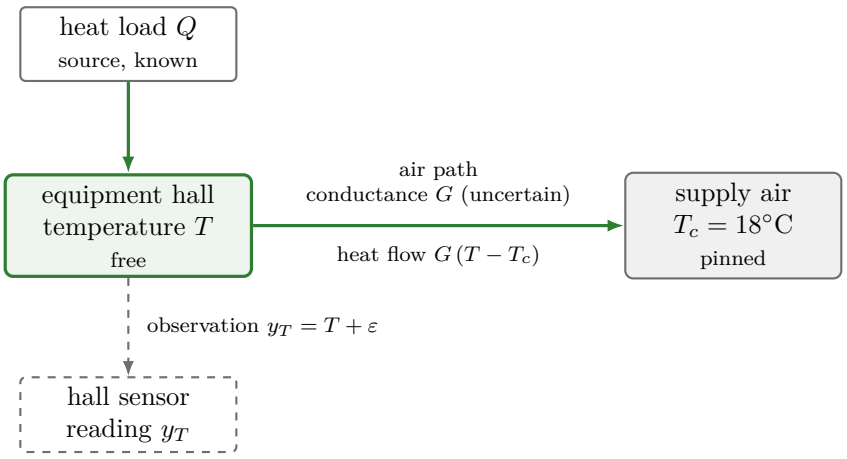


Figure 1.2: Plant graph of the equipment hall in steady state. Free: the hall air temperature T and the conductance G of the air path. Pinned: the supply-air temperature $T_c = 18°\text{C}$, held by the cooling unit's discharge controller, and the heat load Q. Observed: T, through the reading y_T with error ε. Solid arrows are physical relations; the dashed arrow is an observation.

Table 1.1 lists the elements of the graph. The first column names each element, the second its role in the graph, the third its symbol and the fourth its unit; the last column states how the element enters this first example.

Table 1.1: The elements of the plant graph of Figure 1.2.

Element	Role	Symbol	Unit	In the example
Equipment hall	free node	T	°C	air temperature unknown
Supply air	pinned node	T_c	°C	held at 18°C by the discharge controller
Heat load	source	Q	kW	thermal power supplied by the servers, known
Air path	path (edge)	G	kW/K	effective conductance, uncertain
Hall sensor	observation	y_T	°C	a reading related to the hall temperature

For now the supply-air temperature and the heat load are assumed known, and the effective conductance is uncertain. The discharge controller holds T_c at its set point, which is what makes the supply air a pinned boundary; the choice fails if the unit saturates and can no longer hold it (Section 1.6.5). The conductance lumps every mechanism by which the hall's heat reaches the supply air, the airflow through the cooling unit and the mixing of the hall air alike, into one number; how well that lumping represents the hall is a modeling assumption, examined in Section 1.6.5. This is the starting physical model.

Conservation of energy at the hall node. A graph becomes useful for computation when the mathematical relations between its elements are specified. Assume steady state: the hall stores no more and no less heat from one moment to the next, so that its temperature is constant. Conservation of energy at the hall node then requires the heat entering to equal the heat leaving,

$$Q_{\text{in}} = Q_{\text{out}},$$

where Q_{in} and Q_{out} are thermal powers in kilowatts. The heat entering is the load, $Q_{\text{in}} = Q$. The heat leaving is modeled as proportional to the temperature difference across the path,

$$Q_{\text{out}} = G\,(T - T_c),$$

the defining relation of the effective conductance: G is the heat flow per kelvin of difference between the hall and the supply air. Combining the two,

$$Q = G\,(T - T_c), \tag{1.1}$$

where Q is the heat generated by the servers in kW, G the effective thermal conductance in kW/K, T the hall air temperature in °C and T_c the supply-air temperature in °C. A temperature difference has the same value in kelvin and in degrees Celsius, so the units balance: kW/K times K is kW.

Equation (1.1) is a *physical relation*, not a probability statement. It states what must hold if the simplified thermal model is correct, whatever is or is not known about its quantities. Since $G > 0$, dividing by G and adding T_c gives

$$T = T_c + \frac{Q}{G}, \tag{1.2}$$

which gives the temperature when the other three quantities are specified. The hall is warmer than the supply air by Q/G: more load or less conductance raises it.

The forward calculation. The hall carries the load $Q = 30\,\text{kW}$; take as the reference for the forward calculation a conductance of $G = 6\,\text{kW/K}$ (the reference G_{ref}, not the handbook's $7.5\,\text{kW/K}$). With $T_c = 18°\text{C}$, equation (1.2) gives

$$T = 18 + \frac{30\,\text{kW}}{6\,\text{kW/K}} = 18 + 5 = 23°\text{C},$$

the steady-state temperature of the hall at its design conductance. The calculation used no probability, estimated no uncertainty and used no observational evidence. It is a deterministic *forward calculation*: from pinned inputs and a specified conductance to the temperature they imply. Every later chapter needs it, because every inference compares temperatures computed this way with what the sensor reports. Doubling the load doubles the excess $T - T_c$ at every conductance: at $G = 6\,\text{kW/K}$ the hall at 60 kW would be at 28°C.

The state vector. The forward calculation required the conductance to be known. In the inference problems of this book it is not, and neither is the hall temperature. The *state vector* is the collection of the unknown quantities that the inference estimates at once. In the equipment hall the unknowns are the hall temperature and the thermal conductance, and one possible representation is

$$\mathbf{z} = \begin{bmatrix} T \\ G \end{bmatrix}.$$

The state holds everything the inference treats as free; the pinned quantities T_c and Q are not in it, and neither is the reading y_T, which is evidence. The concept recurs throughout the book: the graph fixes the state, the evidence constrains it, and each question is a statement about it. Quantities that are entries of the state, such as T here, are answered directly by the inference; quantities that are functions of the state, such as the heat flow through the path, require one further transformation.

The log-conductance. A conductance is positive, and an estimate of it must be positive too. In practice the conductance is therefore often represented by a logarithmic entry,

$$g = \ln\!\left(\frac{G}{G_{\mathrm{ref}}}\right),$$

where $G_{\mathrm{ref}} > 0$ is a reference conductance in kW/K and g is dimensionless. Inverting,

$$G = G_{\mathrm{ref}}\, \mathrm{e}^{g},$$

which is positive for every real g: an inference free to move g anywhere on the real line can never produce a negative or zero conductance. A change in g is a relative change in G, since $\mathrm{d}G = G_{\mathrm{ref}}\, \mathrm{e}^{g}\, \mathrm{d}g = G\, \mathrm{d}g$, so that $\mathrm{d}g = \mathrm{d}G/G$; uncertainty in g reads directly as a fractional uncertainty in G. The reference value only sets the origin of g. In this chapter it is taken as $G_{\mathrm{ref}} = 6\,\mathrm{kW/K}$, the design conductance of the forward calculation, so that $g = 0$ there, $G = 7.5\,\mathrm{kW/K}$ corresponds to $g = \ln 1.25 = 0.2231$. The state vector becomes

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} T \\ g \end{bmatrix}, \tag{1.3}$$

with entries $z_1 = T$ in °C and $z_2 = g$, dimensionless. It is the one state of this chapter: every question below is a statement about these two entries.

The physics row. The thermal relation can now be written in terms of the state. Substituting $G = G_{\mathrm{ref}}\, \mathrm{e}^{g}$ into equation (1.1),

$$Q = G_{\mathrm{ref}}\, \mathrm{e}^{g}\, (T - T_c),$$

or, moving everything to one side,

$$r_{\mathrm{phys}}(\mathbf{z}) = G_{\mathrm{ref}}\, \mathrm{e}^{g}\, (T - T_c) - Q = 0. \tag{1.4}$$

The function $r_{\mathrm{phys}}(\mathbf{z})$ is a *residual*, in kW. A residual measures how far a candidate state is from satisfying a relation: it is zero when the relation holds exactly and grows as the candidate departs from it. With $G_{\mathrm{ref}} = 6\,\mathrm{kW/K}$, $T_c = 18°\mathrm{C}$ and $Q = 30\,\mathrm{kW}$, the candidate $\mathbf{z} = (23°\mathrm{C},\ 0)$ gives $r_{\mathrm{phys}} = 6 \times 1 \times 5 - 30 = 0$: it is the forward calculation. The candidate $\mathbf{z} = (23.2°\mathrm{C},\ 0)$ gives $6 \times 5.2 - 30 = 1.2\,\mathrm{kW}$: at that temperature and conductance the path would remove 1.2 kW more than the servers supply, so the hall could not be in steady state. The candidate $(23.2°\mathrm{C},\ \ln(30/31.2)) = (23.2°\mathrm{C},\ -0.0392)$ again gives zero: the temperature 23.2°C is consistent with the physics if the conductance is $30/5.2 = 5.769\,\mathrm{kW/K}$. The zero set

of r_{phys} is the curve $T = T_c + Q/(G_{\text{ref}}\,\mathrm{e}^{g})$ in the plane of (g, T); the physics alone allows every point of that curve and no point off it.

The same row has a second form, in the units of the quantity it fixes. Dividing equation (1.4) by $G_{\text{ref}}\,\mathrm{e}^{g} = G$, which is positive, leaves the zero set unchanged and gives the row in kelvin, the departure of a candidate temperature from the temperature the physics implies. Divided further by a *model tolerance* σ_p, the temperature error in kelvin that the physics is allowed, it becomes a dimensionless row,

$$r_1(\mathbf{z}) = \frac{T - T_c - (Q/G_{\text{ref}})\,\mathrm{e}^{-g}}{\sigma_p}, \tag{1.5}$$

which counts the violation of the physics in multiples of its tolerance. This is the form in which the physics row enters every computation of the book. The formulation extends to systems of any size. A larger physical system is represented by a collection of residuals, one per relation: an electrical network has rows for Kirchhoff's current and voltage laws, a fluid network rows for mass conservation at its junctions and for the pressure losses along its pipes, and a cooling plant rows for the energy balance at each of its nodes. Each row has the form of equation (1.4), and the inference treats them alike.

Exact constraints and model discrepancy. Physical rows and reading rows should not automatically be treated as equivalent kinds of noisy measurement. Conservation of energy is a law: if the hall is in steady state and the model has captured every path, equation (1.4) holds exactly, and it can be imposed as an *exact constraint*, $r_{\text{phys}}(\mathbf{z}) = 0$, which restricts the state to the zero set of the row. In a computation the exact constraint is carried as the physics row (1.5) with a model tolerance σ_p far smaller than every other spread of the problem, and exactness is the limit $\sigma_p \to 0$; the limit is checked by showing that the results stop moving as σ_p shrinks, never by using the row to eliminate an entry of the state. A reading row, by contrast, is meant to be violated by an amount of the order of the sensor's error: the reading is not the temperature.

Approximate physics is a third case. The single effective conductance of equation (1.1) is an idealization; the real hall has several paths, a conductance that varies with airflow, and stored heat. When the relation is only approximately true, the honest model is

$$G_{\text{ref}}\,\mathrm{e}^{g}\,(T - T_c) - Q = \delta,$$

with an explicit *model discrepancy* δ in kW, an entry of its own with a stated prior, for example $\delta \sim \mathcal{N}(0, \sigma_\delta^2)$. Algebraically this is a physics row with the weight $1/\sigma_\delta^2$, and in the objective it looks like a reading row. It behaves differently. A measurement error is new at each reading and averages out over many readings; a discrepancy is a property of the model and repeats wherever the same operating condition repeats, so that many readings at one condition do not reduce it. Treating a discrepancy as measurement noise therefore makes the inference overconfident, and treating a law as noisy lets the inference trade physics against data. The choice between an exact constraint and an explicit discrepancy is made when the graph is built, and it is stated with the model. In this chapter conservation of energy is exact.

1.1.5 The record and the handbook value

The record of the hall is one reading. The hall sensor, the one temperature sensor installed in the hall, reported at commissioning, under today's load of 30 kW,

$$y_T = 23.2^{\circ}\mathrm{C}.$$

The sensor is not perfectly accurate. Its measurement model is

$$y_T = T + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_T^2),$$

where ε is the measurement error in kelvin, assumed normal with zero mean, and σ_T its standard deviation, which the sensor's specification gives as

$$\sigma_T = 0.5\,\mathrm{K}.$$

The reading is the only observation: the supply-air temperature and the load are pinned, and nothing measures the conductance.

Knowledge of the conductance that does not come from the reading is also in hand. A design handbook, the commissioning documents or earlier tests of similar halls give an estimate of the effective conductance with an uncertainty; for this hall (illustrative) the handbook gives 7.5 kW/K, with the true conductance within a factor of about 1.35 of it with probability of about 0.68. The handbook value is not a measurement of this hall. It is information held before the reading, and it enters the model as a prior.

1.1.6 The questions (step 1)

The planner's questions can now be written over the graph, each as a probability or an expectation over the state, with the unknown on the left of the bar and the evidence on the right. Table 1.2 lists the five that this chapter answers: the first column is the question in the planner's words; the second the general form that every question of its category takes, written in the book's notation for a quantity $f(\mathbf{z})$ derived from the state, a threshold f^* on it and the posterior density $p(\mathbf{z} \mid \mathbf{y})$; the third its mathematical form for the hall; the fourth its category among the five named at the front of the book, which Chapter 2 develops; and the last the subsection where it is answered. The first two are estimation, of an entry of the state and of a quantity derived from one, the conductance $G = G_{\mathrm{ref}}\,\mathrm{e}^{g}$. The third and fourth are predictions under a threshold, tail probabilities of the posterior: the third at the present load, the fourth at a load of 60 kW, the design load, which the hall has never carried, whose temperature T_2 is a quantity the graph acquires in step 2. The fifth is a design question about a measurement not yet taken; it asks nothing new of the present evidence, and it is answered as the next measurement for the first four.

The questions differ in what they compute and agree in what they are computed over: the same two entries of the state, under the same posterior. Two of the planner's questions are not in the table. Whether the hall sensor is biased compares two complete models of the same reading, the plant with an unbiased sensor and the plant with a bias entry on the reading row, and its answer is a ratio of their evidences rather than a property of either posterior; when every sensor of a larger plant is a candidate, the same comparison run over the candidates is a diagnosis. *Operations and Control by Derivation* answers it in its Chapter 7, and the residual check of Section 1.4.3 takes a first, coarse look at it. What happens if the cooling unit loses half its airflow is a question about a transient that the steady-state graph cannot represent, and Section 1.6.5 states what the graph would need for it.

1.1.7 The evidence as rows (step 2)

Step 2 attaches the reading and the handbook value to the graph as rows on the state $\mathbf{z} = (T, g)^{\mathsf{T}}$, and extends the graph with the operating condition the design-load question names. The three rows, the physics row of Section 1.1.4 among them, are Gaussian, and each is written normalized, with a residual and a Jacobian row.

The modeling choices. One choice makes the physics row, and it buys something and can fail. The row carries the balance of Section 1.1.4 in the limit $\sigma_p \to 0$: this is what confines the posterior to the physics curve, so that a reading of the temperature informs the conductance, and it fails if heat leaves the hall by a mechanism the single conductance does not lump or if the hall is not in steady state (Section 1.6.5).

Table 1.2: The questions asked of the equipment hall, the general form of their category, their mathematical form for the hall, their category and where each is answered. T_2 is the hall temperature at the load $Q_2 = 60\,\mathrm{kW}$ (Section 1.1.7); $\boldsymbol{\Sigma}_a$ the covariance expected after a candidate row with Jacobian row $\mathbf{J}_a$.

Planner's question	General form	Mathematical form	Category	Answered in
What is the hall air temperature at today's load?	$\mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}]$, $\mathrm{Var}(f(\mathbf{z}) \mid \mathbf{y})$	$p(T \mid \mathbf{y})$, $T = z_1$	estimation	1.3.1
What is the effective conductance of the air path?	$\mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}]$, $\mathrm{Var}(f(\mathbf{z}) \mid \mathbf{y})$	$p(G \mid \mathbf{y})$, $G = G_{\mathrm{ref}}\,\mathrm{e}^{z_2}$	estimation of a derived quantity	1.3.1
Is the hall below 24°C now?	$\Pr(f(\mathbf{z}) < f^* \mid \mathbf{y})$	$\Pr(T < 24^\circ\mathrm{C} \mid \mathbf{y})$	prediction under a threshold	1.3.2
Will the hall carry its 60 kW design load inside 27.5°C?	$\Pr(f(\mathbf{z}) < f^* \mid \mathbf{y})$	$\Pr(T_2 < 27.5^\circ\mathrm{C} \mid \mathbf{y})$, $T_2 = T_c + (Q_2/G_{\mathrm{ref}})\,\mathrm{e}^{-z_2}$	prediction under a threshold	1.3.2
Would a measurement before the sign-off narrow the answers?	preposterior covariance $(\boldsymbol{\Sigma}^{-1} + \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a)^{-1}$	$\boldsymbol{\Sigma}_a$ for each candidate row	design	1.6.1

The reading row. The reading of Section 1.1.5 becomes the normalized observation residual

$$r_{\mathrm{obs}}(\mathbf{z}) = \frac{T - y_T}{\sigma_T}, \tag{1.6}$$

which is dimensionless: the difference between the candidate temperature and the reading, in units of the sensor's standard deviation. At $T = 23.2^\circ\mathrm{C}$ it is zero; at the design temperature $T = 23^\circ\mathrm{C}$ it is $(23 - 23.2)/0.5 = -0.4$. The row carries the probability of the reading. Given the state, the reading differs from T only by the error, so its density is the normal density of ε evaluated at $y_T - T$:

$$p(y_T \mid \mathbf{z}) = \frac{1}{\sqrt{2\pi}\,\sigma_T} \exp\left[-\frac{(y_T - T)^2}{2\sigma_T^2}\right] = \frac{1}{\sqrt{2\pi}\,\sigma_T} \exp\left[-\tfrac{1}{2}\, r_{\mathrm{obs}}(\mathbf{z})^2\right].$$

Read as a function of the state for the fixed reading, this is the *likelihood* of the state. A candidate state with a small $|r_{\mathrm{obs}}|$ makes the reading probable; the ratio of the likelihoods of $T = 23^\circ\mathrm{C}$ and $T = 23.2^\circ\mathrm{C}$ is $\mathrm{e}^{-0.4^2/2} = \mathrm{e}^{-0.08} = 0.923$, so the reading alone hardly distinguishes the two. The likelihood does not depend on g: the sensor sees the temperature, and it informs the conductance only through the physics row that ties the two together.

The prior row. The handbook value becomes a row of the same form. Written as a normal distribution on the log-conductance, $g \sim \mathcal{N}(g_0, \sigma_g^2)$, it is the row $(g - g_0)/\sigma_g$: a handbook value is treated as a reading of g taken before the hall's own. The handbook conductance of 7.5 kW/K gives $g_0 = \ln(7.5/6) = 0.2231$, and $\sigma_g = 0.3$ states that the true conductance is within a factor $\mathrm{e}^{0.3} = 1.35$ of it with probability of about 0.68. The prior is written $p(g)$ in general; its form is a modeling choice that must be stated and justified, and Section 1.4.4 tests how much this one weighs.

Three kinds of row. The model now holds three very different kinds of relation. Table 1.3 sets them side by side. Each row of the table is one kind of relation; the columns give its equation for the hall, its residual and its role.

Table 1.3: The three kinds of row in the equipment hall model, with $G = G_{\text{ref}}\,\mathrm{e}^{g}$.

Row	Equation	Residual	Role
Physics	$G\,(T - T_c) - Q = 0$	$G_{\text{ref}}\,\mathrm{e}^{g}(T - T_c) - Q$	constrains what is physically possible
Reading	$y_T = T + \varepsilon$	$(T - y_T)/\sigma_T$	provides imperfect information about the state
Prior	$g \sim \mathcal{N}(g_0, \sigma_g^2)$	$(g - g_0)/\sigma_g$	carries knowledge held before the readings

The distinction between the first two rows is foundational. The physical equation constrains the system: it says which combinations of temperature and conductance can occur at all. The reading provides evidence about the unknown quantities: it says which of the possible combinations are plausible in light of what was measured. The prior row has the form of a reading row and the role of knowledge held in advance. A probabilistic inference method combines the three kinds of relation to obtain a distribution over the plausible states.

Rows and weights. Chapter 3 shows that, for normal errors, the negative logarithm of the posterior is, up to a constant, the objective $\mathcal{L}(\mathbf{z}) = \frac{1}{2}\sum_k w_k\, r_k(\mathbf{z})^2$, a weighted sum of the squared rows with weights $w_k = 1/\sigma_k^2$. A reading row may be written either normalized, as in equation (1.6), with unit weight, or in kelvin, as $T - y_T$ with weight $1/\sigma_T^2$; both put the same term $(T - y_T)^2/\sigma_T^2$ into the objective. Every row of this chapter is written normalized: the physics row in the form (1.5) with the tolerance σ_p, the reading row (1.6) and the prior row, so that every row counts in standard deviations and none dominates because of its unit.

The scenario node. The design-load question doubles the electrical load from 30 kW to 60 kW, an operating condition the hall has never carried. A second operating condition is added to the graph, sharing the same unknown conductance. Figure 1.3 draws it. The upper row is the present condition, at which the hall has been operated and observed: the free node at $Q_1 = Q = 30\,\text{kW}$ with temperature T, its path to the pinned supply air, and its sensor. The lower row is the design-load condition, a scenario node at $Q_2 = 60\,\text{kW}$ with temperature T_2 and its own path to the same supply air. The vertical dashed line between the two paths marks that they carry one and the same conductance. The scenario node has no sensor: it is in the graph only because the question asks about it.

The two operating conditions satisfy the energy balance of equation (1.1) each at its own load,

$$Q_1 = G\,(T - T_c), \tag{1.7}$$

$$Q_2 = G\,(T_2 - T_c), \tag{1.8}$$

with $Q_1 = Q$. The two temperatures are different, but the conductance is shared, and that connection is what makes the second condition predictable from the first. The scenario node adds no evidence and no free quantity: by equation (1.8) its temperature is fixed by the conductance, $T_2 = T_c + (Q_2/G_{\text{ref}})\,\mathrm{e}^{-g}$, a function of the state $\mathbf{z} = (T, g)^{\mathsf{T}}$, which keeps its two entries and its three rows. Chapter 4 carries T_2 instead as a third entry with a second physics row, the same formulation; with the same rows it gives the same answers.

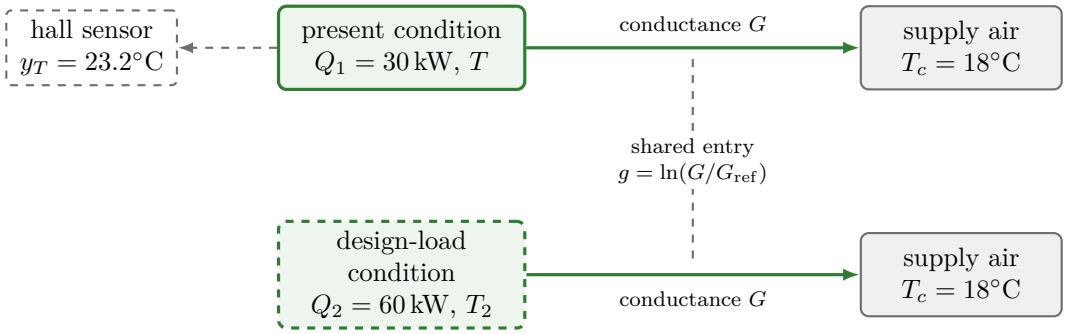


Figure 1.3: Two operating conditions of the equipment hall linked through the same physical conductance. Free: the present-condition temperature T and the shared log-conductance g, the two entries of the state; the design-load temperature T_2 (dashed node) is a function of g. Pinned: the supply-air temperature and the two loads. Observed: T only, through the reading y_T; the scenario node is unobserved and enters the graph because the question asks about it.

Which row acts on which entry. The three rows act on the two entries of the state in a pattern fixed by the graph, before any number is computed. The physics row ties the temperature to the conductance and involves both; the sensor sees the temperature and not the conductance; the handbook speaks about the conductance and not the temperature; the scenario node adds no row. The Jacobian of the rows, the matrix of their derivatives with respect to the entries, can therefore be nonzero only where marked $*$:

$$\mathbf{J} = \begin{bmatrix} * & * \\ * & 0 \\ 0 & * \end{bmatrix} \begin{matrix} \text{physics row} \\ \text{reading row,} \\ \text{prior row} \end{matrix} \qquad \text{columns } z_1 = T,\ z_2 = g. \tag{1.9}$$

The zeros are structural: they hold at every state. Only the physics row has entries in both columns, and it is the one row through which a reading of the temperature reaches the conductance.

The factor graph of the hall. The pattern (1.9) is what a factor graph (Section 1.1.3) draws. Figure 1.4 draws it for the hall with its scenario node. The two solid circles are the entries of the state, T and g. The black squares are the rows: the physics row of equation (1.5), linked to both entries; the reading row, linked to T alone; and the prior row, linked to g alone. The dashed circle is the design-load temperature T_2, joined to g by a dashed line: it is a function of g through equation (1.8), not an entry, and no row acts on it. The figure makes visible what the physical graph only implies: T_2 has no reading of its own and is connected to the evidence only through g, and g is connected to the reading only through the physics row. The physical drawing has two supply-air boxes and two paths; the factor graph has no node for the supply air at all, since T_c is pinned and appears only as a constant inside the physics row.

The residual vector. The state is the column vector of equation (1.3),

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} T \\ g \end{bmatrix},$$

with T in °C and g dimensionless. Three rows act on it, the physics row (1.5), the reading row (1.6) and the prior row, each divided by its standard deviation. Stacked into one residual

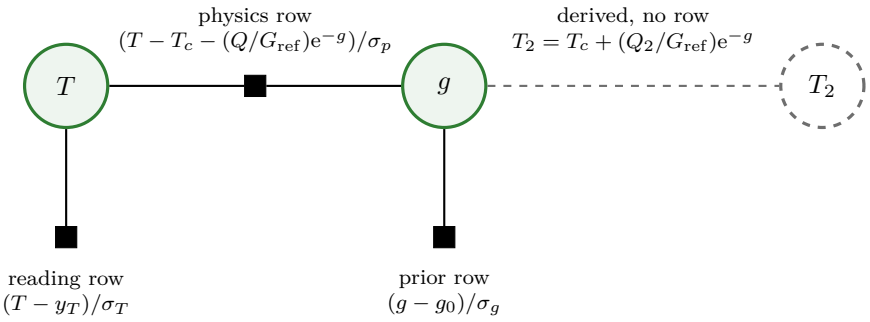


Figure 1.4: Factor graph of the equipment hall with its scenario node. Solid circles: the entries of the state, T and g. Squares: the rows of the model, each linked to the entries it involves, the pattern (1.9) of the Jacobian. Dashed: the design-load temperature T_2, a function of g with no row of its own, which reaches the evidence only through the shared g.

vector, the three rows are

$$\mathbf{r}(\mathbf{z}) = \begin{bmatrix} r_1(\mathbf{z}) \\ r_2(\mathbf{z}) \\ r_3(\mathbf{z}) \end{bmatrix} = \begin{bmatrix} (T - T_c - (Q/G_{\mathrm{ref}})\,\mathrm{e}^{-g})/\sigma_p \\ (T - y_T)/\sigma_T \\ (g - g_0)/\sigma_g \end{bmatrix}, \tag{1.10}$$

with $T_c = 18°\mathrm{C}$, $Q = 30\,\mathrm{kW}$, $G_{\mathrm{ref}} = 6\,\mathrm{kW/K}$, the reading $y_T = 23.2°\mathrm{C}$ with $\sigma_T = 0.5\,\mathrm{K}$, and the handbook prior $g_0 = \ln(7.5/6) = 0.2231$ with $\sigma_g = 0.3$. Each entry of $\mathbf{r}$ is one row: a dimensionless number stating by how many standard deviations a candidate state violates one relation. The objective is half the sum of their squares,

$$\mathcal{L}(\mathbf{z}) = \tfrac{1}{2}\,\mathbf{r}(\mathbf{z})^{\mathsf{T}}\,\mathbf{r}(\mathbf{z}) = \tfrac{1}{2}\sum_{k=1}^{3} r_k(\mathbf{z})^2, \tag{1.11}$$

which Chapter 3 shows to be minus the logarithm of the posterior up to a constant, so that the most probable state $\hat{\mathbf{z}}$, the maximum a posteriori (MAP) state, is the minimizer of $\mathcal{L}$. The computation takes $\sigma_p = 0.01\,\mathrm{K}$, fifty times tighter than the sensor, and Section 1.4.1 checks that the limit has been reached.

The Jacobian. The Jacobian $\mathbf{J}$ collects the derivative of every row with respect to every entry of the state: its entry in row k and column i is $\partial r_k/\partial z_i$, so it has one row per row of the model and one column per entry of the state, here three by two. Differentiating equation (1.10) entry by entry,

$$\mathbf{J}(\mathbf{z}) = \frac{\partial \mathbf{r}}{\partial \mathbf{z}} = \begin{bmatrix} 1/\sigma_p & (Q/G_{\mathrm{ref}})\,\mathrm{e}^{-g}/\sigma_p \\ 1/\sigma_T & 0 \\ 0 & 1/\sigma_g \end{bmatrix}, \tag{1.12}$$

the pattern (1.9) with its entries filled in. Each entry is a sensitivity in standard deviations: how far one row moves when one entry of the state moves by one unit. Only the entry in the first row and second column depends on the state; the others are constants. The first row is the physics row's own sensitivities in its physical unit, rescaled. Differentiating the kilowatt form of equation (1.4),

$$\frac{\partial r_{\mathrm{phys}}}{\partial T} = G_{\mathrm{ref}}\,\mathrm{e}^{g} = G, \qquad \frac{\partial r_{\mathrm{phys}}}{\partial g} = G_{\mathrm{ref}}\,\mathrm{e}^{g}\,(T - T_c) = G\,(T - T_c),$$

in kW/K and kW; on the zero set the second equals the load Q by equation (1.1). A one-kelvin error in the temperature unbalances the hall by G kilowatts, and a unit change in g, a factor e

in the conductance, by about Q kilowatts to first order. Divided by G, to reach the temperature form, and by σ_p, the pair becomes $(1/\sigma_p,\ (Q/G)/\sigma_p)$, the first row of equation (1.12) on the zero set.

1.2 The posterior of the state (step 3)

All the questions of Table 1.2 are read from one posterior: the posterior of the state that the physics row, the reading row and the prior row of step 2 give, approximately the Laplace normal of (T, g) and exactly a density along the physics curve. This section derives it and computes it for the hall; Section 1.3 reads every answer from it.

The objects of the fit recur in the specialized forms, and they are named here by their roles. The *MAP state* $\hat{\mathbf{z}} = (\hat{T}, \hat{g})^{\mathsf{T}}$, the maximum a posteriori state, is the minimum of the objective that the three rows put on the state, the most probable state; the *information matrix* $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ is the curvature that the rows give that objective at the minimum, with $\mathbf{J}$ the 3×2 Jacobian of the normalized rows; the *Laplace covariance* $\mathbf{\Sigma}$, a 2×2 matrix, is the inverse of that curvature after a term from the physics row's own curvature is added to it, and it makes the posterior of the state approximately normal, $\mathbf{z} \mid \mathbf{y} \approx \mathcal{N}(\hat{\mathbf{z}}, \mathbf{\Sigma})$; and a derived quantity f is carried by its *gradient* $\mathbf{g} = \partial f/\partial \mathbf{z}$ at $\hat{\mathbf{z}}$ (bold, to tell it from the entry g), with the variance $\mathbf{g}^{\mathsf{T}}\mathbf{\Sigma}\mathbf{g}$. Section 1.2.1 derives them from the rows. Beside them stands the exact posterior that this problem admits in the limit $\sigma_p \to 0$: on the physics curve the state has one free direction, and the posterior is the density of one entry, evaluated by quadrature. The Laplace route, a normal distribution centered on the MAP state, is what lets every answer be read from $\hat{\mathbf{z}}$ and $\mathbf{\Sigma}$ by operations that serve a plant of any size, and it misleads where the posterior is skewed along the curved ridge, which Section 1.4.2 measures against the exact posterior. The chapter reports the Laplace values, sets the exact value beside each answer with the gap, and collects the comparison in Section 1.4.

1.2.1 The posterior from the rows

The answers use the objects of one posterior: the MAP state $\hat{\mathbf{z}}$, the covariance $\mathbf{\Sigma}$ and the Jacobian $\mathbf{J}$ of the rows. The Jacobian is written out in Section 1.1.7; the other two are derived here for the hall. Chapter 3 derives these objects from the probability rules for Gaussian rows over any state, and Chapter 4 dissects the same objects in full on the hall with its scenario node carried as an entry, including the Laplace approximation and its curvature term. Here each is written out once for a reader meeting it for the first time.

The posterior. The ingredients are in hand: the unknown state $\mathbf{z} = (T, g)^{\mathsf{T}}$; the observation $y_T = 23.2°\mathrm{C}$; the physical relation $Q = G\,(T - T_c)$; and the prior information $p(g)$. What is ultimately wanted is

$$p(\mathbf{z} \mid \mathbf{y}), \tag{1.13}$$

the *posterior distribution* of the state, conditional on the evidence $\mathbf{y}$, here the reading y_T and the handbook value, and on the assumed physical model. It states which values of the state remain plausible after the physics, the reading and the prior information have been combined. Chapter 3 derives Bayes' theorem, by which the posterior is proportional to the likelihood times the prior, $p(\mathbf{z} \mid \mathbf{y}) \propto p(y_T \mid \mathbf{z})\, p(\mathbf{z})$; here the reading row plays the part of the likelihood, and the physics row and the prior row on g together play the part of the prior $p(\mathbf{z})$.

The normal equations and the Gauss–Newton step. Near a state $\mathbf{z}$, each row is replaced by its tangent, $\mathbf{r}(\mathbf{z} + \Delta\mathbf{z}) \approx \mathbf{r}(\mathbf{z}) + \mathbf{J}\,\Delta\mathbf{z}$, and the objective by $\frac{1}{2}\,\|\mathbf{r} + \mathbf{J}\,\Delta\mathbf{z}\|^2$, a quadratic in the

step $\Delta\mathbf{z}$. Its gradient with respect to the step is $\mathbf{J}^\mathsf{T}(\mathbf{r} + \mathbf{J}\,\Delta\mathbf{z})$, and setting it to zero gives the *normal equations*

$$(\mathbf{J}^\mathsf{T}\mathbf{J})\,\Delta\mathbf{z} = -\mathbf{J}^\mathsf{T}\mathbf{r}, \tag{1.14}$$

a linear system of two equations in the two entries of the step. The Gauss–Newton method solves it, moves to $\mathbf{z} + \Delta\mathbf{z}$, evaluates $\mathbf{r}$ and $\mathbf{J}$ there, and repeats until the step no longer changes the state at the precision wanted. The vector $\mathbf{J}^\mathsf{T}\mathbf{r}$ is the gradient of the objective (1.11), so the iteration stops where that gradient vanishes, at the MAP state. The iteration starts here from the reference state $\mathbf{z}^0 = (23°\mathrm{C},\ 0)^\mathsf{T}$ of the forward calculation, at which the physics row is zero.

The information matrix. The matrix $\mathbf{J}^\mathsf{T}\mathbf{J}$ of the normal equations is the *information matrix* of the rows. It is a sum over the rows: writing $\mathbf{J}_k$ for the k-th row of $\mathbf{J}$,

$$\mathbf{J}^\mathsf{T}\mathbf{J} = \sum_{k=1}^{3} \mathbf{J}_k^\mathsf{T}\mathbf{J}_k,$$

so that each row contributes the outer product of its own sensitivities, and the precisions of independent rows add. For the hall the three outer products are of two kinds. The reading row and the prior row each have one nonzero entry, so each contributes one diagonal entry, $1/\sigma_T^2$ in the T–T position and $1/\sigma_g^2$ in the g–g position. The physics row has two, and its outer product $\mathbf{J}_1^\mathsf{T}\mathbf{J}_1$ fills the whole matrix, off-diagonal entries included: it is the only source of the coupling between the two entries, and it has rank one, with all its information along its own row and none along the tangent of the curve. The eigenvectors of $\mathbf{J}^\mathsf{T}\mathbf{J}$ give the directions in which the rows fix the state, and its eigenvalues how sharply: a large one across the curve, set by $1/\sigma_p^2$, and a small one along it, set by the reading and the prior, as the ridge below makes exact.

The Laplace covariance and the curvature term. The inverse of the information matrix is the covariance of the posterior: exactly so for rows linear in the state (Chapter 3), and as the *Laplace approximation*, the normal distribution fitted at the MAP state, for curved rows (Chapter 4). The Gauss–Newton matrix $\mathbf{J}^\mathsf{T}\mathbf{J}$ drops the curvature of the rows themselves. The full Hessian of the objective (1.11), its matrix of second derivatives, is

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{z}\,\partial \mathbf{z}^\mathsf{T}} = \mathbf{J}^\mathsf{T}\mathbf{J} + \sum_{k=1}^{3} r_k \frac{\partial^2 r_k}{\partial \mathbf{z}\,\partial \mathbf{z}^\mathsf{T}}, \tag{1.15}$$

and of the three rows only the physics row is curved, through e^{-g}, with the single nonzero second derivative

$$\frac{\partial^2 r_1}{\partial g^2} = -\frac{(Q/G_\mathrm{ref})\,\mathrm{e}^{-g}}{\sigma_p}.$$

The *curvature term* $r_1\,\partial^2 r_1/\partial g^2$ therefore adds to the g–g entry alone. It does not vanish in the exact-physics limit, because r_1 shrinks in proportion to σ_p while $\partial^2 r_1/\partial g^2$ grows as $1/\sigma_p$: at the MAP state the gradient $\mathbf{J}^\mathsf{T}\mathbf{r}$ is zero, its T entry reads $r_1/\sigma_p + r_2/\sigma_T = 0$, so $r_1 = -(\sigma_p/\sigma_T)\,r_2$, and

$$r_1 \frac{\partial^2 r_1}{\partial g^2} = \frac{r_2}{\sigma_T}\frac{Q}{G_\mathrm{ref}}\,\mathrm{e}^{-\hat{g}}$$

at every σ_p. It is the curvature of the physics curve itself, and it acts in the one direction, along the curve, in which the information is small. The *Laplace covariance* is the inverse of the full Hessian at the MAP state,

$$\boldsymbol{\Sigma} = \Big(\mathbf{J}^\mathsf{T}\mathbf{J} + \sum_k r_k \frac{\partial^2 r_k}{\partial \mathbf{z}\,\partial \mathbf{z}^\mathsf{T}}\Big)^{-1}_{\hat{\mathbf{z}}}, \qquad \mathbf{z} \mid \mathbf{y} \approx \mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma}); \tag{1.16}$$

a Newton iteration, which uses the full Hessian in place of $\mathbf{J}^\mathsf{T}\mathbf{J}$ in equation (1.14), reaches the same $\hat{\mathbf{z}}$.

The ridge. The spreads of the two entries are not independent of each other, and the reason is read off the physics row. Its Jacobian row, the derivative of equation (1.5) with respect to (T, g), and the vector it annihilates are

$$\mathbf{J}_1 = \frac{1}{\sigma_p}\begin{bmatrix} 1 & (Q/G_{\mathrm{ref}})\,\mathrm{e}^{-g} \end{bmatrix}, \qquad \mathbf{n} = \begin{bmatrix} (Q/G_{\mathrm{ref}})\,\mathrm{e}^{-g} \\ -1 \end{bmatrix}, \qquad \mathbf{J}_1\,\mathbf{n} = 0. \tag{1.17}$$

A step along $\mathbf{n}$ leaves the physics row unchanged to first order, so $\mathbf{n}$ is the tangent to the zero set of the row, a change of $(Q/G_{\mathrm{ref}})\,\mathrm{e}^{-g}$ kelvin of temperature per unit of g with the opposite sign. The information the physics row contributes, $\mathbf{J}_1^{\mathsf{T}}\mathbf{J}_1$, is a matrix of rank one: it grows as $1/\sigma_p^2$ across the curve and is zero along $\mathbf{n}$. Along the curve only the reading row and the prior row inform the state, and by the pattern (1.9) each acts on one entry, with the contributions $\mathbf{J}_2^{\mathsf{T}}\mathbf{J}_2 = \mathrm{diag}(1/\sigma_T^2,\ 0)$ and $\mathbf{J}_3^{\mathsf{T}}\mathbf{J}_3 = \mathrm{diag}(0,\ 1/\sigma_g^2)$. With $n_1 = (Q/G_{\mathrm{ref}})\,\mathrm{e}^{-\hat{g}}$ the first entry of $\mathbf{n}$ at the MAP state and r_2 the reading row there, the physics row's own curvature adds the term $\mathbf{C} = \mathrm{diag}(0,\ n_1 r_2/\sigma_T)$, derived above. With $\mathbf{u} = \mathbf{n}/\|\mathbf{n}\|$ the unit tangent, the information along the curve is

$$\lambda_{\|} = \mathbf{u}^{\mathsf{T}}(\mathbf{J}_2^{\mathsf{T}}\mathbf{J}_2 + \mathbf{J}_3^{\mathsf{T}}\mathbf{J}_3 + \mathbf{C})\mathbf{u} = \frac{n_1^2/\sigma_T^2 + 1/\sigma_g^2 + n_1 r_2/\sigma_T}{n_1^2 + 1}, \tag{1.18}$$

which does not depend on σ_p. The tangent $\mathbf{u}$ is orthogonal to $\mathbf{J}_1$, so in the orthonormal basis $(\mathbf{J}_1^{\mathsf{T}}/\|\mathbf{J}_1\|,\ \mathbf{u})$ the curvature of the objective has the entry $\lambda_{\|}$ along $\mathbf{u}$, an entry across the curve that grows as $1/\sigma_p^2$, and a coupling that stays bounded; its inverse, by the Schur complement of Chapter 3, leaves $1/\lambda_{\|}$ along $\mathbf{u}$ and nothing else as $\sigma_p \to 0$. The covariance therefore tends to the rank-one matrix $\boldsymbol{\Sigma} \to \mathbf{u}\,\mathbf{u}^{\mathsf{T}}/\lambda_{\|}$: the posterior collapses onto a segment of the tangent, a *ridge* of the objective. Then $\sqrt{\Sigma_{11}} \to |u_1|/\sqrt{\lambda_{\|}}$ and $\sqrt{\Sigma_{22}} \to |u_2|/\sqrt{\lambda_{\|}}$ are one spread along the curve seen through the two components of the tangent, and the correlation $\Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$ tends to the sign of $u_1 u_2$, which is -1. The two entries are not known separately: a hall slightly warmer than $\hat{T}$ is a hall with a slightly lower conductance, and the posterior allows the one only together with the other. The reading row has a zero in the column of g, and it narrows the conductance all the same, through the first component of $\mathbf{u}$.

> **Principle 1.2.** *Physical rows constrain the system; reading rows provide evidence about its unknown quantities. The posterior is where the two agree, weighted by the prior, and neither kind of row alone determines it. An exact physical row leaves the posterior only along its own zero set, and its signature in* $\boldsymbol{\Sigma}$ *is a correlation near* -1 *or* $+1$.

The exact posterior along the curve. In the limit $\sigma_p \to 0$ the posterior lies on the physics curve, and along the curve it can be computed without approximation. On the curve the temperature is a function of the conductance entry, $T(g) = T_c + (Q/G_{\mathrm{ref}})\,\mathrm{e}^{-g}$. The rows define the posterior relative to a base flat in (T, g). Integrated across the curve, the physics row contributes a factor that does not depend on g, because the row is written in kelvin and its slope in T, $1/\sigma_p$, does not depend on g; the base is therefore flat in g along the curve, and what remains is the density of g along the curve, the reading row and the prior row evaluated on it:

$$p(g \mid \mathbf{y}) \propto \exp\left[-\frac{1}{2}\left(\frac{T(g) - y_T}{\sigma_T}\right)^2 - \frac{1}{2}\left(\frac{g - g_0}{\sigma_g}\right)^2\right]. \tag{1.19}$$

One-dimensional quadrature of this density over g gives every probability and moment of a quantity derived from the state exactly, for T through $T(g)$ and for G through $G_{\mathrm{ref}}\,\mathrm{e}^{g}$.

The answer read off. An answer that is an entry of the state is read from $\hat{\mathbf{z}}$ and $\boldsymbol{\Sigma}$ directly: the entry $\hat{z}_i$ of $\hat{\mathbf{z}}$, with the spread $\sqrt{\Sigma_{ii}}$. The relation between two entries is the correlation $\Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$, and the directions of the posterior are the eigenvectors of $\boldsymbol{\Sigma}$. A derived quantity f is carried by its gradient, with the spread $\sqrt{\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}}$, and a threshold on an entry is a normal tail of its marginal. Section 1.2.2 computes $\hat{\mathbf{z}}$, $\mathbf{J}$ and $\boldsymbol{\Sigma}$ for the hall.

1.2.2 The posterior of the hall

The hall's state estimation, the fit of the state to the three rows, has as inputs the pinned values $T_c = 18°\mathrm{C}$, $Q = Q_1 = 30\,\mathrm{kW}$ and $G_{\mathrm{ref}} = 6\,\mathrm{kW/K}$; the reading $y_T = 23.2°\mathrm{C}$ with $\sigma_T = 0.5\,\mathrm{K}$; the handbook prior $g_0 = 0.2231$ with $\sigma_g = 0.3$; and the tolerance $\sigma_p = 0.01\,\mathrm{K}$ of the physics row.

The first step. At the reference state $\mathbf{z}^0 = (23°\mathrm{C},\ 0)^{\mathsf{T}}$ the Jacobian (1.12) is

$$\mathbf{J}(\mathbf{z}^0) = \begin{bmatrix} 100 & 500 \\ 2 & 0 \\ 0 & 3.333 \end{bmatrix} :$$

a one-kelvin change in T moves the physics row by 100 and the reading row by 2, and a unit change in g moves the physics row by 500 and the prior row by 3.33. The first row is the pair of kilowatt sensitivities $(G,\ Q) = (6,\ 30)$ divided by $G = 6\,\mathrm{kW/K}$ and by σ_p. The residual vector there is $\mathbf{r} = (0,\ -0.400,\ -0.744)$: the physics holds, the temperature lies 0.4 standard deviations below the reading, and g lies 0.74 standard deviations below the handbook value. The two sides of the normal equations (1.14) are

$$\mathbf{J}(\mathbf{z}^0)^{\mathsf{T}}\mathbf{J}(\mathbf{z}^0) = \begin{bmatrix} 10004 & 50000 \\ 50000 & 250011 \end{bmatrix}, \qquad -\mathbf{J}(\mathbf{z}^0)^{\mathsf{T}}\mathbf{r}(\mathbf{z}^0) = \begin{bmatrix} 0.800 \\ 2.479 \end{bmatrix},$$

and the first step is their solution,

$$\Delta\mathbf{z} = \begin{bmatrix} 0.0684 \\ -0.0137 \end{bmatrix}.$$

The iterations. Table 1.4 lists the iterations. Each row is one iterate $\mathbf{z}^k$; the columns give its two entries, its three rows and the objective of equation (1.11).

Table 1.4: Gauss–Newton iterations for the posterior of the equipment hall (equations (1.10) to (1.14)), with $\sigma_p = 0.01\,\mathrm{K}$, from the reference state $(23°\mathrm{C},\ 0)$.

k	T [°C]	g [–]	r_1 [–]	r_2 [–]	r_3 [–]	$\mathcal{L}$ [–]
0	23.0000	0.0000	0.000	−0.400	−0.744	0.3566
1	23.0684	−0.0137	−0.042	−0.263	−0.789	0.3471
2	23.0701	−0.0139	0.005	−0.260	−0.790	0.3460
3	23.0701	−0.0139	0.005	−0.260	−0.790	0.3460

The first step does almost all the work, because over a step this short the rows are nearly linear. It leaves the physics row slightly violated, $r_1 = -0.042$, through the curvature of e^{-g} that the tangent ignores; the second step restores it, and the third changes nothing at the printed precision. The MAP state is

$$\hat{\mathbf{z}} = \begin{bmatrix} \hat{T} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 23.070°\mathrm{C} \\ -0.0139 \end{bmatrix}, \qquad \hat{G} = 6\,\mathrm{e}^{-0.0139} = 5.917\,\mathrm{kW/K}. \tag{1.20}$$

There the physics holds to $r_1\,\sigma_p = 5\times10^{-5}\,\mathrm{K}$, and the reading and the prior share the disagreement: the temperature lies 0.26 standard deviations below the reading and g lies 0.79 standard deviations below the handbook value.

The information matrix. At the MAP state the Jacobian (1.12) is

$$\mathbf{J}(\hat{\mathbf{z}}) = \begin{bmatrix} 100 & 507.0 \\ 2 & 0 \\ 0 & 3.333 \end{bmatrix}, \qquad \frac{(Q/G_{\mathrm{ref}})\,\mathrm{e}^{-\hat{g}}}{\sigma_p} = \frac{5\,\mathrm{e}^{0.0139}}{0.01} = 507.0,$$

and the information matrix is the sum of the three outer products,

$$\mathbf{J}^{\mathsf{T}}\mathbf{J} = \underbrace{\begin{bmatrix} 10000 & 50701 \\ 50701 & 257056 \end{bmatrix}}_{\text{physics}} + \underbrace{\begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}}_{\text{reading}} + \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & 11.11 \end{bmatrix}}_{\text{prior}} = \begin{bmatrix} 10004 & 50701 \\ 50701 & 257067 \end{bmatrix}.$$

The physics term has rank one: all its information lies along its own row (100, 507.0), across the physics curve, and none along the curve; only the reading's 4 and the prior's 11.11 inform along it. The eigenvalues of $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ are 267067 across the curve and 4.27 along it, with the eigenvector $(0.9811,\ -0.1935)^{\mathsf{T}}$ of the small one pointing along the curve; the first two terms of equation (1.18) give it directly, $(5.070^2\times4+11.11)/(5.070^2+1) = 4.27$. The Gauss–Newton covariance is the inverse,

$$\mathbf{\Sigma}_{\mathrm{GN}} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1} = \begin{bmatrix} 0.2256 & -0.0445 \\ -0.0445 & 0.0088 \end{bmatrix},$$

with the standard deviations $\sqrt{0.2256} = 0.475\,\mathrm{K}$ and $\sqrt{0.0088} = 0.094$.

The curvature term and the Laplace covariance. The second derivative of the physics row at the MAP state is $\partial^2 r_1/\partial g^2 = -507.0$; multiplied by $r_1 = 0.0052$, it adds -2.63 to the g–g entry, $257067 \to 257064$, the same value as $(r_2/\sigma_T)(Q/G_{\mathrm{ref}})\,\mathrm{e}^{-\hat{g}} = (-0.2598/0.5)\times5.0701$. Projected on the along-curve eigenvector $(0.9811,\ -0.1935)^{\mathsf{T}}$, the term is $-2.63\times0.1935^2 = -0.099$ against the along-curve information 4.27, a loss of 2.3 % in the one direction in which the information is small; it raises the variances by 2.4 % and the standard deviations by 1.2 %. The Laplace covariance (1.16) is

$$\mathbf{\Sigma} = \begin{bmatrix} 0.23096 & -0.045552 \\ -0.045552 & 0.0089881 \end{bmatrix}, \qquad \sqrt{\Sigma_{11}} = 0.4806\,\mathrm{K}, \quad \sqrt{\Sigma_{22}} = 0.09481. \tag{1.21}$$

These are the Laplace values used in the rest of the chapter, quoted as 0.481 K and 0.0948; Gauss–Newton alone gives 0.475 K and 0.094. The correlation of the two entries, computed from the unrounded $\mathbf{\Sigma}$, is

$$\frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\,\Sigma_{22}}} = \frac{-0.045552}{0.48058\times0.094806} = -0.9998.$$

The ridge. The physics row of $\mathbf{J}(\hat{\mathbf{z}})$, (100, 507.0), is annihilated by $\mathbf{n} = (5.070,\ -1)^{\mathsf{T}}$, equation (1.17) at $\hat{g}$, and the long axis of $\mathbf{\Sigma}$ is $\mathbf{u} = (0.9811,\ -0.1935)^{\mathsf{T}}$, the same direction normalized. By equation (1.18),

$$\lambda_{\parallel} = \frac{5.070^2\times4 + 11.11 + 5.070\times(-0.260)/0.5}{5.070^2+1} = 4.168,$$

$$\frac{0.9811}{\sqrt{4.168}} = 0.481\,\mathrm{K}, \quad \frac{0.1935}{\sqrt{4.168}} = 0.0948,$$

the square roots of the two diagonal entries of equation (1.21).

Figure 1.5 shows the geometry. The horizontal axis is the log-conductance g and the vertical axis the hall temperature T, the two entries of the state, so that every point of the plane is a candidate state. The green curve is the zero set of the physics row: the states the physics allows. The gray band is the set of temperatures within one standard deviation of the reading, $23.2 \pm 0.5°\mathrm{C}$, and the dashed line inside it the reading itself. Points are labeled by their plane coordinates (g, T), horizontal first. The three black dots are the reference state, $(g, T) = (0, 23)$; the state on the curve that matches the reading exactly, $(g, T) = (-0.039, 23.2)$; and, between them, the MAP state of equation (1.20), $(g, T) = (-0.014, 23.07)$. The purple segment is the one-standard-deviation ellipse of the covariance of equation (1.21), centered on the MAP state; it is drawn as a thick line because its half-width across the curve, about 0.002 in g, is below the resolution of the figure.

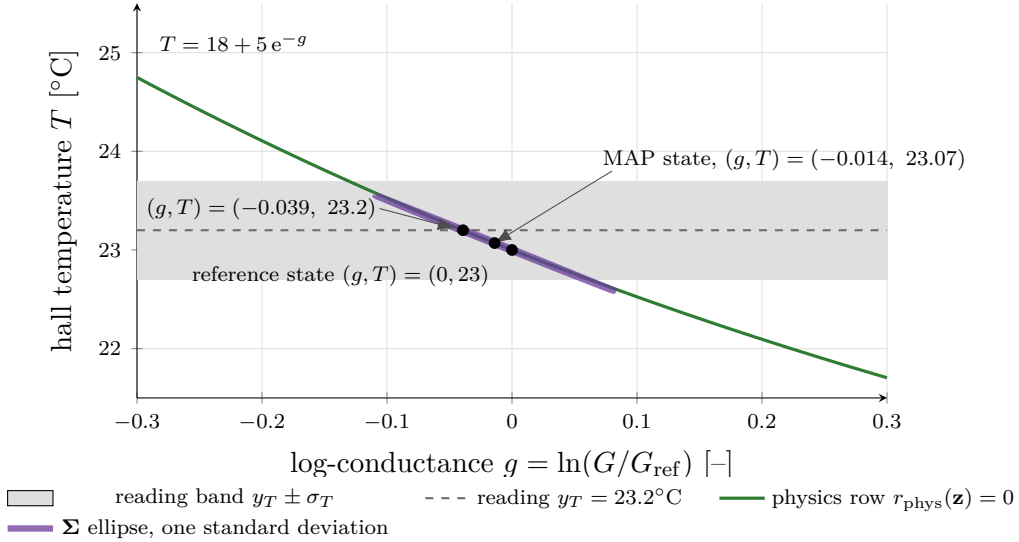


Figure 1.5: The state plane of the equipment hall with $T_c = 18°\mathrm{C}$, $Q = 30\,\mathrm{kW}$ and $G_{\mathrm{ref}} = 6\,\mathrm{kW/K}$. Green: the zero set of the physics row, $T = 18 + 5\,\mathrm{e}^{-g}$ (equation (1.4)). Gray: the reading $y_T = 23.2°\mathrm{C}$ with its band of one standard deviation, $\sigma_T = 0.5\,\mathrm{K}$ (equation (1.6)). The plausible states lie on the arc of the curve inside the band. Purple: the one-standard-deviation ellipse of the Laplace covariance (equation (1.21)) around the MAP state, $(g, T) = (-0.014,\ 23.07)$, collapsed onto the curve by the physics row. Points are labeled (g, T), horizontal coordinate first.

The plausible states lie on the arc of the curve inside the band, from $g = -0.131$ to $g = 0.062$ ($G = 5.26$ to $6.38\,\mathrm{kW/K}$), and the MAP state sits on it between the reading-match point and the reference state, pulled toward the handbook value $g_0 = 0.2231$ by the prior row.

The pair $(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$ of equations (1.20) and (1.21) is what step 4 uses: every answer of the chapter is computed from it.

1.3 The answers (step 4)

The first four questions of Table 1.2 take two general forms, written in the book's notation: $p(\mathbf{z} \mid \mathbf{y})$ is the posterior density of the state, $f(\mathbf{z})$ a quantity derived from the state and f^* a threshold on it. The next-measurement question, a design question about a row not yet attached, is derived in Section 1.6.1. Each is answered from the posterior of Section 1.2: its

planner's question, the general form of its category, the form it takes for the hall, the quantities of the posterior it uses, and the hall's numbers. The questions name the limit of 24°C at the present load and the limit of 27.5°C at $Q_2 = 60\,\mathrm{kW}$.

1.3.1 The hall temperature and the conductance

Planner's questions. What is the hall air temperature at today's load? What is the effective conductance of the air path? The first asks for the temperature of the hall at the present load, given its one reading and the handbook value; the second for the conductance of its air path, the calibrated value the design engineer compares with the handbook value.

General form. An estimation question about a quantity $f(\mathbf{z})$ derived from the state asks for its posterior distribution, the probability that the quantity lies at or below each value f^* (the value at which it equals one half is the median),

$$\Pr(f(\mathbf{z}) \le f^* \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) \le f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \tag{1.22}$$

where $\mathbf{1}[\cdot]$ is one when the event holds and zero otherwise, summarized by a location and a spread,

$$\mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}] = \int f(\mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \qquad \mathrm{Var}(f(\mathbf{z}) \mid \mathbf{y}) = \mathbb{E}\big[(f(\mathbf{z}) - \mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}])^2 \mid \mathbf{y}\big]. \tag{1.23}$$

An entry of the state is the case $f(\mathbf{z}) = z_i$.

Specialized form. The two questions are equations (1.22) and (1.23) with the state $\mathbf{z} = (T, g)^\mathsf{T}$ of the hall, the derived quantity $f(\mathbf{z}) = z_1 = T$ for the temperature and $f(\mathbf{z}) = z_2 = g$, or $f(\mathbf{z}) = G = G_{\mathrm{ref}}\, \mathrm{e}^{z_2}$, for the conductance, and $p(\mathbf{z} \mid \mathbf{y})$ the posterior of step 3, the Laplace normal $\mathbf{z} \mid \mathbf{y} \approx \mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$ of equations (1.20) and (1.21).

The temperature. Here $f(\mathbf{z}) = z_1 = T$, the first entry of the state. Under the Laplace posterior the marginal of an entry of a normal vector is normal, with the entry's mean and the diagonal entry of the covariance, so

$$T \mid \mathbf{y} \approx \mathcal{N}(\hat{z}_1, \Sigma_{11}),$$

and the answer is read off without any transformation: the first entry of $\hat{\mathbf{z}}$, with the spread $\sqrt{\Sigma_{11}}$.

The conductance. The second entry $z_2 = g$ is read off in the same way, $g \mid \mathbf{y} \approx \mathcal{N}(\hat{z}_2, \Sigma_{22})$. The conductance itself, $G = G_{\mathrm{ref}}\, \mathrm{e}^{z_2}$, is a derived quantity of the state. Linearized at the MAP state, its variance is $\mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g}$ with the gradient

$$\mathbf{g} = \frac{\partial G}{\partial \mathbf{z}} = \begin{bmatrix} \partial G/\partial T \\ \partial G/\partial g \end{bmatrix}_{\hat{\mathbf{z}}} = \begin{bmatrix} 0 \\ G_{\mathrm{ref}}\, \mathrm{e}^{\hat{g}} \end{bmatrix} = \begin{bmatrix} 0 \\ \hat{G} \end{bmatrix}, \qquad \sigma_G^2 = \mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g} = \hat{G}^2\, \Sigma_{22}, \tag{1.24}$$

the delta-method spread, written σ_G. The ratio $\sigma_G/\hat{G} = \sqrt{\Sigma_{22}}$: the spread of the log-conductance is the fractional spread of the conductance, as Section 1.1.4 anticipated.

The value beside each comes from the exact posterior: the moments (1.23) under the density (1.19), evaluated by one-dimensional quadrature over g, with T through $T(g) = T_c + (Q/G_{\mathrm{ref}})\, \mathrm{e}^{-g}$ and G through $G_{\mathrm{ref}}\, \mathrm{e}^{g}$, and the median of each, the value f^* at which equation (1.22) equals one half. The Laplace location, the entry of $\hat{\mathbf{z}}$, is the median of the Laplace normal, and it is set beside the exact median, like with like; the exact mode of g is $\hat{g}$ itself, and the exact mean is given as well, labeled as a mean. They differ by the skew of the posterior along the curve.

Building blocks. The MAP state $\hat{\mathbf{z}} = (23.070\,^{\circ}\mathrm{C},\ -0.0139)^{\mathsf{T}}$ and the Laplace covariance $\boldsymbol{\Sigma}$, curvature term included, with $\sqrt{\Sigma_{11}} = 0.4806\,\mathrm{K}$ and $\sqrt{\Sigma_{22}} = 0.09481$ (equations (1.20) and (1.21)); the gradient (1.24); and, for the exact values beside them, the density (1.19) and a one-dimensional quadrature over g.

The hall temperature now. The hall temperature is the first entry of the state: $\hat{T} = 23.070\,^{\circ}\mathrm{C}$ with $\sqrt{\Sigma_{11}} = 0.481\,\mathrm{K}$ (Gauss–Newton alone, 0.475 K). The value beside it comes from the exact posterior: the median of T along the curve is 23.028°C, 0.042 K below $\hat{T}$ (the mean is 23.030°C), with the standard deviation 0.4813 K, 0.0007 K wider.

The effective conductance. The conductance entry is $\hat{g} = -0.0139$ with $\sqrt{\Sigma_{22}} = 0.0948$. By equation (1.24), with $\mathbf{g} = (0,\ 5.917)^{\mathsf{T}}$,

$$\hat{G} = 5.917\,\mathrm{kW/K}, \qquad \sigma_G = \sqrt{\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\,\mathbf{g}} = 5.917\,\sqrt{0.0089881} = 0.561\,\mathrm{kW/K},$$

a band of 9.5 % of its value, against a factor $\mathrm{e}^{0.3} = 1.35$ in the handbook before the reading. The values beside them come from the exact posterior: the conductance entry has the median -0.0056, 0.0083 above $\hat{g}$ (the mean is -0.0013), and the standard deviation 0.0967, 0.0019 wider; the conductance has the median 5.967 kW/K, 0.049 kW/K above $\hat{G}$ (the mean is 6.020 kW/K), and the standard deviation 0.591 kW/K, 0.030 kW/K or 5 % wider. The exact posterior extends further toward large conductances than the normal at $\hat{\mathbf{z}}$ does, the skew that the design-load question below meets in its tail.

1.3.2 The limits at the present load and at 60 kilowatts

Planner's questions. Is the hall below 24°C now? Will the hall carry its 60 kW design load inside 27.5°C? The first asks about the hall as it runs today. The second is the design review's question, which for the planner reads

Will the hall carry its 60 kW *design load inside the* 27.5°C *limit, and how sure is that?*

Its unknown is the temperature T_2 of the scenario node at the design load of 60 kW, and over the state of the graph it is

$$\Pr(T_2 < 27.5\,^{\circ}\mathrm{C} \mid \mathbf{y}), \tag{1.25}$$

with the unknown, the design-load temperature, on the left of the bar and the evidence, the reading of the present condition and the handbook value, on the right; the graph and its pinned inputs are conditions of the model throughout.

General form. A prediction under a threshold is a tail probability of a derived quantity under the posterior, the posterior mass of the states that satisfy the event,

$$\Pr(f(\mathbf{z}) < f^{*} \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) < f^{*}\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}. \tag{1.26}$$

Specialized form. The two questions are equation (1.26) with the state $\mathbf{z} = (T, g)^{\mathsf{T}}$ of the hall, $p(\mathbf{z} \mid \mathbf{y})$ the posterior of step 3, the Laplace normal $\mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$ of equations (1.20) and (1.21), and the derived quantity $f(\mathbf{z})$ and the threshold f^{*} given for each below: $f(\mathbf{z}) = T$ with $f^{*} = 24\,^{\circ}\mathrm{C}$, and $f(\mathbf{z}) = T_2$ with $f^{*} = T_{\max} = 27.5\,^{\circ}\mathrm{C}$. Two choices enter here. The design-load condition is taken to share the conductance of the present condition, with the load and the supply air pinned at both: this is what makes a condition the hall has never carried predictable from a reading at 30 kW, and it fails if the conductance changes with the load, for instance through an airflow that rises with it (Section 1.1.4), or if the unit saturates and can no longer

hold T_c at 60 kW (Section 1.6.5), and no check of this chapter can test either because nothing has been observed at 60 kW. The design-load answer is the steady temperature at 60 kW: this is what lets the event be written on the conductance alone, and it says nothing about the period just after the load rises, while the hall's stored heat still matters (Section 1.6.5).

The present load. Here $f(\mathbf{z}) = z_1 = T$ and $f^* = 24°\mathrm{C}$. Under the Laplace posterior the tail is a normal tail of the first entry,

$$\Pr(T < 24°\mathrm{C} \mid \mathbf{y}) \approx \Phi\!\left(\frac{24 - \hat{z}_1}{\sqrt{\Sigma_{11}}}\right), \tag{1.27}$$

where Φ is the standard normal cumulative distribution function. The value beside it comes from the exact posterior: on the curve the event $T < 24°\mathrm{C}$ is the event $T(g) < 24°\mathrm{C}$, that is $g > -\ln\!\big((24 - T_c)\,G_{\mathrm{ref}}/Q\big) = -\ln 1.2$, and its probability is the integral of equation (1.19) above that value.

The design load. Here $f(\mathbf{z}) = T_2$, the temperature of the scenario node. By equation (1.8) it is a function of the second entry alone, $T_2 = T_c + (Q_2/G_{\mathrm{ref}})\,\mathrm{e}^{-z_2}$, and on the zero set of the physics row it is also a function of the first: dividing equation (1.8) by equation (1.7) removes the conductance,

$$\frac{T_2 - T_c}{T - T_c} = \frac{Q_2}{Q_1} \qquad \Longrightarrow \qquad T_2 = T_c + \frac{Q_2}{Q_1}\,(T - T_c). \tag{1.28}$$

At twice the load the excess over the supply air is twice as large, whatever the conductance: a present temperature of 23°C predicts $T_2 = 18 + 2 \times 5 = 28°\mathrm{C}$, and a present temperature of 23.2°C predicts 28.4°C. Readings taken during the first operating condition therefore inform the second, although the hall has never carried its design load; this is how a plant model links a measurement of the hall as it runs today to predictions under operating conditions that have not yet occurred. In the vector form the link is the tangent $\mathbf{n}$ of equation (1.17): a step along the curve moves T by $(Q_1/G_{\mathrm{ref}})\,\mathrm{e}^{-g}$ and T_2 by $(Q_2/G_{\mathrm{ref}})\,\mathrm{e}^{-g}$ per unit of $-g$, in the ratio Q_2/Q_1.

The event itself needs no posterior to be rewritten. The design-load temperature $T_2 = T_c + Q_2/G$ decreases in G, so for a limit $T_{\max} > T_c$ the event $T_2 < T_{\max}$ is, in turn, an event on the conductance, on the conductance entry and, by equation (1.28), on the present temperature:

$$T_2 < T_{\max} \iff G > G^* \iff g > g^* = \ln\frac{G^*}{G_{\mathrm{ref}}} \iff T < T^*, \tag{1.29}$$

with the thresholds

$$G^* = \frac{Q_2}{T_{\max} - T_c}, \tag{1.30}$$

$$T^* = T_c + \frac{Q_1}{Q_2}\,(T_{\max} - T_c). \tag{1.31}$$

For $T_{\max} \le T_c$ the probability is zero, since $T_2 > T_c$ for every positive conductance. The inequalities of equation (1.29) describe the same set of states on the physics curve, so their probabilities are equal under any distribution on the curve; the equivalence holds under the assumed steady-state model, with known load and supply-air temperature. The question about a load the hall has never carried has become a question about the unknown conductance, an entry of the state that the present condition informs, before any data are processed.

Figure 1.6 draws the equivalence (1.29) on the forward relation (1.2). The horizontal axis is the conductance G in kW/K and the vertical axis the hall temperature in °C. The solid blue curve is the hall at the load of 30 kW, the dashed green curve the same hall at 60 kW; the dot on the blue curve is the forward calculation of Section 1.1.4, $T = 23°\mathrm{C}$ at $G = 6\,\mathrm{kW/K}$, and the dot above it the same conductance at 60 kW, 28°C. The red line is the limit of 27.5°C, and the dotted vertical line marks the threshold G^* of equation (1.30), where the dashed curve crosses

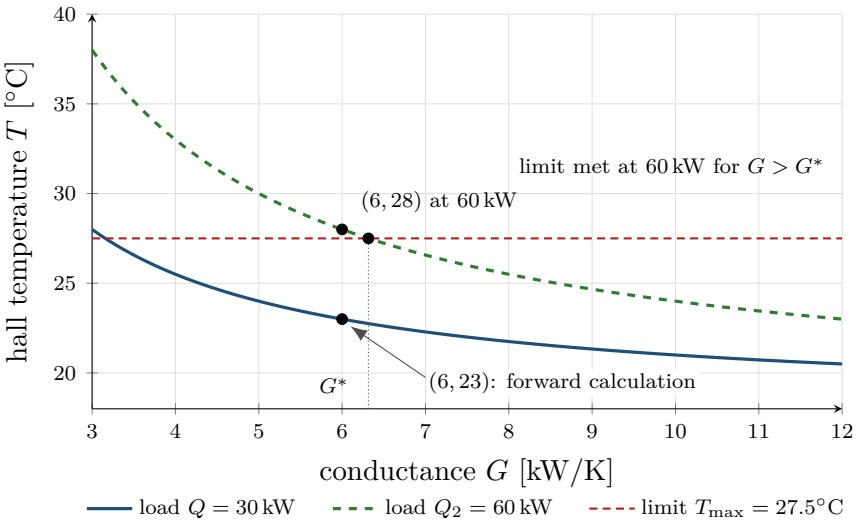


Figure 1.6: The forward relation $T = T_c + Q/G$ (equation (1.2)) with $T_c = 18$°C, for the loads $Q = 30$ kW and $Q_2 = 60$ kW. Dots: the forward calculation $T = 23$°C at $G = 6$ kW/K; the same conductance at 60 kW, 28°C; and the threshold $G^* = Q_2/(T_{\max} - T_c)$ of equation (1.30) at which the 60 kW hall reaches the limit of 27.5°C.

the limit. The dashed green curve lies below the red line exactly to the right of G^*: because the curve is strictly decreasing, the event "below the limit" on the vertical axis is the event "above G^*" on the horizontal axis.

Under the Laplace posterior the probability is a normal tail of the second entry, the entry onto which the physics maps the event,

$$\Pr(T_2 < T_{\max} \mid \mathbf{y}) = \Pr(g > g^* \mid \mathbf{y}) \approx \Phi\left(\frac{\hat{z}_2 - g^*}{\sqrt{\Sigma_{22}}}\right). \tag{1.32}$$

The same event written on the first entry, $T < T^*$, has the same exact probability, but under the Laplace normal, whose ridge is the straight tangent rather than the curve, its tail differs slightly; the chapter reports the tail on g, the entry on which the design-load temperature depends alone. The value beside it comes from the exact posterior: the probability is the integral of equation (1.19) above g^*.

> **Principle 1.3.** *When a quantity in a question is a monotone function of one entry of the state, a threshold on the quantity is a threshold on that entry, and the two events have the same probability under every posterior. The transformation uses the physics only, and it can be made before any data are processed.*

Building blocks. The entries $\hat{z}_1 = 23.070$°C and $\hat{z}_2 = -0.0139$ of the MAP state and the standard deviations $\sqrt{\Sigma_{11}} = 0.481$ K and $\sqrt{\Sigma_{22}} = 0.0948$ of the covariance (equations (1.20) and (1.21)); the threshold of equation (1.30) for the design limit $T_{\max} = 27.5$°C and $Q_2 = 60$ kW, and for other limits and loads; the normal cumulative distribution Φ; and, for the exact values beside them, the density (1.19) integrated over g above each threshold.

The hall below 24 degrees. By equation (1.27),

$$\Pr(T < 24\text{°C} \mid \mathbf{y}) \approx \Phi\left(\frac{24 - 23.070}{0.481}\right) = \Phi(1.93) = 0.973.$$

The value beside it comes from the exact posterior: the mass of equation (1.19) above $g = -\ln 1.2 = -0.1823$ is 0.977, 0.004 higher.

The hall at 60 kilowatts. For the design limit $T_{\max} = 27.5°\mathrm{C}$ at $Q_2 = 60\,\mathrm{kW}$, equations (1.30) and (1.31) give

$$G^* = \frac{60}{27.5 - 18} = \frac{60}{9.5} = 6.316\,\mathrm{kW/K}, \qquad g^* = \ln\frac{6.316}{6} = 0.0513,$$

$$T^* = 18 + \tfrac{1}{2} \times 9.5 = 22.75°\mathrm{C}\,:$$

the temperature at 60 kW stays below 27.5°C if and only if the temperature at 30 kW is below 22.75°C; on the curve of Figure 1.5, the states with $g > g^*$ are those below 22.75°C. By equation (1.32),

$$\Pr(T_2 < 27.5°\mathrm{C} \mid \mathbf{y}) \approx \Phi\left(\frac{-0.0139 - 0.0513}{0.0948}\right) = \Phi(-0.688) = 0.246,$$

with the MAP conductance 5.917 kW/K to the left of G^* in Figure 1.6. The value beside it comes from the exact posterior: the mass of equation (1.19) above $g^* = 0.0513$ is 0.281, 0.036 higher. The event lies in the tail of g toward large conductances, the side on which the exact posterior extends further than the normal at $\hat{\mathbf{z}}$, so the state route understates it. The gap does not change the decision: on both routes the hall carries its design load inside the limit with probability of about one in four. The temperature at a load the hall has not yet carried is the same fit read at a different input; Chapter 4 names this device.

Table 1.5 evaluates the same blocks for other limits and loads. Each row is a limit $T_{\max}$ and a future load Q_2; the columns give the threshold conductance G^*, the threshold log-conductance g^*, the threshold on the present temperature T^*, the probability of equation (1.32), which the chapter reports, and beside it the exact value, the mass of equation (1.19) above g^*. One row of the arithmetic, for $T_{\max} = 27°\mathrm{C}$ and $Q_2 = 60\,\mathrm{kW}$: $G^* = 60/9 = 6.667\,\mathrm{kW/K}$, $g^* = \ln(6.667/6) = 0.1054$, $T^* = 18 + 0.5 \times 9 = 22.5°\mathrm{C}$, and $\Phi((-0.0139 - 0.1054)/0.0948) = \Phi(-1.258) = 0.104$.

Table 1.5: Equivalent thresholds and the design-load probability, with $T_c = 18°\mathrm{C}$, $Q_1 = 30\,\mathrm{kW}$ and $G_{\mathrm{ref}} = 6\,\mathrm{kW/K}$ (equations (1.30) and (1.31)); the probability by the Laplace tail (1.32), with $\hat{g} = -0.0139$ and $\sqrt{\Sigma_{22}} = 0.0948$, and exactly, by quadrature of equation (1.19). The row 27.5°C at 60 kW is the design-load question; the others are added to show the trend.

Limit $T_{\max}$ [°C]	Load Q_2 [kW]	G^* [kW/K]	g^* [–]	T^* [°C]	Laplace [–]	Exact [–]
26	60	7.500	0.2231	22.00	0.0062	0.0155
27	60	6.667	0.1054	22.50	0.1042	0.1354
27.5	60	6.316	0.0513	22.75	0.2458	0.2815
29	60	5.455	−0.0953	23.50	0.8047	0.8355
27.5	45	4.737	−0.2364	24.33	0.9905	0.9964
27.5	75	7.895	0.2744	21.80	0.0012	0.0049

A stricter limit, or a larger future load, demands a larger conductance and a cooler present temperature; the first five columns follow from the physics alone. The Laplace tail is below the exact one in every row, and the gap grows in ratio toward the far tail: a factor of 2.5 to 4 in the two smallest rows.

1.4 The checks (step 5)

The fit rests on two approximations and one assumption: a physics row with a finite tolerance in place of an exact law, a normal distribution in place of the posterior along a curved ridge, and a reading and a handbook value that are compatible with each other. Each is set against a full calculation, and the weight of the handbook value is measured by removing it.

1.4.1 The exact-physics limit

Table 1.6 repeats the computation at three model tolerances. Each row is one σ_p; the columns are the two entries of $\hat{\mathbf{z}}$, their Laplace standard deviations and their correlation. From 0.1 to 0.01 K the conductance entry moves by 0.001, its spread by 0.002 and the correlation from -0.979 to -0.9998; from 0.01 to 0.001 K nothing moves at the printed precision but the last digit of the correlation. The limit is reached at $\sigma_p = 0.01\,\mathrm{K}$, and the correlation tends to -1 as equation (1.18) requires.

Table 1.6: The posterior of the equipment hall at three model tolerances σ_p of the physics row: MAP state, Laplace standard deviations (with the curvature term) and correlation. The exact-physics limit $\sigma_p \to 0$ is reached at 0.01 K.

σ_p [K]	$\hat{T}$ [°C]	$\hat{g}$ [–]	$\sqrt{\Sigma_{11}}$ [K]	$\sqrt{\Sigma_{22}}$ [–]	correlation [–]
0.1	23.0705	−0.0130	0.4806	0.0966	−0.9790
0.01	23.0701	−0.0139	0.4806	0.0948	−0.9998
0.001	23.0701	−0.0139	0.4806	0.0948	−1.0000

1.4.2 The Laplace posterior against the exact posterior

In the limit $\sigma_p \to 0$ the exact posterior is the density (1.19) along the physics curve, and one-dimensional quadrature of it gives every quantity of the questions; the event $T < 24$°C is the event $g > -\ln 1.2 = -0.1823$ on the curve, and the design-load event is $g > g^* = 0.0513$. Table 1.7 collects the pairs that step 4 computed, the Laplace value it reports beside the exact one. Each row is one quantity of the posterior, with its unit; the third column is its Laplace value, read from $\hat{\mathbf{z}}$ and $\boldsymbol{\Sigma}$ of equations (1.20) and (1.21), and the fourth its exact value from the quadrature. The location of each quantity is its median in both columns, the Laplace median being the value at $\hat{\mathbf{z}}$; the exact mean has a row of its own.

Table 1.7: The Laplace approximation of the posterior of the equipment hall against the exact posterior along the physics curve, in the limit $\sigma_p \to 0$.

Quantity	Unit	Laplace, at $\hat{\mathbf{z}}$	Exact, along the curve
Median of T	°C	23.070	23.028
Mean of T	°C	–	23.030
Standard deviation of T	K	0.4806	0.4813
$\Pr(T < 24\,^\circ\mathrm{C} \mid \mathbf{y})$	–	0.973	0.977
Median of g	–	−0.0139	−0.0056
Mean of g	–	–	−0.0013
Standard deviation of g	–	0.0948	0.0967
Median of G	kW/K	5.917	5.967
Mean of G	kW/K	–	6.020
Standard deviation of G	kW/K	0.561	0.591
$\Pr(T_2 < 27.5\,^\circ\mathrm{C} \mid \mathbf{y})$ at 60 kW	–	0.246	0.281

For the temperature the approximation is adequate: the tail probability differs by 0.004 and the standard deviation by 0.0007 K. The posterior is slightly skewed along the curve, so its mode, median and mean differ: the mode of g is $\hat{g}$, the median lies 0.0083 and the mean 0.013 above it, the median conductance lies 0.049 kW/K and the mean 0.10 kW/K above $\hat{G}$, and the exact standard deviation of G is 0.591 kW/K, 5 % wider than the Laplace 0.561 kW/K.

The skew matters most for the design-load question, whose event lies in the tail of g toward large conductances, the side on which the exact posterior extends further: the Laplace tail 0.246 is low by 0.036 against the exact 0.281. Further into the tail the understatement grows: in Table 1.5 the Laplace tail on g is smaller than the exact one by a factor of 2.5 to 4 in the two smallest rows, 0.0062 against 0.0155 and 0.0012 against 0.0049. Written on the first entry instead, as $T < T^* = 22.75°\mathrm{C}$, the design-load event has the Laplace tail 0.253, because the normal follows the straight tangent rather than the curve; the exact value is 0.281 on either entry. The exact values rest on the base of equation (1.19), flat in g along the curve; with a base flat in T along the curve instead, the design-load value is 0.249, a reminder that an exact tail depends on the base measure as well as on the rows. The conductance band and the design-load probability are therefore slightly narrow and slightly low, and the report says so.

A scalar check. Eliminating T through the physics row and differentiating the remaining one-entry objective in g twice gives the same $\sqrt{\Sigma_{22}} = 0.0948$ and, through the slope of the curve, the same $\sqrt{\Sigma_{11}} = 0.481\,\mathrm{K}$. The vector form reaches them without the substitution, and it is the form that carries over unchanged to the cooling network of Section 1.1.3.

1.4.3 The residual check

Three rows act on two entries, and with the physics exact the reading and the handbook both inform the one direction along the curve: one row is redundant, and the rows at the MAP state measure how well the two agree. They are $r_1 = 0.005$, $r_2 = -0.260$ and $r_3 = -0.790$ (Table 1.4), and their squares sum to $2\mathcal{L}(\hat{\mathbf{z}}) = 0.692$. When the model and the standard deviations of the rows are right, that sum follows a chi-square distribution with as many degrees of freedom as there are rows less entries, here one; *Operations and Control by Derivation*, Chapter 9, derives the result, the chi-square test of bad-data detection. A sum of 0.692 or more then has probability 0.41, while a sum above 3.84 would have probability below 0.05 and would say that the reading or the handbook value is wrong. The two agree, which is a first, coarse look at the question of a biased sensor.

1.4.4 The weight of the handbook prior

The handbook value is the one input of the fit that the hall did not supply, and removing its row shows how much the answers lean on it. Without the prior row the residual vector holds the rows r_1 and r_2 of equation (1.10): two rows, two entries, a square Jacobian, both rows zero at the MAP state ($23.2°\mathrm{C},\ -0.0392$), and $\Sigma_{11} = \sigma_T^2$ whatever σ_p, and $\sqrt{\Sigma_{22}} = 0.0962$ (the last exercise of the chapter derives both). On the conductance entry, the route of equation (1.32), the design-load probability is then

$$\Pr(T_2 < 27.5°\mathrm{C} \mid y_T) \approx \Phi\!\left(\frac{-0.0392 - 0.0513}{0.0962}\right) = \Phi(-0.941) = 0.173,$$

against 0.246 with the handbook row, a rise of 0.073. Exactly, with the same base flat in g along the curve and the prior term removed from equation (1.19), it is 0.212 against 0.281, a rise of 0.069. Figure 1.7 draws the two Laplace marginals of the conductance entry. The horizontal axis is g and the vertical axis the probability density; the dashed curve is the marginal under the reading alone, $\mathcal{N}(-0.0392,\ 0.0962^2)$, and the solid curve the marginal with the handbook row, $\mathcal{N}(-0.0139,\ 0.0948^2)$. The dotted red line is the threshold $g^* = 0.0513$, and the shaded areas to its right are the two probabilities. The handbook value, which places the conductance above what the reading alone implies, moves the marginal toward the threshold with nearly the same width.

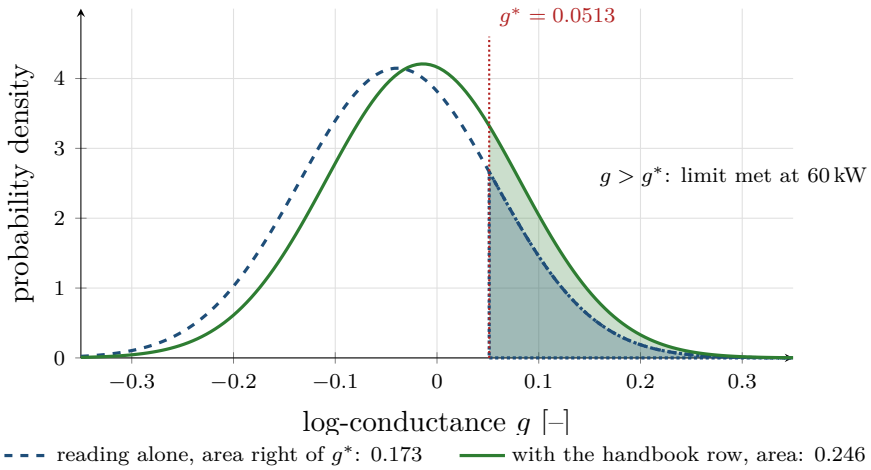


Figure 1.7: The weight of the handbook prior in the design-load answer: Laplace marginals of the conductance entry, $\mathcal{N}(-0.0392,\ 0.0962^2)$ under the reading alone (dashed) and $\mathcal{N}(-0.0139,\ 0.0948^2)$ with the handbook row (solid), densities $\varphi((g-\hat{g})/s)/s$ with φ the standard normal density. The shaded areas right of $g^* = 0.0513$ are the design-load probabilities, 0.173 and 0.246.

The same holds in every row of Table 1.5, on both routes: without the handbook row each probability is lower than with it. The design-load answer therefore depends on the handbook value, and the handbook value is stated with the report.

1.5 The report (step 6)

The record does not support the design sheet's pass. Carried forward, the handbook conductance of 7.5 kW/K put the hall at 26°C at 60 kW, well below the limit of 27.5°C; held against the one reading through the energy balance, it gives way to a conductance of 5.92 kW/K, below the 6.316 kW/K the design load needs, and the hall stays below 27.5°C at 60 kW with probability of about one in four, 0.246 by the Laplace route and 0.281 exactly: neither the comfortable pass of the nameplate forward solve nor the fail of the commissioning reading taken as exact. An airflow and coil test, the measurement that seems to measure the very heat in question, would not change that answer (Section 1.6.2). The answers to the temperature, conductance and limit questions of Table 1.2 are given below in the units in which they were asked, and the sign-off follows from the design-load question.

1.5.1 The answers and their bands

Table 1.8 gives the answers. Each row is one question; the columns give the answer computed from $\hat{\mathbf{z}}$ and $\boldsymbol{\Sigma}$, which the chapter reports, the exact value beside it from the density along the curve, and the band (one standard deviation for the temperature and the conductance, Laplace and exact; for the design-load probability, its value without the handbook row, the check of Section 1.4.4).

The reading has placed the hall temperature to within about half a kelvin, slightly better than the sensor's own 0.5 K because the handbook value adds a little through the physics row. The conductance is known to within 9.5 % of its value, against a factor $e^{0.3} = 1.35$ in the handbook before the reading; the exact posterior widens that band to 0.59 kW/K. The hall is below 24°C with probability of about 0.97. The next-measurement question, whether a measurement before the sign-off would narrow these answers, is answered in Section 1.6.

Table 1.8: The report on the equipment hall from the reading $y_T = 23.2°\mathrm{C}$ and the handbook conductance of 7.5 kW/K, with the physics exact.

Question	Answer (state route)	Exact	Band
What is the hall air temperature at today's load?	$\hat{T} = 23.07°\mathrm{C}$	median 23.03°C	±0.48 K (0.481 K exact)
What is the effective conductance of the air path?	$\hat{G} = 5.92\,\mathrm{kW/K}$	median 5.97 kW/K (mean 6.02 kW/K)	±0.56 kW/K (0.59 kW/K exact)
Is the hall below 24°C now?	$\Pr(T < 24°\mathrm{C} \mid \mathbf{y}) =$ 0.973, about 0.97	0.977; gap 0.004	–
Will the hall carry its 60 kW design load inside 27.5°C?	$\Pr(T_2 < 27.5°\mathrm{C} \mid \mathbf{y}) =$ 0.246, about 0.25	0.281; gap 0.036	without the handbook value 0.173 (0.212 exact)

1.5.2 The design-load decision

The design-load question has the answer of the last row: at 60 kW the hall stays below 27.5°C with probability of about one in four, 0.246 by the Laplace route and 0.281 exactly: on the present cooling the record does not support signing off the 60 kW design load. The answer assumes that the unit still holds the supply air at 18°C at 60 kW; if it cannot, T_c rises with the load and the hall is warmer than this answer says (Section 1.6.5). The reason is read from the equivalent threshold: the limit is met only if the conductance exceeds $G^* = 6.316\,\mathrm{kW/K}$, and the posterior places the conductance at $5.92 \pm 0.56\,\mathrm{kW/K}$; in the conductance entry the threshold $g^* = 0.0513$ lies 0.69 standard deviations above $\hat{g} = -0.0139$. The answer leans on the handbook value: without it the probability would be 0.173 by the Laplace route and 0.212 exactly. A measurement cannot raise the conductance; it can only show where it lies, and with the estimate below the threshold, further evidence is more likely to confirm the shortfall than to remove it. Carrying the 60 kW design load under this limit calls for more cooling in the design: more airflow (a larger conductance) or a colder supply air, if the unit can deliver it (Chapter 14 ranks such design changes); not more measurements.

1.6 The next measurement (step 7)

Whether the bands of Table 1.8 are narrow enough is the planner's decision. If they are not, the next-measurement question of Table 1.2 asks which measurement or test to commission on the existing hall before the design is signed off, to narrow them most, and a measurement not yet taken is a row not yet attached.

1.6.1 The preposterior covariance

The value a candidate measurement will return is unknown, but its Jacobian row at $\hat{\mathbf{z}}$, its sensitivity to each entry divided by its standard deviation, is known before the measurement is taken, and that is enough. The information matrix is a sum of one outer product per row (Section 1.2.1), so a candidate row with the Jacobian row $\mathbf{J}_a$ adds $\mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a$ to the inverse of the present covariance, and the covariance expected after the measurement is

$$\boldsymbol{\Sigma}_a = \left(\boldsymbol{\Sigma}^{-1} + \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a\right)^{-1}, \tag{1.33}$$

with $\boldsymbol{\Sigma}$ of equation (1.21), curvature term included. The value the measurement will return does not appear in it. Chapter 2 derives $\boldsymbol{\Sigma}_a$ as the preposterior covariance, and Chapter 18 ranks

candidate tests by it, before the build is committed, for quantities derived from the state. The band on the temperature after the measurement is $\sqrt{\Sigma_{a,11}}$, and the band on the conductance, the quantity on which the design-load answer rests, is $\sigma_{G,a} = \sqrt{\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}_a\mathbf{g}}$ with the gradient of equation (1.24).

1.6.2 The candidate measurements

Four candidate measurements on the existing hall are considered, each written as a row over $\mathbf{z} = (T, g)^{\mathsf{T}}$:

- a second hall sensor of the same accuracy, with an error independent of the first: the row $(T - y'_T)/\sigma_T$, with the Jacobian row $\mathbf{J}_a = (2,\ 0)$, the same as the present reading row;
- a reference sensor of 0.2 K in the hall: $\mathbf{J}_a = (5,\ 0)$;
- a reading of the supply-air temperature, or a load meter on the servers: T_c and Q are pinned, not entries of the state, so the row has no entry in either column, $\mathbf{J}_a = (0,\ 0)$;
- an airflow and coil test, which measures the heat the unit removes (the airflow through the unit times the temperature drop across its coil), $\sigma_q = 1\,\text{kW}$ on the heat removed $G\,(T - T_c)$: $\mathbf{J}_a = (\hat{G},\ \hat{G}\,(\hat{T} - T_c))/\sigma_q = (5.917,\ 30.00)$.

Table 1.9 evaluates equation (1.33) for each. Each row is one candidate; the columns give its Jacobian row, the standard deviation of T after it, $\sqrt{\Sigma_{a,11}}$, and the width it removes from the present 0.481 K, then the conductance band after it, $\sigma_{G,a}$ with the gradient $\mathbf{g} = (0,\ 5.917)^{\mathsf{T}}$, and the width it removes from the present 0.561 kW/K. One row of the arithmetic, for the second sensor: $\mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a$ has the single entry 4, in its T–T position, and

$$\boldsymbol{\Sigma}_a = \begin{bmatrix} 0.12005 & -0.023678 \\ -0.023678 & 0.0046738 \end{bmatrix}, \qquad \sqrt{0.12005} = 0.346\,\text{K}, \quad 5.917\,\sqrt{0.0046738} = 0.405\,\text{kW/K}.$$

Table 1.9: Candidate measurements for the equipment hall ranked by the preposterior covariance of equation (1.33), evaluated at $\hat{\mathbf{z}}$. Present bands: 0.481 K on T and 0.561 kW/K on G.

Candidate	$\mathbf{J}_a$ [K^{-1}, –]	$\sqrt{\Sigma_{a,11}}$ [K]	removed [K]	$\sigma_{G,a}$ [kW/K]	removed [kW/K]
Second hall sensor, 0.5 K	(2, 0)	0.346	0.134	0.405	0.156
Reference hall sensor, 0.2 K	(5, 0)	0.185	0.296	0.216	0.345
Supply-air reading or load meter	(0, 0)	0.481	0	0.561	0
Airflow and coil test, 1 kW	(5.917, 30.00)	0.481	$< 10^{-6}$	0.561	$< 10^{-6}$

Only the two sensors in the hall remove anything, and the reason is visible in their Jacobian rows. The posterior is wide in one direction only, along the physics curve, $(0.9811,\ -0.1935)^{\mathsf{T}}$, where the present information is 4.17: the 4.27 of $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ less the 0.099 of the curvature term. The second sensor's row projects on that direction as $2 \times 0.9811 = 1.962$ and adds $1.962^2 = 3.85$, nearly doubling the information, so that the variances along the curve fall by the factor $4.17/8.02 = 0.52$ and both standard deviations by the factor 0.72. The supply-air reading and the load meter read pinned quantities: in this model they are known already, their rows are zero, and they would inform only a graph in which T_c or Q were free. The airflow and coil test is the subtler case. Its row (5.917, 30.00) is proportional to the physics row (100, 507.0), both in the ratio 5.070, so it points across the curve, where the physics already fixes the state to within σ_p; it measures again what the energy balance asserts, and it adds nothing along the

curve. This row is the product alone; an airflow reading by itself would be a different row, on g, and does not fix G because G also lumps mixing. A measurement is worth what it adds in the direction in which the posterior is wide.

1.6.3 The recommendation

The recommendation, and the answer to the next-measurement question, is a second sensor in the hall: it removes 0.134 K of the 0.481 K on the temperature and 0.156 kW/K of the 0.561 kW/K on the conductance, 28 % of each, and a 0.2 K reference sensor, if one is available, removes 62 %. A supply-air reading, a load meter and an airflow and coil test are not worth commissioning for these answers. Two conditions apply. The second sensor's error must be independent of the first: a second reading from the same sensor at the same condition would share any bias of the first, and a shared error is a discrepancy that repeats (Section 1.1.4), which a repeated reading does not reduce. And the gain is computed at $\hat{\mathbf{z}}$ under the Laplace approximation, which the checks found adequate for T and slightly narrow for G. Once the measurement is taken, its row joins the three of equation (1.10), and the method returns to step 3 with four rows. For the design-load decision the narrower conductance band sharpens the answer but does not move the threshold: the decision of Section 1.5.2 turns on more cooling in the design, more airflow or a colder supply air if the unit can deliver it (Chapter 14), not on more measurements.

1.6.4 What the analysis showed

The design sheet's nameplate forward solve put the hall at 26°C at its 60 kW design load, a comfortable pass, and its commissioning reading, taken as exact, at 28.4°C, a fail; the analysis, which holds the two against each other through the energy balance, finds neither, but a probability of about one in four that the hall carries its design load below 27.5°C, 0.246 from the posterior of the state (0.281 exact). It also overturns the obvious candidate: an airflow and coil test, although it measures the very heat the question is about, removes nothing, because it repeats what the energy balance already asserts. One reading and a handbook value pin the hall temperature to about half a degree but the conductance of its air path only to about ten percent, and the two are not known separately: the posterior lives on the physics curve, a ridge along which a warmer hall is a hall with a weaker path, which is why the correlation of the two entries is -0.9998. The design review's question about the design load is a question about that conductance. The hall meets 27.5°C at 60 kW only if its path conducts more than 6.316 kW/K; the evidence places it below that, and the answer leans on the handbook value. What would narrow the answers is a second sensor in the hall, which removes 28 % of both bands. What would change the design-load answer is more cooling in the design: more airflow (a larger conductance) or a colder supply air, if the unit can deliver it (Chapter 14); not more measurements.

1.6.5 Limits of the model

The posterior is exact only under the model, the steady-state graph of one hall with one path to a supply air that the discharge controller holds, and that model can be wrong before any inference is run. The supply air is pinned only while the controller holds it: a cooling unit driven to full output can no longer hold T_c, which then rises with the load and would have to become a free quantity of the graph, including at the 60 kW of the scenario node, which the design-load answer assumes the unit can still hold at 18°C. The record itself is illustrative: the 23.2°C reading and the handbook value are constructed for the example, not field data, and no comparison with a true state is claimed.

Modeling mistakes. Several errors can be made before any inference algorithm is run, in the construction of the graph and its rows. Table 1.10 lists five. Each row names the mistake, gives its form in the equipment hall, and states the correction.

Table 1.10: Five modeling mistakes made before any inference is run.

Mistake	In the equipment hall	Correction
Treating a reading as the true state	the hall sensor reports 23.2°C, and $T = 23.2$°C is assumed	the reading is evidence with error σ_T; T stays a free entry of the state
Confusing parameters and observations	the conductance and the reading are treated alike	G is a physical property, an entry of the state; y_T is evidence about T, connected to G through the physics row
Assuming the physical model is complete	the steady-state relation is used just after a fan trip	stored heat matters in a transient, $C_{\mathrm{th}}\,\mathrm{d}T/\mathrm{d}t = Q - G\,(T - T_c)$
Confusing structural with parameter uncertainty	doubt whether the hall has one or two effective air paths is folded into the spread of G	the value of G is a parameter question within one model; the number of paths is a question between two models
Believing the graph alone determines the posterior	a correct graph is expected to give a narrow answer	useful inference also needs suitable evidence, assumptions and constraints; some entries may remain poorly identified

The first two mistakes confuse the roles of Table 1.3. A sensor reporting 23.2°C does not establish that the true temperature is exactly 23.2°C: the observation has an uncertainty, and the physics row of Section 1.1.4 showed that 23.2°C and 23°C imply conductances of 5.77 and 6 kW/K, a difference of 4 %. The conductance is a physical property of the path; the reading is evidence about the temperature, which is connected to the conductance through the model. The fourth mistake mixes two kinds of uncertainty that require different treatments. Uncertainty about the numerical value of G within the model of Figure 1.2 is represented by the posterior of g. Uncertainty about whether the hall has one or two effective air paths is uncertainty about the model M itself, and it is represented by comparing the evidence of the two drawings. The fifth mistake is the converse of the first. A graph provides relations; the posterior also requires evidence, and some entries may remain poorly identified even when the model is perfectly specified. In the two-path drawing, for example, a reading of the hall temperature determines only the total conductance of the two paths and leaves their split to the prior: the Jacobian of the physics and reading rows has a null vector along the split, and only a prior row adds information in that direction.

The steady-state assumption. The third mistake concerns the physics itself. The model of equation (1.1) assumes steady conditions and a single effective conductance. During transient operation the hall stores and releases heat, and the energy balance gains a storage term,

$$C_{\mathrm{th}}\,\frac{\mathrm{d}T}{\mathrm{d}t} = Q - G\,(T - T_c),$$

where C_{th} is the thermal capacitance of the hall, its air, structure and servers, in kJ/K, and t is time in seconds. For constant Q, G and T_c, write the departure from the steady temperature as $u = T - T_c - Q/G$. Then $\mathrm{d}u/\mathrm{d}t = \mathrm{d}T/\mathrm{d}t$ and the balance becomes

$$C_{\mathrm{th}}\,\frac{\mathrm{d}u}{\mathrm{d}t} = -G\,u,$$

whose solution is $u(t) = u(0)\,\mathrm{e}^{-t/\tau}$ with the time constant $\tau = C_{\mathrm{th}}/G$. Therefore

$$T(t) = T_c + \frac{Q}{G} + \Big(T(0) - T_c - \frac{Q}{G}\Big)\,\mathrm{e}^{-t/\tau}, \qquad \tau = \frac{C_{\mathrm{th}}}{G}. \tag{1.34}$$

The steady-state relation of equation (1.2) is the limit $t \to \infty$. Suppose (illustratively) that at $t = 0$ the cooling unit loses half its airflow (one of its two fans trips), so that the conductance falls from 6 to 3 kW/K while the hall is at 23°C under 30 kW. The new steady temperature is $18 + 30/3 = 28$°C, and equation (1.34) gives $T(t) = 28 - 5\,\mathrm{e}^{-t/\tau}$: after one time constant the hall is at $28 - 5\,\mathrm{e}^{-1} = 26.16$°C. A steady-state model would place it at 28°C at once. It is therefore inadequate for predicting temperatures immediately after a fan trip, where the time constant decides whether the hall has minutes or seconds before it reaches its limit; the graph for the airflow question, the partial loss of cooling, needs the storage term and C_{th} as a further entry or a pinned input.

The chapter has built the representation on which every question will be asked. The next chapter, Chapter 2, turns to the questions themselves: how a planner's question stated in the design review's language is recognized as estimation, prediction, comparison, diagnosis or design, and written as a precise probability or expectation over the state of such a graph; Chapter 3 derives the probability rules the posterior rests on, and Chapter 4 dissects the inference and names the scenario node of every design what-if. From Part II the hall becomes a data hall with three units and five racks, and the question becomes how much load it can take (Chapter 5).

1.7 What to retain

Definition 1.1 (Plant graph)**.** A plant graph represents a physical system as free and pinned nodes, paths and sources, with the physical relations that must hold among them written as residual rows; its free quantities, physical states and parameters alike, form the state vector $\mathbf{z}$; readings, records and prior knowledge are attached as further rows; and every planner's question is then a probability or an expectation over $\mathbf{z}$, conditional on the evidence $\mathbf{y}$, computed from the posterior $p(\mathbf{z} \mid \mathbf{y})$ that the rows define.

The forward calculation, the posterior and the question are three different things computed over the same graph. The forward calculation uses no evidence and gives a number; the posterior combines the physics, the readings and the prior; the question picks what is read from the posterior. The posterior is computed in state-vector form: the rows, each divided by its standard deviation, form one residual vector over $\mathbf{z}$; the normal equations $(\mathbf{J}^{\mathsf{T}}\mathbf{J})\,\Delta\mathbf{z} = -\mathbf{J}^{\mathsf{T}}\mathbf{r}$ give the Gauss–Newton steps to the MAP state $\hat{\mathbf{z}}$; the information matrix $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ is a sum of one outer product per row; the covariance $\boldsymbol{\Sigma}$ is the inverse of $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ plus the rows' own curvature, here the curvature of the physics curve; and an answer is an entry of $\hat{\mathbf{z}}$ with its $\sqrt{\Sigma_{ii}}$, or a derived quantity with $\sqrt{\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\,\mathbf{g}}$. An exact physical law is the limit $\sigma_p \to 0$ of its row, and its signature in $\boldsymbol{\Sigma}$ is a correlation near -1: a ridge along the zero set of the row, in the direction of the null vector of the row's Jacobian. A question about a quantity that is a monotone function of one entry of the state is transformed into a threshold on that entry by the physics alone, as the design-load question became $G > 6.316$ kW/K, and its probability is a normal tail of that entry's marginal. The answer is computed from $\hat{\mathbf{z}}$ and $\boldsymbol{\Sigma}$, with the exact value along the curve set beside it and the gap stated, checked against the full calculation (the limit $\sigma_p \to 0$, the exact posterior along the curve, the agreement of the redundant rows, the weight of the prior), reported with its band in the units of the question, and followed by the next measurement, ranked by the preposterior covariance $\boldsymbol{\Sigma}_a = (\boldsymbol{\Sigma}^{-1} + \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a)^{-1}$ of a row not yet taken; a measurement of a pinned quantity, or a row parallel to a physics row, removes nothing. Table 1.11 lists the chapter's terms with their expressions, units and the subsection that defines each.

Table 1.11: The terms of the plant graph and where each is defined. Units for the equipment hall.

Term	Expression	Units	Meaning	Defined in
Plant graph	nodes, paths, sources, rows	–	the system before any question	1.1.4
Free, pinned	T, g free; T_c, Q pinned	–	inferred against known	1.1.4
Energy balance	$Q = G\,(T - T_c)$	kW	conservation at the hall node	1.1.4
Forward calculation	$T = T_c + Q/G$	°C	temperature from known inputs	1.1.4
State vector	$\mathbf{z} = (T, g)^{\mathsf{T}}$	°C, –	all free quantities at once	1.1.4
Log-conductance	$g = \ln(G/G_{\mathrm{ref}})$	–	keeps G positive	1.1.4
Physics row	$(T - T_c - (Q/G_{\mathrm{ref}})\mathrm{e}^{-g})/\sigma_p$	–	a relation that must hold, tolerance σ_p	1.1.4
Model discrepancy	$r_{\mathrm{phys}}(\mathbf{z}) = \delta$	kW	model error, not sensor error	1.1.4
Reading row	$(T - y_T)/\sigma_T$	–	evidence about the state	1.1.7
Prior row	$(g - g_0)/\sigma_g$	–	knowledge before the readings	1.1.7
Jacobian pattern	which row acts on which entry	–	a property of the graph	1.1.7
Scenario node	$T_2 = T_c + (Q_2/G_{\mathrm{ref}})\mathrm{e}^{-g}$	°C	an unobserved condition sharing g	1.1.7
Factor graph	entries and rows	–	the graph of the inference	1.1.3
Delta-method band	$\sigma_G^2 = \mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}$, $\mathbf{g} = (0, \hat{G})^{\mathsf{T}}$	kW/K	spread of a derived quantity	1.3.1
Ridge	$\mathbf{J}_1\mathbf{n} = 0$, $\boldsymbol{\Sigma} \to \mathbf{u}\mathbf{u}^{\mathsf{T}}/\lambda_{\parallel}$, $\lambda_{\parallel} = (n_1^2/\sigma_T^2 + 1/\sigma_g^2 + n_1 r_2/\sigma_T)/(n_1^2 + 1)$	–	posterior along the physics curve	1.2.1
Exact posterior on the curve	$p(g \mid \mathbf{y})$, equation (1.19)	–	the limit $\sigma_p \to 0$ by quadrature	1.2.1
Threshold probability	$\Pr(T_2 < T_{\max} \mid \mathbf{y})$	–	prediction under a threshold	1.3.2
Equivalent threshold	$G^* = Q_2/(T_{\max} - T_c)$	kW/K	the same event on an entry	1.3.2
Posterior	$p(\mathbf{z} \mid \mathbf{y})$	–	plausible states after the evidence	1.2.1
Residual vector	$\mathbf{r}(\mathbf{z})$, $\mathcal{L} = \frac{1}{2}\mathbf{r}^{\mathsf{T}}\mathbf{r}$	–	all rows in standard deviations	1.1.7
Jacobian	$\mathbf{J} = \partial\mathbf{r}/\partial\mathbf{z}$	varies	sensitivity of each row to each entry	1.1.7
Normal equations	$(\mathbf{J}^{\mathsf{T}}\mathbf{J})\,\Delta\mathbf{z} = -\mathbf{J}^{\mathsf{T}}\mathbf{r}$	varies	the Gauss–Newton step to $\hat{\mathbf{z}}$	1.2.1
MAP state	$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}} \mathcal{L}(\mathbf{z})$	°C, –	the most probable state	1.2.1
Information matrix	$\mathbf{J}^{\mathsf{T}}\mathbf{J} = \sum_k \mathbf{J}_k^{\mathsf{T}}\mathbf{J}_k$	varies	precisions of the rows, added	1.2.1
Laplace covariance	$\boldsymbol{\Sigma} = \left(\mathbf{J}^{\mathsf{T}}\mathbf{J} + \sum_k r_k\,\partial^2 r_k/\partial\mathbf{z}\,\partial\mathbf{z}^{\mathsf{T}}\right)^{-1}$	varies	spreads and correlation of the entries	1.2.1
Preposterior covariance	$\boldsymbol{\Sigma}_a = (\boldsymbol{\Sigma}^{-1} + \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a)^{-1}$	varies	spreads after a row not yet taken	1.6.1
Time constant	$\tau = C_{\mathrm{th}}/G$	s	transient omitted by steady state	1.6.5

1.8 Exercises

Exercise 1.1. (a) Which statement best describes the role of the hall sensor: it defines the physical law; it reports the exact true state; or it supplies evidence about an uncertain state? (The third: the sensor provides noisy evidence, while the physical model relates the state to the inputs

and the conductance.) (b) A hall loses heat to the same supply air through two cooling units' air paths, with conductances G_1 and G_2. Name the free and pinned quantities, write the physics row, and say what a temperature reading alone can determine. (Free: T, g_1, g_2; pinned: T_c, Q; the physics row, normalized as in equation (1.5), is $(T - T_c - Q/(G_{\text{ref}}(\mathrm{e}^{g_1} + \mathrm{e}^{g_2})))/\sigma_p$; on its zero set its Jacobian row over $\mathbf{z} = (T, g_1, g_2)^\mathsf{T}$ is $(1,\ (T-T_c)G_1/(G_1+G_2),\ (T-T_c)G_2/(G_1+G_2))/\sigma_p$ and the reading row's is $(1/\sigma_T,\ 0,\ 0)$, so the two-by-three Jacobian of the two rows has the null vector $(0,\ G_2,\ -G_1)^\mathsf{T}$, the step $\Delta G_1 = -\Delta G_2$ that leaves the total unchanged; $\mathbf{J}^\mathsf{T}\mathbf{J}$ is singular along it, the reading determines only $G_1 + G_2$, and the split rests on the priors.) (c) Write "will the hall stay below 27°C if the load rises to 45 kW?" as a probability, and say what the graph needs. ($\Pr(T_3 < 27°\text{C} \mid \mathbf{y})$, with a scenario node $T_3 = T_c + (45/G_{\text{ref}})\,\mathrm{e}^{-g}$ at 45 kW sharing g, unobserved.)

Exercise 1.2. (a) A room is cooled by supply air at 20°C; it has a heat load of 24 kW and a conductance of 4 kW/K. Compute its steady-state temperature. ($T = 20 + 24/4 = 26$°C.) (b) A room is cooled by supply air at 19°C; it has a load of 40 kW and a strict limit of 27°C. What conductance keeps it below the limit? ($19 + 40/G < 27$ gives $40/G < 8$, so $G > 5$ kW/K.) (c) For the hall of the chapter, with $G_{\text{ref}} = 6$ kW/K, evaluate r_{phys} and r_{obs} at $\mathbf{z} = (23.2°\text{C},\ 0)$ and at $\mathbf{z} = (23°\text{C},\ 0)$. (1.2 kW and 0; 0 and -0.4: the first state matches the reading and violates the physics, the second satisfies the physics and lies 0.4 standard deviations from the reading.)

Exercise 1.3. Suppose both the supply-air temperature T_c and the conductance G are uncertain. Starting from $T = T_c + Q/G$, write the probability that the temperature exceeds a limit $T_{\max}$, and state what joint distribution is needed to evaluate it. ($\Pr(T > T_{\max} \mid \mathbf{y}) = \iint \mathbf{1}\{T_c + Q/G > T_{\max}\}\, p(T_c, G \mid \mathbf{y})\, \mathrm{d}T_c\, \mathrm{d}G$, which needs the joint posterior of T_c and G; if the two are independent given $\mathbf{y}$, it reduces to $\int p(T_c \mid \mathbf{y})\, \Pr\big(G < Q/(T_{\max} - T_c) \mid \mathbf{y}\big)\, \mathrm{d}T_c$ over $T_c < T_{\max}$, plus $\Pr(T_c \geq T_{\max} \mid \mathbf{y})$.)

Exercise 1.4. For the two-condition hall of Figure 1.3, with a proposed load of $Q_2 = 50$ kW and a limit of 28°C: (a) derive the threshold conductance and the equivalent threshold on the present temperature. ($T_2 = 18 + 50/G < 28$ gives $G > 5$ kW/K; by equation (1.31), $T < 18 + (30/50) \times 10 = 24$°C.) (b) Show that the present temperature determines the temperature at the proposed load without knowledge of G. (On the physics curve both temperatures are functions of g, $T = T_c + (Q_1/G_{\text{ref}})\,\mathrm{e}^{-g}$ and $T_2 = T_c + (Q_2/G_{\text{ref}})\,\mathrm{e}^{-g}$, so a step along the null vector of the physics row moves T_2 by $Q_2/Q_1 = 5/3$ times the move of T; in closed form, dividing the two balances gives $T_2 - T_c = (Q_2/Q_1)(T - T_c)$, equation (1.28).) (c) If the reading $y_T = 23.2$°C with $\sigma_T = 0.5$ K were the only information about T, what would the probability of meeting the limit be? ($\Phi((24 - 23.2)/0.5) = \Phi(1.6) = 0.945$.)

Exercise 1.5. The hall is at 23°C under 30 kW with $T_c = 18$°C when its conductance falls from 6 to 3 kW/K; its thermal capacitance is $C_{\text{th}} = 3600$ kJ/K (illustrative). (a) Compute the time constant and the new steady temperature. ($\tau = 3600/3 = 1200$ s $= 20$ min; 28°C.) (b) Compute the temperature after 10 and 60 minutes, and the time at which it reaches 27.5°C. ($28 - 5\,\mathrm{e}^{-0.5} = 24.97$°C; $28 - 5\,\mathrm{e}^{-3} = 27.75$°C; $5\,\mathrm{e}^{-t/\tau} = 0.5$ gives $t = \tau \ln 10 = 46.1$ min.) (c) State what a steady-state model would predict and when it is adequate. (It predicts 28°C immediately; it is adequate for conditions held for several time constants, not for the first minutes after the fan trip.)

Exercise 1.6. Remove the prior row from the hall of Section 1.2.1, so that the residual vector holds the rows r_1 and r_2 of equation (1.10) over $\mathbf{z} = (T, g)^\mathsf{T}$, with $\sigma_p = 0.01$ K and $\sigma_T = 0.5$ K. (a) Find the MAP state and write the two-by-two Jacobian there. (Both rows vanish at $T = y_T = 23.2$°C and $g = \ln(5/5.2) = -0.0392$; there $\mathbf{J} = \left[\begin{smallmatrix} 100 & 520 \\ 2 & 0 \end{smallmatrix}\right]$, since $5\,\mathrm{e}^{0.0392}/0.01 = 520$.) (b) Show that $\mathbf{\Sigma} = (\mathbf{J}^\mathsf{T}\mathbf{J})^{-1}$ has $\Sigma_{11} = \sigma_T^2$ for every σ_p, and compute $\sqrt{\Sigma_{22}}$. (The Jacobian is square and invertible, so $\mathbf{\Sigma} = \mathbf{J}^{-1}\mathbf{J}^{-\mathsf{T}}$. Writing $\mathbf{J} = \left[\begin{smallmatrix} a & b \\ c & 0 \end{smallmatrix}\right]$, $\mathbf{J}^{-1} = \left[\begin{smallmatrix} 0 & 1/c \\ 1/b & -a/(bc) \end{smallmatrix}\right]$;

its first row gives $\Sigma_{11} = 1/c^2 = \sigma_T^2 = 0.25\,\mathrm{K}^2$, and its second $\Sigma_{22} = (\sigma_p^2 + \sigma_T^2)/(5\,\mathrm{e}^{-\hat{g}})^2$, so $\sqrt{\Sigma_{22}} = \sqrt{0.2501}/5.2 = 0.0962$.) (c) Explain why no curvature term is needed here, and how the result gives the reading-alone probability 0.173 of Section 1.4.4. (Both rows are zero at $\hat{\mathbf{z}}$, so $\sum_k r_k\,\partial^2 r_k/\partial\mathbf{z}\,\partial\mathbf{z}^{\mathsf{T}}$ vanishes; the entry g then has the Laplace marginal $\mathcal{N}(-0.0392,\ 0.0962^2)$, and its tail above $g^* = 0.0513$ is $\Phi((-0.0392 - 0.0513)/0.0962) = \Phi(-0.941) = 0.173$.)

Chapter 2

The Language of Capacity Questions

2.1 Planning questions and their mathematical form

Chapter 1 drew the equipment hall as a plant graph: one hall node whose air temperature balances a heat load against one heat-removal path to the supply air, which the cooling unit's discharge controller holds at a pinned 18°C, with the conductance of that path unknown and a temperature sensor attached to the hall. The graph was built before any question was asked of it. At the design review the hall's line on the design sheet reads 23.2°C at today's 30 kW, 0.8 K inside the 24°C limit, and the design load is 60 kW; on the design sheet alone the hall has margin and can take the load. With the hall's log of committed-load exceedances beside the sheet, the planner asks several questions. They are not the same kind of question even though they concern the same plant: what the conductance of the air path is, whether the hall will exceed its limit at the design load, whether the hall sensor is reading high, whether a second sensor is worth installing, whether the hall will stay within its committed load through the peak hours of a design day, and whether to rent a temporary chiller for the peak season. A design sheet gives each of them a number, and an inference algorithm can return a mathematically correct result to a question nobody asked; the step that prevents both is the translation of the planner's question into the mathematical quantity that answers it. The question of this chapter is

how a planner's question about a plant to be built or expanded is translated into a mathematical expression that probabilistic inference can evaluate, and how that expression determines what must be computed.

The answer is short. Once a probability statement names its event, its evidence and its model, every question about the plant is a probability or an expectation over the state of its plant graph, with the unknown on the left of the conditioning bar and the evidence on the right. Planning questions about a plant then fall into five categories, estimation, prediction under a threshold, hypothesis comparison, diagnosis and design, each with its general forms, integrals or ratios written without reference to any particular plant. Five computations supply every quantity those forms contain: the posterior of the state; the trend, which reads that posterior along time; the evidence of a model; the preposterior of a measurement not yet taken; and the expected consequence of an action. Two of them, the trend and the expected consequence, have general forms of their own, which the categories call on. What differs from one plant to the next is the specialization: which entry or function of the state the form is applied to, which event or family of models it names, and which objects of the fit carry the answer.

Jargon: event, conditional probability, indicator function, committed load, exceedance log, exceedance rate, coincidence, posterior predictive, trend, crossing time, evidence, Bayes factor, expected consequence, risk, value of information

An *event* is a statement about the plant that is either true or false, such as "the hall temperature exceeds 28°C"; a probability assigns it a number between zero and one. The *exceedance log* is the record of the times at which a hall's metered heat load rose above its committed load, and the *exceedance rate* is the expected number of such exceedances per hour of logged operation. The *committed load* is the load allocated to the hall's tenants, 30 kW for the equipment hall today against its 60 kW design load, and brief workload peaks exceed it: the committed load is the planner's promise, and the exceedance rate is the evidence on how often the promise is broken. Two loads are *coincident* when they peak together; the design sheet carries their coincidence as a coincidence (diversity) factor, and the probability that both exceed at once is not the product of their separate probabilities when a common cause drives both. A *conditional probability* $\Pr(A \mid B)$ is the probability of event A when the information B is treated as known. The *indicator function* $\mathbf{1}_A(\mathbf{z})$ of an event equals one for the states at which the event is true and zero elsewhere, so that the probability of the event is the expectation of its indicator. A *posterior predictive* probability is the probability of a future event that stays random even when the state is known, averaged over the posterior of the state. A *trend* is the change per unit time of a quantity the state carries, and the *crossing time* is the time at which a quantity that moves reaches a stated level. The *evidence*, or marginal likelihood, of a model is the probability of the observed evidence averaged over the prior of that model's unknowns; the *Bayes factor* is the ratio of the evidences of two models. The *expected consequence* of an action, also called its expected loss, is the probability-weighted average of the consequences that may follow it; summed over the events that may follow, it is each event's probability times its consequence, the *risk* of the action. The *value of information*, here the expected value of sample information (EVSI), is the reduction in expected loss that a measurement is expected to bring because the decision can be taken after its result is known.

Key idea

Every planning question about a plant is answered by a specific mathematical object computed from the same plant graph and its rows: a marginal, a tail probability, a predictive average, the posterior read along time, a ratio of evidences, a preposterior or an expected consequence minimized over actions. Five computations produce these objects, the posterior, the trend, the evidence, the preposterior and the expected consequence. The question, not the line on the design sheet or the inference tool, decides which object, and so which computation, is required.

The questions are put to one plant, the equipment hall of Chapter 1, and to the log of its committed-load exceedances, so that the differences between them lie in the questions alone.

2.1.1 The plant graph of the equipment hall

Figure 2.1 draws the plant graph of Chapter 1, with the scenario node that the hall example of Section 2.4.1 adds to it. The upper green box is the free node of the hall at the present load, whose temperature T_1 is an entry of the state; the lower green box is the same hall at the design load of 60 kW, a scenario node whose temperature T_2 is also an entry and which has no sensor; the gray box on the right is the pinned node, the supply air at $T_c = 18$°C, held there by the discharge controller of the cooling unit; the two solid green arrows are the heat-removal path at the two loads, which carry one and the same effective thermal conductance G, held in the state as the log-conductance $g = \ln(G/G_{\text{ref}})$; the white boxes are the sources, the heat loads Q_1 and

Q_2; and the dashed box on the left, joined to the hall by a dashed arrow, is the observation, the reading y_T of the present-load temperature. Each hall node carries one physical row, its steady-state energy balance, written for a node of temperature T and load Q as

$$Q = G\,(T - T_c), \tag{2.1}$$

where Q is the heat load in kW, G the conductance in kW/K, T the hall air temperature and T_c the supply-air temperature, both in °C. The supply air is pinned because the unit's controller holds it; the pin fails when the unit saturates, at a load its coil cannot carry, which the questions of this chapter do not reach. Solved for the hall temperature, it gives the relation used throughout the chapter,

$$T = T_c + \frac{Q}{G}. \tag{2.2}$$

The relation is deterministic: once G is known, T is known. What makes the hall a probabilistic problem is that G is not known, and that a reading y_T differs from T by a measurement error.

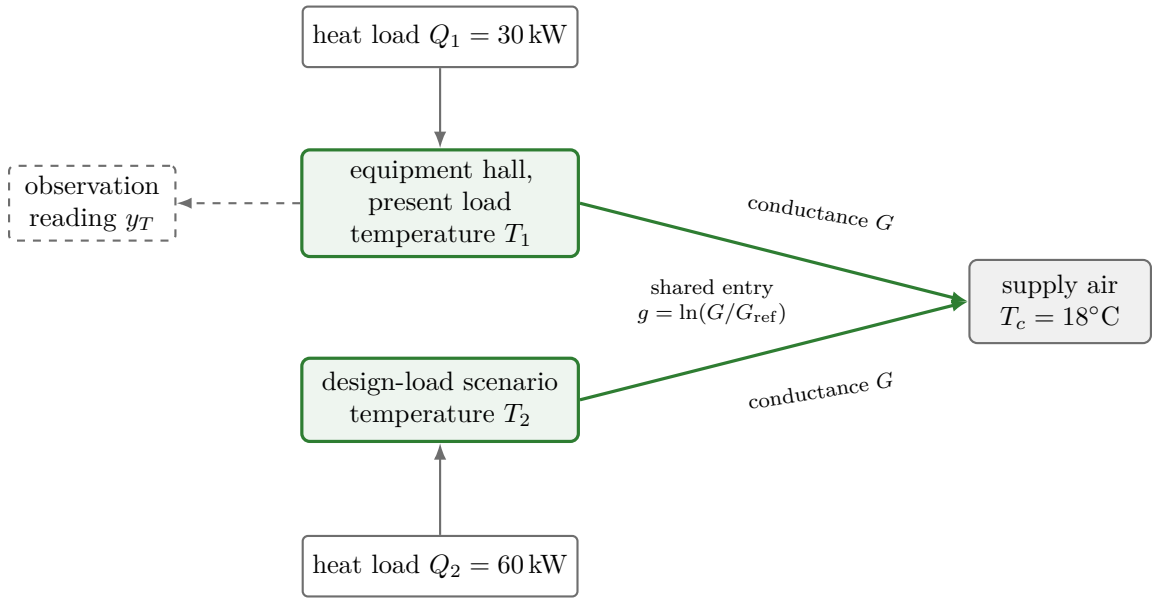


Figure 2.1: The plant graph of the equipment hall with the design-load scenario node. Free: the present-load temperature T_1 and the design-load temperature T_2 (green nodes) and the log-conductance g shared by the two heat-removal paths (solid green edges). Pinned: the supply-air temperature $T_c = 18°\mathrm{C}$, held by the cooling unit's discharge controller (gray node). Sources: the heat loads $Q_1 = 30\,\mathrm{kW}$ and $Q_2 = 60\,\mathrm{kW}$. Observed: the reading y_T of the present-load temperature (dashed box and arrow); the scenario node has no sensor.

2.1.2 Four questions about one plant

Suppose the planner of the hall asks the four questions of Table 2.1. Each row is a question in the planner's words; the second column names the kind of calculation it requires and the third column the quantity that answers it, in the notation that Section 2.2 defines.

Question A estimates an unknown physical quantity. Question B asks for the probability of an event at a load not yet applied. Question C distinguishes between two explanations of the same readings, one in which the sensor reads the hall temperature with zero-mean error and one in which it carries a bias. Question D predicts how a reading that does not yet exist would change a design decision. All four concern the same graph, and all four are computed from the same physical model; none of them is answered by the same calculation as another.

Table 2.1: Four questions about the equipment hall and the calculation each requires.

Planner's question	Calculation	Answering quantity
A. What is the effective conductance of the hall's air path?	estimation of an unknown physical quantity	$p(G \mid \mathbf{y})$
B. Can the hall carry its 60 kW design load and stay below 27.5°C?	probability of an event	$\Pr(T_2 < 27.5^\circ\mathrm{C} \mid \mathbf{y})$
C. Is the hall sensor reading high?	comparison of competing explanations	$\Pr(M_{\text{bias}} \mid \mathbf{y})$ against $\Pr(M_{\text{unbiased}} \mid \mathbf{y})$
D. Is a second hall sensor worth installing before the design is signed?	prediction of what a future reading would change	the preposterior, and the expected loss after the reading

Principle 2.1. *The same physical graph does not imply the same probabilistic calculation. The formulation of the question determines which calculation is needed.*

2.2 The grammar of probability statements

A question can be classified only when the statement that expresses it is complete: an event or an unknown quantity, the evidence it is conditioned on, and the model under which its probability is computed.

2.2.1 Events

An event is a statement that can be true or false. Three events about the plant are

$$T > 28^\circ\mathrm{C}, \qquad G < 5\,\mathrm{kW/K}, \qquad N_e \geq 2,$$

where N_e is the number of exceedances of the hall's committed load during the peak hours of a design day. A probability assigns to an event a number between zero and one. The statement

$$\Pr(T > 28^\circ\mathrm{C}) = 0.10$$

means that the probability assigned to the temperature-exceedance event is 10 %. The statement is incomplete in one respect: it does not say how the probability was calculated or what information it incorporates. A probability computed before any reading of the hall was taken and one computed after thirty readings are different numbers for the same event, and the notation must say which is meant. That is the role of conditioning.

2.2.2 Conditional probability and the evidence

The probability of event A given the information B is written

$$\Pr(A \mid B), \tag{2.3}$$

and read "the probability of A, given B". In capacity planning the information on the right of the bar is the evidence: the operating record of the existing plant (temperatures, loads, meter readings), the log of committed-load exceedances, capacity tests and surveys, the drawings and the run states they assume, nameplate and handbook values, and what commissioning

established. The book collects all of it in the bold symbol $\mathbf{y}$, and keeps a plain y with a subscript, such as y_T, for a single reading. The statement

$$\Pr(T > 28^\circ\mathrm{C} \mid \mathbf{y})$$

asks for the probability that the hall temperature exceeds 28°C, conditional on the available evidence. It differs from $\Pr(T > 28^\circ\mathrm{C})$, which does not condition on that evidence explicitly. The unknown, the event about the plant, stands on the left of the bar; everything treated as known stands on the right. Keeping that order fixed is what prevents questions from becoming entangled when several are asked of one plant.

2.2.3 The role of the model

A complete probability statement also names the model under which it is computed:

$$\Pr(A \mid \mathbf{y}, M), \tag{2.4}$$

where A is the event of interest, $\mathbf{y}$ the evidence, and M the assumed mathematical model, which comprises the physical relations of the plant graph and the probability assumptions about errors and unknowns. The thermal model of equation (2.2) assumes steady state; a model that adds the thermal storage of the hall would assign a different probability to the same temperature event under the same evidence, because it predicts a different temperature after a change of load. The book sometimes suppresses M to keep the notation readable. The dependence on the model remains in every probability, and a probability is a statement about the plant only to the extent that the model is.

2.2.4 Unknown quantities that are not events

An unknown conductance G is a numerical quantity, not an event; it has a value, not a truth value. What can be asked of it is its probability distribution, described for a continuous quantity by a probability density,

$$p(G \mid \mathbf{y}, M),$$

where the lowercase p denotes a density. Alternatively, a question can ask for the probability that the conductance falls within a range, for example

$$\Pr(5 < G < 7 \mid \mathbf{y}, M),$$

and that is an event probability, because "$5 < G < 7$" is a statement that is true or false. The two are connected by integration of the density over the range that defines the event:

$$\Pr(5 < G < 7 \mid \mathbf{y}, M) = \int_5^7 p(G \mid \mathbf{y}, M)\,\mathrm{d}G, \tag{2.5}$$

with the limits in kW/K. A distribution describes uncertainty across all possible values of a quantity; an event probability describes how much of that probability falls within a specified set of values. A density is not a probability: it has units, $\mathrm{kW}^{-1}\,\mathrm{K}$ here, and it can exceed one. Only its integral over a set is a probability.

2.2.5 A reusable notation

The questions of this chapter, and of the book, are written with the symbols of Table 2.2. The first column is the symbol, the second its meaning. The state $\mathbf{z}$ holds every free quantity of the plant graph at once, the physical quantities such as the log-conductance as well as the temperatures; there is no separate parameter vector. The action a stands for a design

decision such as renting a temporary chiller for the peak season, adding a unit, choosing a unit size or commissioning a capacity test; the symbol $\tilde{y}_a$ stands for the outcome of a candidate measurement a that has not yet been taken, and y_{new} for a future outcome of the plant itself, such as the time of the next exceedance of the committed load.

Table 2.2: The reusable notation of probabilistic questions.

Symbol	Meaning
$\mathbf{z}$	state vector of the plant graph: every unknown quantity, physical parameters included
$\mathbf{y}$	observed evidence: readings, counts, records, priors
M, M_j	assumed model or system structure; one of several competing models
$p(\mathbf{y} \mid M)$	evidence of model M: the probability of the evidence with the model's state integrated out
B_{AB}	Bayes factor of model M_A against M_B, $p(\mathbf{y} \mid M_A)/p(\mathbf{y} \mid M_B)$; B_{10} is M_1 against M_0
A	an event
$f(\mathbf{z})$, f^*	a quantity derived from the state; a requirement or limit on it
κ	assurance level: the posterior probability a requirement must reach
y_{new}	a future observation that stays random when the state is known
a	a design action or decision, or a candidate measurement
$\tilde{y}_a$	the outcome of candidate reading a, not yet observed
u, $U(a)$	a quantity of the posterior after $\tilde{y}_a$; its expectation over the outcomes, which ranks the candidates
$\mathbf{\Sigma}_a$	covariance of the state after candidate measurement a (the preposterior)
$L(a, \mathbf{z})$	loss or cost of action a when the state is $\mathbf{z}$: its consequence
$\mathcal{R}(a)$	expected loss of action a over the posterior, $\mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$
c_a, C_e	the fixed cost of action a; the consequence of event e
$\mathbf{g}_a$, $\mathbf{H}_a$	gradient and curvature of $L(a, \mathbf{z})$ in the state at $\hat{\mathbf{z}}$
$f_t(\mathbf{z})$, $\mathbf{g}(t)$	a quantity at time t carried by a state indexed by time; its gradient
α, γ	the level and the trend entries of such a state: the quantity at the record's center $\bar{t}$ and its change per unit time
t^*	crossing time: the time at which f_t reaches a level f^*

With these symbols, questions as different as the conductance of an air path, a design day within the committed load and the choice of a capacity test are written in one mathematical language, and the differences between them become visible as differences in the expression, not in the words.

2.3 The probability over the state, the general forms and the five computations

Every question of Section 2.1 is an integral over the posterior of the state, a ratio of such integrals, or the minimum of one over a set of actions. The categories of question differ in the integrand, and the integrands draw on five computations: the posterior, the trend, the evidence, the preposterior and the expected consequence.

2.3.1 The probability of an event as an integral over the state

Suppose that inference has produced the posterior distribution of the state,

$$p(\mathbf{z} \mid \mathbf{y}, M),$$

and that a question asks whether the state lies in some region $\mathcal{A}$ of the state space. For the hall, $\mathcal{A}$ might be the set of all combinations of temperature and log-conductance that satisfy a design limit. The probability of the region is the posterior integrated over it,

$$\Pr(\mathbf{z} \in \mathcal{A} \mid \mathbf{y}, M) = \int_{\mathcal{A}} p(\mathbf{z} \mid \mathbf{y}, M)\, \mathrm{d}\mathbf{z}, \tag{2.6}$$

which is equation (2.5) with a vector in place of one quantity. The same expression is written with the indicator function of the event,

$$\mathbf{1}_A(\mathbf{z}) = \begin{cases} 1 & \text{if event } A \text{ holds at } \mathbf{z}, \\ 0 & \text{otherwise,} \end{cases}$$

which extends the integral to the whole state space and lets the indicator select the region:

$$\Pr(A \mid \mathbf{y}, M) = \int \mathbf{1}_A(\mathbf{z})\, p(\mathbf{z} \mid \mathbf{y}, M)\, \mathrm{d}\mathbf{z} = \mathbb{E}\big[\mathbf{1}_A(\mathbf{z}) \mid \mathbf{y}, M\big]. \tag{2.7}$$

The last form reads the probability of an event as the posterior expectation of its indicator. The equation states the central rule of the chapter: *the probability of an event is the posterior integrated over the states at which the event is true.* Many planning calculations are variations of it, differing only in the event and in how the integral is evaluated.

2.3.2 Future events and the posterior predictive

Some events remain random even when the state is known exactly. A workload peak whose rate is known still arrives at a random time; knowing the rate fixes only the probability that a design day passes without one. For such a future event A_{future}, the model supplies the conditional probability $\Pr(A_{\text{future}} \mid \mathbf{z}, M)$, and the question asks for its probability given the evidence. The derivation uses three rules. By the law of total probability, the probability of the event is the integral of its joint probability with the state:

$$\Pr(A_{\text{future}} \mid \mathbf{y}, M) = \int \Pr(A_{\text{future}}, \mathbf{z} \mid \mathbf{y}, M)\, \mathrm{d}\mathbf{z}.$$

By the product rule, the joint factors into a conditional and a marginal:

$$\Pr(A_{\text{future}} \mid \mathbf{y}, M) = \int \Pr(A_{\text{future}} \mid \mathbf{z}, \mathbf{y}, M)\, p(\mathbf{z} \mid \mathbf{y}, M)\, \mathrm{d}\mathbf{z}.$$

Finally, the model assumes that the future event depends on the past evidence only through the state: once $\mathbf{z}$ is known, the readings and records of the past add nothing about it. Under that conditional independence, $\Pr(A_{\text{future}} \mid \mathbf{z}, \mathbf{y}, M) = \Pr(A_{\text{future}} \mid \mathbf{z}, M)$, and

$$\Pr(A_{\text{future}} \mid \mathbf{y}, M) = \int \Pr(A_{\text{future}} \mid \mathbf{z}, M)\, p(\mathbf{z} \mid \mathbf{y}, M)\, \mathrm{d}\mathbf{z}. \tag{2.8}$$

This is the posterior predictive probability, the foundation of every prediction of a future exceedance, reading or excursion. Equation (2.7) is the special case in which the conditional probability is an indicator, zero or one at each state. The difference between the two matters as soon as the event is a future exceedance rather than a property of the present plant, which Section 2.4 shows with the hall's exceedance log.

2.3.3 The general forms of the five categories

Planning questions about a plant are many, but the mathematical objects that answer them are few, and the book classifies questions by the object rather than by the design-review word in the question. Five categories cover every question the book asks, each with its general forms, written here in the notation of Table 2.2: $p(\mathbf{z} \mid \mathbf{y})$ is the posterior density of the state; $f(\mathbf{z})$ is a quantity derived from the state, an entry being the case $f(\mathbf{z}) = z_i$, and f^* a value of it, a requirement in a threshold question; y_{new} is a future observation; M_j is a complete model of the plant, a drawing with its rows; and $\mathbf{\Sigma}_a$ is the covariance of the state after a candidate measurement a. Nothing of any particular plant appears in the forms. Two further forms, of a trend (Section 2.3.4) and of an expected consequence (Section 2.3.5), belong to computations that the categories call on rather than to a category of their own. Every later chapter opens its derivation with one of these forms and then specializes it to its plant.

Estimation. An estimation question asks what a quantity is and how well it is known. Its general form is the posterior distribution of the quantity, the probability that it lies at or below each value f^*,

$$\Pr(f(\mathbf{z}) \le f^* \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) \le f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \tag{2.9}$$

which is equation (2.7) for the event $f(\mathbf{z}) \le f^*$, read as a function of f^*. It is summarized by a named location, the mean, the median or the mode, and a spread; the mean and the variance are

$$\begin{aligned} \mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}] &= \int f(\mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \\ \mathrm{Var}(f(\mathbf{z}) \mid \mathbf{y}) &= \int \big(f(\mathbf{z}) - \mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}]\big)^2\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \end{aligned} \tag{2.10}$$

the median is the value f^* at which equation (2.9) equals $\frac{1}{2}$, and the mode is the value at which the density of $f(\mathbf{z})$ is largest. The uncertainty is also stated as the central 95 % credible interval $[f_L, f_U]$, whose endpoints leave 2.5 % of posterior probability in each tail:

$$\Pr(f(\mathbf{z}) \le f_L \mid \mathbf{y}) = 0.025, \qquad \Pr(f(\mathbf{z}) \le f_U \mid \mathbf{y}) = 0.975. \tag{2.11}$$

The quantiles of a monotone function of a quantity are the images of that quantity's quantiles, in the same order for an increasing function and in reverse order for a decreasing one; the median is a quantile and maps the same way, while the mean and the mode in general do not. Under the Laplace posterior $\mathbf{z} \mid \mathbf{y} \approx \mathcal{N}(\hat{\mathbf{z}}, \mathbf{\Sigma})$ an entry is read off directly, with location $\hat{z}_i$ and spread $\sqrt{\Sigma_{ii}}$, and a derived quantity through its gradient $\mathbf{g} = \partial f / \partial \mathbf{z}$ at $\hat{\mathbf{z}}$, with location $f(\hat{\mathbf{z}})$ and spread $\sqrt{\mathbf{g}^\top \mathbf{\Sigma} \mathbf{g}}$ by the delta method. The form is the posterior seen through one quantity: a distribution, of which any single number reported is a named summary.

Prediction under a threshold. A prediction under a threshold asks whether the plant meets a requirement. The requirement is written as the event $f(\mathbf{z}) \ge f^*$, "the requirement holds", with f^* the required value, and its general form is the posterior probability of that event, compared with an assurance level κ, the posterior probability the requirement must reach:

$$\Pr(f(\mathbf{z}) \ge f^* \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) \ge f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \qquad \text{accept if } \Pr(f(\mathbf{z}) \ge f^* \mid \mathbf{y}) \ge \kappa. \tag{2.12}$$

A requirement that a quantity stay below an upper limit, such as a temperature $T \le T_{\max}$, is the same form with $f(\mathbf{z}) = -T$ and $f^* = -T_{\max}$; equivalently, the probability of the requirement is one minus the probability of the complementary event, the exceedance $T > T_{\max}$, and the rule is applied to the requirement, never to the exceedance. A joint event is the same integral with the indicator of an intersection. The required value is the design limit and the assurance

level is the design policy's, and the form keeps the two apart. When the event lies in the future and stays random even when the state is known, the tail is taken of the predictive distribution of the future observation,

$$p(y_{\text{new}} \mid \mathbf{y}) = \int p(y_{\text{new}} \mid \mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \tag{2.13}$$

of which equation (2.8) is the case of an event, with y_{new} the indicator of A_{future}. The predictive is the randomness the model gives the future outcome, averaged over what is not known about the state. When the quantity moves in time and the requirement concerns a date, the event is a crossing, and its form is the trend form of Section 2.3.4.

Hypothesis comparison. A hypothesis comparison asks which of two descriptions of the plant is true. Its general form is the ratio of the evidences of two complete models fitted to the same evidence, the Bayes factor,

$$B_{AB} = \frac{p(\mathbf{y} \mid M_A)}{p(\mathbf{y} \mid M_B)}, \tag{2.14}$$

where the evidence $p(\mathbf{y} \mid M)$ of a model is the probability of the observed evidence with the model's own state integrated out over its prior. The ratio multiplies the prior odds of the two models into their posterior odds, and because each evidence integrates its own state out, it compares the drawings, not the values of their entries. The Bayes factor B_{10} of the notation table is the case $M_A = M_1$ against $M_B = M_0$.

Diagnosis. A diagnosis asks where and what a fault is. Its general form is hypothesis comparison run over a family $M_1, \ldots, M_J$ of candidate faults generated from the plant graph, one for each component or reading that could be at fault, with the answer returned as the ranking of

$$\Pr(M_j \mid \mathbf{y}) \propto p(\mathbf{y} \mid M_j)\, \Pr(M_j), \qquad j = 1, \ldots, J. \tag{2.15}$$

The answer is the order of the candidates with the posterior probability of each, not a single verdict.

Design. A design question asks what to measure or do next, before the measurement exists. Its general form is the preposterior expectation: for a candidate measurement a whose outcome $\tilde{y}_a$ is not yet known, the expected value, over the outcomes weighted by their predictive distribution, of a quantity u of the posterior that the outcome would produce,

$$U(a) = \mathbb{E}_{\tilde{y}_a \mid \mathbf{y}}\Big[\, u\big(p(\mathbf{z} \mid \mathbf{y}, \tilde{y}_a)\big)\Big] = \int u\big(p(\mathbf{z} \mid \mathbf{y}, \tilde{y}_a)\big)\, p(\tilde{y}_a \mid \mathbf{y})\, \mathrm{d}\tilde{y}_a, \tag{2.16}$$

where $p(\tilde{y}_a \mid \mathbf{y})$ is the predictive (2.13) of the outcome; for an outcome that takes discrete values, such as a number of exceedances in a trial period, the integral is a sum over those values weighted by their predictive probabilities. The quantity u is one of four: the spread of a derived quantity after the outcome; the information gained; the indicator that an acceptance rule (2.12) is met after the outcome, which makes $U(a)$ the probability that the trial ends with the requirement accepted; or the expected loss of the best action once the outcome is known. The candidates are ranked by $U(a)$, the largest first when u is an information gain or an acceptance indicator, the smallest first when u is a spread or a loss, and a candidate's cost enters as a loss of its own. When the posterior is Gaussian and the candidate is a linear Gaussian row, $r_a(\mathbf{z}) = (\mathbf{h}_a\mathbf{z} - \tilde{y}_a)/\sigma_a$ with $\mathbf{h}_a$ the row vector of its sensitivities and σ_a the standard deviation of its noise, the covariance of the updated posterior does not depend on

the outcome, and for a spread or an information gain the form needs only two objects: the preposterior covariance,

$$\boldsymbol{\Sigma}_a = \left(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_a\right)^{-1}, \qquad \boldsymbol{\Lambda}_a = \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a = \frac{\mathbf{h}_a^{\mathsf{T}}\mathbf{h}_a}{\sigma_a^2}, \tag{2.17}$$

with $\mathbf{J}_a = \mathbf{h}_a/\sigma_a$ the Jacobian of the candidate row, and the expected information gain, in nats,

$$\mathrm{EIG}(a) = \tfrac{1}{2}\ln\frac{\det\boldsymbol{\Sigma}}{\det\boldsymbol{\Sigma}_a}. \tag{2.18}$$

The spread of a derived quantity f after the measurement is then $\sqrt{\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}_a\mathbf{g}}$, with the same gradient $\mathbf{g}$ at $\hat{\mathbf{z}}$. For a count, demand or lifetime row the candidate's information is the expected information: $\boldsymbol{\Lambda}_a = \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a$ evaluated at $\hat{\mathbf{z}}$ with the expected counts of the trial in place of observed ones, a count row entering $\mathbf{J}_a$ through the line $\sqrt{\mu_k}\,\partial\ln\mu_k/\partial\mathbf{z}$ of its working form (Section 4.2.7); equation (2.17) is then the Laplace preposterior, exact only to the order of the Laplace approximation, and it still does not depend on the outcome. A design question values a posterior that does not exist yet, through what the reading would do to it. When the candidate is an action that brings no new evidence, renting a chiller or adding a unit, there is no outcome to average over, the posterior after the action is the present one, and the form reduces to the expected consequence of Section 2.3.5.

2.3.4 The general form of a trend

Some questions concern a quantity that moves in time: whether an exceedance rate or a unit's derate is rising, when it will cross a planning limit, what it will be at the end of the planning horizon. By its integrand each is an estimation or a prediction under a threshold; what sets it apart is that the quantity is indexed by time, so that the state must carry how the quantity changes as well as where it stands. Write the quantity, on the scale on which the state carries it (a log-rate, a log-derate, a temperature), as $f_t(\mathbf{z})$, its value at time t. In the simplest drawing it is a straight line in time about the center $\bar{t}$ of the record, with a level entry α, the quantity at $\bar{t}$, and a *trend entry* γ, its change per unit time:

$$\mathbf{z} = \begin{bmatrix}\alpha\\ \gamma\end{bmatrix}, \qquad f_t(\mathbf{z}) = \alpha + \gamma\,(t-\bar{t}) = \mathbf{g}(t)^{\mathsf{T}}\mathbf{z}, \qquad \mathbf{g}(t) = \frac{\partial f_t}{\partial\mathbf{z}} = \begin{bmatrix}1\\ t-\bar{t}\end{bmatrix}. \tag{2.19}$$

The evidence arrives in windows, one row per window. Window k, centered at the time t_k, gives the row $r_k(\mathbf{z})$, which sees the state only through the quantity at its own time, $f_{t_k}(\mathbf{z})$, so its Jacobian row is the gradient $\mathbf{g}(t_k)^{\mathsf{T}}$ times the row's sensitivity $d_k = \partial r_k/\partial f_{t_k}$: $1/\sigma_k$ for a reading row of standard deviation σ_k, and $\sqrt{\mu_k}$, the square root of the window's expected count, for a count row in the working form of Section 4.2.7. The rows of the record stack to

$$\begin{gathered}\mathbf{J} = \mathbf{D}\,\mathbf{X}, \qquad \mathbf{X} = \begin{bmatrix}\mathbf{g}(t_1)^{\mathsf{T}}\\ \mathbf{g}(t_2)^{\mathsf{T}}\\ \vdots\end{bmatrix}, \qquad \mathbf{D} = \mathrm{diag}(d_1, d_2, \ldots),\\ \boldsymbol{\Lambda} = \mathbf{J}^{\mathsf{T}}\mathbf{J} = \begin{bmatrix}\sum_k d_k^2 & \sum_k d_k^2\,(t_k-\bar{t})\\ \sum_k d_k^2\,(t_k-\bar{t}) & \sum_k d_k^2\,(t_k-\bar{t})^2\end{bmatrix}, \qquad \boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}.\end{gathered} \tag{2.20}$$

The trend entry of the information, $\sum_k d_k^2(t_k-\bar{t})^2$, weighs each window's distance from the center squared, while the windows and their precisions d_k^2 add linearly: widening the span of a record raises the information on the trend faster than adding windows inside it.

The posterior of this state is read along time in three ways, and the three make up the general form of a trend:

$$\begin{aligned}
\Pr(\gamma > 0 \mid \mathbf{y}) &= \int \mathbf{1}[\gamma > 0]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \\
\Pr(t^* \le s \mid \mathbf{y}) &= \int \mathbf{1}\big[f_s(\mathbf{z}) \ge f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \qquad s \ge \bar{t}, \\
\Pr(f_t(\mathbf{z}) \le f^* \mid \mathbf{y}) &= \int \mathbf{1}\big[f_t(\mathbf{z}) \le f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \qquad t \text{ beyond the record.}
\end{aligned} \tag{2.21}$$

The first line is the sign of the trend, the probability that the quantity is rising: the threshold form on the trend entry. The second is the distribution function of the crossing time $t^*(\mathbf{z})$, the time at which the quantity reaches the level f^*, for an upward crossing of a level that the quantity stands below at $\bar{t}$: such a quantity, if it rises, has crossed by a time $s \ge \bar{t}$ if and only if it is at or above f^* at s, so each value of the distribution function is a threshold probability at one time, and a state with $\gamma \le 0$ never crosses, a mass $\Pr(\gamma \le 0 \mid \mathbf{y})$ at $t^* = \infty$. The third is the forecast, the estimation form at a time beyond the last window of the record. Nothing of a particular plant appears in the three lines; a plant enters through the scale of f_t, the rows of its windows and the level f^*.

Under the Laplace posterior $\mathbf{z} \mid \mathbf{y} \approx \mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$ the quantity at any time is linear in the state, so each line is an exact normal tail in two numbers, the center $\mathbf{g}(t)^{\mathsf{T}}\hat{\mathbf{z}}$ and the variance $\mathbf{g}(t)^{\mathsf{T}}\boldsymbol{\Sigma}\,\mathbf{g}(t)$:

$$\Pr(\gamma > 0 \mid \mathbf{y}) \approx \Phi\left(\frac{\hat{\gamma}}{\sqrt{\Sigma_{\gamma\gamma}}}\right), \qquad \Pr(t^* \le s \mid \mathbf{y}) \approx \Phi\left(\frac{\mathbf{g}(s)^{\mathsf{T}}\hat{\mathbf{z}} - f^*}{\sqrt{\mathbf{g}(s)^{\mathsf{T}}\boldsymbol{\Sigma}\,\mathbf{g}(s)}}\right), \tag{2.22}$$
$$\mathbf{g}(t)^{\mathsf{T}}\boldsymbol{\Sigma}\,\mathbf{g}(t) = \Sigma_{\alpha\alpha} + 2\,(t - \bar{t})\,\Sigma_{\alpha\gamma} + (t - \bar{t})^2\,\Sigma_{\gamma\gamma}.$$

The sign is the same calculation with the gradient $(0,\ 1)^{\mathsf{T}}$, which reads the trend entry alone, so that its variance is $\Sigma_{\gamma\gamma}$; the forecast is the same with the time asked. The variance of the quantity grows with the square of the distance $t - \bar{t}$, the lever arm: a forecast pays for the uncertainty of the trend with that square. The crossing time itself, $t^* = \bar{t} + (f^* - \alpha)/\gamma$ for $\gamma > 0$, is a ratio with the trend entry in its denominator. When the posterior of γ has a positive density at zero, as a Gaussian posterior always has, t^* has no posterior mean and no variance, so it is reported by its median, the plug-in date $t^*(\hat{\mathbf{z}})$ at which the argument of the normal tail is zero, by the band whose ends solve the second line at 0.025 and 0.975, and by the mass that never crosses; the delta band $t^*(\hat{\mathbf{z}}) \pm 1.96\sqrt{\mathbf{g}_*^{\mathsf{T}}\boldsymbol{\Sigma}\,\mathbf{g}_*}$, with $\mathbf{g}_* = \partial t^*/\partial \mathbf{z}$ at $\hat{\mathbf{z}}$, is a finite number standing in for an infinite one. For the exceedance log of a sister hall whose load is growing, its committed-load exceedances counted in six periods of forty days and rising from 5 to 14 a period (Exercise 2.5), the band of the crossing of the planning limit of 0.5 exceedances per logged day runs from 5.90 to 28.3 periods by inversion, with period j centered at $t = j$ and the log ending at $t = 6.5$, against a delta band from 4.44 to 11.45 periods that is wrong at both ends. The rising counts are the growth of the hall's load seen from below, through the hours in which it already rises above the committed load. Chapter 11 reads a crossing of the same kind on the site's load path, where the crossing date less the lead time is the date by which the next unit must be ordered.

2.3.5 The general form of an expected consequence

Other questions ask what to do: whether to rent a temporary chiller for the peak season, whether to build to N+1 or N+2, which unit size to buy. Each compares actions a from a set of candidates by what they are expected to cost. Let $L(a, \mathbf{z})$ be the loss of taking action a when the state is $\mathbf{z}$, its consequence in whatever unit the decision is taken in: reais, hours above the

limit, energy of load shed. It is a derived quantity of the state, one for each action. The general form of an expected consequence is its posterior expectation, the expected loss $\mathcal{R}(a)$ of the action, and the action that minimizes it. When the action costs c_a for certain and leaves events e that occur with probability $\Pr(e \mid \mathbf{z}, a)$ and have the consequence C_e each, the loss given the state is the cost of the action plus the consequence of each event weighted by its probability, and by the predictive (2.8) the posterior expectation turns each conditional probability into its predictive one:

$$\begin{aligned} \mathcal{R}(a) &= \mathbb{E}[L(a,\mathbf{z}) \mid \mathbf{y}] = \int L(a,\mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \qquad a^* = \arg\min_a \mathcal{R}(a), \\ L(a,\mathbf{z}) &= c_a + \sum_e C_e \Pr(e \mid \mathbf{z}, a) \quad \Longrightarrow \quad \mathcal{R}(a) = c_a + \sum_e C_e \Pr(e \mid \mathbf{y}, a). \end{aligned} \tag{2.23}$$

The second line is the rule that risk is probability times consequence, summed over the events and added to the certain cost of the action. Nothing of a particular plant appears in it either; a plant enters through the events, the probabilities its posterior gives them and the consequence schedule c_a, C_e that the planner supplies.

Under the Laplace posterior the expected consequence is read from the same objects as every other answer. With $\mathbf{g}_a = \partial L(a,\mathbf{z})/\partial \mathbf{z}$ and $\mathbf{H}_a = \partial^2 L(a,\mathbf{z})/\partial \mathbf{z}\, \partial \mathbf{z}^{\mathsf{T}}$ at $\hat{\mathbf{z}}$, the expansion of the loss to second order about $\hat{\mathbf{z}}$, averaged over $\mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$, keeps the constant and the curvature term, since $\mathbb{E}[\mathbf{z} - \hat{\mathbf{z}} \mid \mathbf{y}] = \mathbf{0}$ and $\mathbb{E}[(\mathbf{z} - \hat{\mathbf{z}})^{\mathsf{T}} \mathbf{H}_a (\mathbf{z} - \hat{\mathbf{z}}) \mid \mathbf{y}] = \operatorname{tr}(\mathbf{H}_a \boldsymbol{\Sigma})$; the spread follows from the gradient, and the comparison of two actions from the gradient of their difference:

$$\begin{aligned} \mathcal{R}(a) &\approx L(a,\hat{\mathbf{z}}) + \tfrac{1}{2} \operatorname{tr}(\mathbf{H}_a \boldsymbol{\Sigma}), \qquad \operatorname{sd}(L(a,\mathbf{z}) \mid \mathbf{y}) \approx \sqrt{\mathbf{g}_a^{\mathsf{T}} \boldsymbol{\Sigma}\, \mathbf{g}_a}, \\ \Pr(L(a,\mathbf{z}) < L(a',\mathbf{z}) \mid \mathbf{y}) &\approx \Phi\left(\frac{\mathcal{R}(a') - \mathcal{R}(a)}{\sqrt{(\mathbf{g}_a - \mathbf{g}_{a'})^{\mathsf{T}} \boldsymbol{\Sigma}\, (\mathbf{g}_a - \mathbf{g}_{a'})}} \right). \end{aligned} \tag{2.24}$$

The loss at the most probable state, $L(a,\hat{\mathbf{z}})$, misses the curvature term. The curvature is taken in the coordinates of the state, and its sign there, not in a physical quantity, sets the direction of the second-order term. The exceedance log of Section 2.4.4 shows the difference over the next 100 h of operation, about eight design days of twelve peak hours. The probability of no exceedance, $\mathrm{e}^{-\lambda\tau}$, is convex in the rate λ, so by Jensen's inequality the predictive 0.7725 lies above the plug-in 0.7659; but the state carries the log-rate ω, in which the probability is concave, with curvature -0.150 at $\hat{\omega}$, and the second-order reading with $\sigma_\omega^2 = 0.25$ moves the plug-in down to 0.747, close to the average 0.748 over the Laplace normal of the log-rate and away from the exact 0.7725. The term is right about the Laplace posterior and wrong about the exceedances, because the error lies in the Laplace posterior itself; where the second-order term and an exact route move the plug-in in opposite directions, the value read from the state is stated with the exact one beside it, and the gap is what the Laplace posterior costs. The second line of equation (2.24) asks how firm a ranking is: the probability that action a costs less than a' at the true state. The variance of the difference is the sum of the two variances minus $2\, \mathbf{g}_a^{\mathsf{T}} \boldsymbol{\Sigma}\, \mathbf{g}_{a'}$, so the band of the comparison is narrower than the two bands taken apart exactly when $\mathbf{g}_a^{\mathsf{T}} \boldsymbol{\Sigma}\, \mathbf{g}_{a'} > 0$, when the two losses move together across the posterior. Section 2.5.4 applies the form to the choice between renting a temporary chiller for the peak season and waiting, and Chapters 23 and 24 apply it to the design itself.

2.3.6 The five computations beneath the forms

Five computations sit underneath the forms, each defined by the object it returns: the posterior of the state, the trend, the evidence, the preposterior and the expected consequence.

The posterior. The posterior of the state follows from Bayes' rule, the likelihood of the evidence times the prior divided by their integral over the state. For Gaussian rows divided by their standard deviations it is the exponential of minus the objective of the rows,

$$p(\mathbf{z} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{z})\, p(\mathbf{z})}{p(\mathbf{y})} \propto \exp[-\mathcal{L}(\mathbf{z})], \qquad \mathcal{L}(\mathbf{z}) = \tfrac{1}{2}\, r(\mathbf{z})^{\mathsf{T}} r(\mathbf{z}), \tag{2.25}$$

and a count, demand or lifetime row puts its negative log-probability term into $\mathcal{L}$ instead. Under the Laplace approximation the posterior is the normal $\mathcal{N}(\hat{\mathbf{z}}, \mathbf{\Sigma})$ centered at the MAP state $\hat{\mathbf{z}}$, the minimum of $\mathcal{L}$, with the covariance $\mathbf{\Sigma} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}$ built from the Jacobian $\mathbf{J}$ of the rows; Chapter 3 derives the normal equations and the covariance, and Chapter 4 the Laplace approximation. The estimation, threshold and predictive forms are integrals against this posterior, read for an entry $\hat{z}_i$ with $\sqrt{\Sigma_{ii}}$ or for a derived quantity through $\mathbf{g}^{\mathsf{T}}\mathbf{\Sigma}\mathbf{g}$.

The trend. The trend is the posterior over a state indexed by time or by window, with a trend entry among its entries, read along time: as the probability that the trend entry is positive, as a crossing time with its band, or as a forecast, the three lines of equation (2.21). Its building blocks are those of the posterior, $\hat{\mathbf{z}}$ and $\mathbf{\Sigma}$ from the rows of the windows (2.20), with the gradient $\mathbf{g}(t)$ of the quantity at the time asked. It is a reading of the posterior, and it is named as a computation of its own because its object, its check and its failures are its own: the object is a function of the time asked rather than one number; the check orders the record's residuals in time and inverts the threshold line for the band of the crossing instead of linearizing the crossing time; and it fails where no other reading does, in a forecast that pays for the uncertainty of the trend with the square of its lever arm, in a crossing time without a mean when the posterior of the trend reaches zero, and in the shape of the line, which the record tests only inside its own span.

The evidence. The denominator of equation (2.25), written with its model explicit, is the evidence of the model, the probability of the observed evidence with the model's own unknowns integrated out over their prior:

$$p(\mathbf{y} \mid M_j) = \int p(\mathbf{y} \mid \mathbf{z}, M_j)\, p(\mathbf{z} \mid M_j)\, \mathrm{d}\mathbf{z}. \tag{2.26}$$

It is the normalizing constant of the posterior within model M_j, and it is the only place where the comparison between models enters: the posterior inside either model, however sharp, does not say which model is true. Under the Laplace approximation it is read from the model's own fit, as the volume of its posterior times the height of the integrand at its MAP state, $p(\mathbf{y} \mid M_j) \approx (2\pi)^{d_j/2} \det(\mathbf{\Sigma}_j)^{1/2}\, p(\mathbf{y} \mid \hat{\mathbf{z}}_j, M_j)\, p(\hat{\mathbf{z}}_j \mid M_j)$ with d_j the number of entries of its state, exact when the objective is quadratic (Chapter 4). The comparison and diagnosis forms are a ratio and a ranking of it.

The preposterior. The preposterior is the expected posterior under a measurement not yet taken: the posterior that each outcome $\tilde{y}_a$ would produce, averaged over the predictive distribution of the outcomes, the inner object of the design form (2.16). For a linear Gaussian candidate row its covariance does not depend on the outcome, and equation (2.17) follows from the rows. Let the current posterior of the state be approximately Gaussian with mean $\hat{\mathbf{z}}$ and covariance $\mathbf{\Sigma} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}$, and let the proposed measurement depend on the state through its sensitivities, $\tilde{y}_a = \mathbf{h}_a\mathbf{z} + \varepsilon$ to first order, with independent Gaussian noise ε of variance σ_a^2. Divided by its standard deviation, the candidate row is $r_a(\mathbf{z}) = (\mathbf{h}_a\mathbf{z} - \tilde{y}_a)/\sigma_a$, and its Jacobian is the row vector $\mathbf{J}_a = \partial r_a/\partial \mathbf{z} = \mathbf{h}_a/\sigma_a$. The objective after the measurement is the current one plus that row,

$$\mathcal{L}_a(\mathbf{z}) = \tfrac{1}{2}\, r(\mathbf{z})^{\mathsf{T}} r(\mathbf{z}) + \tfrac{1}{2}\, r_a(\mathbf{z})^2,$$

and its curvature, the information, is the curvature of the stacked Jacobian $[\mathbf{J};\ \mathbf{J}_a]$, the sum $\mathbf{J}^\top\mathbf{J} + \mathbf{J}_a^\top\mathbf{J}_a$: the information of independent rows adds. With $\mathbf{J}^\top\mathbf{J} = \mathbf{\Sigma}^{-1}$, the covariance after the measurement under the linear-Gaussian approximation is the inverse of that sum, which is equation (2.17). When the current rows include exact constraints, $\mathbf{\Sigma}$ is singular in the limit $\sigma_p \to 0$ and $\mathbf{\Sigma}^{-1}$ does not exist there. The limit is then taken through the equivalent update form, which uses $\mathbf{\Sigma}$ alone,

$$\mathbf{\Sigma}_a = \mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{J}_a^\top\left(\mathbf{I} + \mathbf{J}_a\mathbf{\Sigma}\mathbf{J}_a^\top\right)^{-1}\mathbf{J}_a\mathbf{\Sigma}, \tag{2.27}$$

where $\mathbf{I}$ is the identity of the size of the number of candidate rows. The value $\tilde{y}_a$ of the measurement does not appear in either form, only its Jacobian, its sensitivity divided by its noise. That is why equation (2.17) predicts how the uncertainty would shrink before the measurement is taken, and why it is one foundation of sensor placement and experimental design. The information gain (2.18) is the standard result for a Gaussian prior and a linear Gaussian measurement; where both determinants vanish in the limit of exact rows, the determinant identity $\det\mathbf{\Sigma}/\det\mathbf{\Sigma}_a = \det(\mathbf{I} + \mathbf{J}_a\mathbf{\Sigma}\mathbf{J}_a^\top)$, which also uses $\mathbf{\Sigma}$ alone, evaluates it.

The expected consequence. This computation returns the expected loss of each candidate action over the posterior, $\mathcal{R}(a) = \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$, and the choice of the action that minimizes it, equation (2.23); summed over events, it is probability times consequence. Its building blocks are $\hat{\mathbf{z}}$ and $\mathbf{\Sigma}$ with the gradient $\mathbf{g}_a$ and the curvature $\mathbf{H}_a$ of each action's loss, equation (2.24). It is the expectation of one more derived quantity, and it is named as a computation of its own because its object is a choice among actions rather than a description of the state, because its check is the curvature term that separates it from the loss at the most probable state together with the probability that the ranking holds, and because its failures are its own: a second-order reading whose sign follows the coordinates of the state, a ranking that is clear in expectation and uncertain at the true state, and a consequence schedule that the posterior does not contain, whose own uncertainty no band of the state carries.

Table 2.3 sets the five categories against the five computations: its first column is the category, its second the computation with the kind of question in the category that calls on it, and its third the object the computation returns. A category calls on more than one computation when its questions differ in kind.

In this book the posterior is used in every chapter, and capacity planning is mostly the posterior read at a scenario node, with the expected consequence of a design leading the decisions: the posterior at the scenarios of Chapters 5 to 8, 13 to 16 and 20 to 22, and over the enumeration of unit states in Chapters 9, 10 and 15; the trend on the exceedance log of Section 2.3.4, with a crossing of the same kind dating the next unit's order in Chapter 11; the evidence in Chapter 4; the preposterior at step 7 of Chapters 1 and 4 onward, and in Chapters 18 and 19; and the expected consequence in Section 2.5.4, in Chapter 9 for the cost of the extra unit, and in Chapters 23 and 24.

2.3.7 The finer question types inside the five categories

A finer list of question types is useful in practice, because each type has a characteristic wording. Table 2.4 lists twelve of them. The first column is the type, the second an example in the planner's words, the third the mathematical target, and the last two the category of the book it belongs to and the computation beneath it. The list does not limit the applications; it collects the mathematical patterns that recur across them.

Most rows fall into a category directly; five need a sentence. A *joint event* such as "both halls stay within their design loads" is an event like any other, a region of the state space, so its probability is equation (2.7) with the indicator of an intersection; the wording "both" changes the region, not the category. A *trend* question asks for threshold events, on the trend entry and

Table 2.3: The five categories of question, the computations that answer them, and the object each computation returns.

Category	Computation	Object returned
estimation	posterior	an entry, $\hat{z}_i$ with $\sqrt{\Sigma_{ii}}$; a derived quantity, $f(\hat{\mathbf{z}})$ with $\sqrt{\mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g}}$
	trend, for the quantity at a stated time and for the distribution of a crossing time	$\mathbf{g}(t)^\mathsf{T}\hat{\mathbf{z}}$ with $\mathbf{g}(t)^\mathsf{T}\boldsymbol{\Sigma}\,\mathbf{g}(t)$; the median crossing time $t^*(\hat{\mathbf{z}})$ and its band by inversion
	expected consequence, for the expected loss of one course of action	$\mathcal{R}(a) = \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$
prediction under a threshold	posterior	a tail of an entry or of a derived quantity; a predictive
	trend, for the sign of a trend and for a crossing by a date	$\Pr(\gamma > 0 \mid \mathbf{y})$; $\Pr(t^* \le s \mid \mathbf{y})$
hypothesis comparison	evidence	a Bayes factor $B_{AB} = p(\mathbf{y} \mid M_A)/p(\mathbf{y} \mid M_B)$
diagnosis	evidence	the ranking of $\Pr(M_j \mid \mathbf{y})$ over a family generated from the graph
design	preposterior, for the next measurement	$\boldsymbol{\Sigma}_a$, the information gain, the value of information
	expected consequence, for the choice of an action	$a^* = \arg\min_a \mathcal{R}(a)$; the probability that the ranking holds

on the quantity at a date, so its category is prediction under a threshold; its computation is the trend, because the state is indexed by time and the crossing time is a ratio that the threshold line must be inverted to bound. An *expectation* of a consequence, $\mathbb{E}[L \mid \mathbf{y}] = \int L(\mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}$, is the mean of the marginal of the derived quantity $L(\mathbf{z})$: its category is estimation, applied to a loss, and its computation is the expected consequence, whose curvature term separates it from the loss at the most probable state; its tail $\Pr(L > l \mid \mathbf{y})$ would be a prediction under a threshold. An *importance* measure (in capacity planning, the gradient of headroom with respect to a rating, which orders the debottlenecking) is a derivative of a derived quantity with respect to an entry of the state, the gradient that the delta method already uses, and it belongs with estimation. A *decision* among actions is a design question answered by the expected consequence, with the preposterior as well when a test comes before the action.

2.3.8 The twenty questions of the design review

The book is organized around twenty questions that a capacity planner brings to a design review, grouped by the decision on the table: whether the plant will carry the load, how much to build, where the bottleneck is, what the uncertainty in the design inputs costs, how the design holds under failure, and what its expected cost and risk are. Table 2.5 lists them in the planner's words. Its first column is the number of the question; the second the question as the planner puts it; the third the general form it takes, by equation number, with what the specialization adds; the fourth the computation beneath it; and the last the chapter that leads with it. Chapters 5 to 8 put the questions to one data hall with three cooling units and five racks, hall B, calibrated on its own record. Chapters 9 to 24 put them to the whole site: two halls on one chilled-water plant with its chillers, pump, shared supply pipe and power feed, and the design alternatives the planner must choose between.

Table 2.4: Twelve question types, their mathematical targets, and the category and computation each belongs to.

Question type	Example	Mathematical target	Category	Computation
Estimation	What is the hall's committed-load exceedance rate?	$p(\lambda \mid \mathbf{y})$	estimation	posterior
Threshold probability	Will the hall stay within its limit at the design load?	$\Pr(T \le T_{\max} \mid \mathbf{y})$	prediction under a threshold	posterior
Joint and conditional events	Will both halls stay within their committed loads through the peak?	$\Pr(\bar{A}_A \cap \bar{A}_B \mid \mathbf{y})$	prediction under a threshold	posterior
Prediction	Will a design day's peak hours pass within the committed load?	$\Pr(T > 12\,\mathrm{h} \mid \mathbf{y})$	prediction under a threshold	posterior
Trend and crossing	Is the exceedance rate rising, and when will it pass its planning limit?	$\Pr(\gamma > 0 \mid \mathbf{y})$, $\Pr(t^* \le s \mid \mathbf{y})$	prediction under a threshold	trend
Derived quantities	How much heat can the unit reject on the design day?	$p(Q_u(\mathbf{z}) \mid \mathbf{y})$	estimation	posterior
Model comparison	Is the air path as drawn, or obstructed?	$\Pr(M_j \mid \mathbf{y})$	hypothesis comparison	evidence
Diagnosis	Which unit falls short of its rating?	$\Pr(F_j \mid \mathbf{y})$	diagnosis	evidence
Risk and expectation	What does a load shed at the season's peak cost on average?	$\mathbb{E}[L \mid \mathbf{y}]$	estimation	expected consequence
Importance	Which design input moves the hall temperature most?	$\partial T/\partial z_i$, an entry of the gradient $\mathbf{g}$	estimation	posterior
Information and experiment design	Which capacity test should be run before committing?	expected information gain or value of information	design	preposterior
Decision optimization	Rent a temporary chiller for the season or wait?	$\arg\min_a \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$	design	expected consequence, with the preposterior when a test comes first

Table 2.5: The twenty planning questions of the book, with the general form and the computation of each and the chapter that leads with it.

No.	Planner's question	General form	Computation	Chapter
Will it carry the load				
1	**Headroom.** How much load can hall B take before an inlet crosses the limit, and how sure is that number?	(2.9), the load at which the limit binds at a scenario node	posterior	5
2	**Growth scenario.** If the IT load grows 15 % a year, what is the probability the plant is still inside its envelope in year three?	(2.12), at the year-three scenario node	posterior	6
3	**Peak day.** On the design-day ambient, with every unit running, what is the probability of a violation anywhere?	(2.12), a joint tail at the design-day scenario	posterior	7
4	**Diversity.** The racks never all peak at once; what is the probability the coincident peak exceeds the plant's capacity?	(2.12), on a sum of correlated load entries	posterior	8
How much to build				
5	**Redundancy level.** N+1 or N+2?	(2.12) over the enumeration of unit states; (2.23) for the extra unit	posterior, expected consequence	9
6	**Unit sizing.** Two 600 kW units or three 400 kW?	(2.12) and (2.9) over the enumeration	posterior	10
7	**Expansion timing.** When does the next unit have to be ordered so that it is in before the probability of a shortfall passes 5 %?	(2.12) along the load path, a crossing date as in (2.21)	posterior	11
8	**Oversizing cost.** What is the probability that the plant as drawn is more than 30 % oversized at the ten-year load?	(2.12), the upper tail of a ratio	posterior	12
Where is the bottleneck				
9	**Binding constraint.** Which element binds first as the load grows: the chiller, the pump, the pipe, or the power feed?	(2.9) for each binding load; (2.12) on their order	posterior	13
10	**Debottleneck ranking.** If I could upgrade one element, which one buys the most headroom?	(2.9), an importance measure: a gradient	posterior	14
11	**Shared-path risk.** Two halls share one pipe; what is the probability that the shared pipe, not the halls' own units, is what limits them?	(2.12) over the enumeration, a cut-set reading	posterior	15
12	**Hidden margin.** The design says 20 % margin; what does the as-built model say the margin is?	(2.9), (2.12)	posterior	16
What does uncertainty in the design inputs cost				
13	**Parameter risk.** The conductance of the economizer path is a handbook guess; how much of the headroom band does that one guess carry?	(2.10), a variance budget	posterior	17
14	**What to measure before committing.** Which test or survey on the existing plant would most narrow the build decision?	(2.16), u a spread	preposterior	18
15	**Value of a pilot.** Would a one-hall pilot change the decision on the full build?	(2.16), u the indicator of a flipped decision	preposterior	19
Design under failure				
16	**Failure headroom.** With the largest unit out, what is the probability the remaining plant holds the inlets at the design load?	(2.12), a joint tail at the unit-out scenario	posterior	20
17	**Cascading.** If one unit trips, does the load shift onto the others push a second one past its envelope?	(2.12) along a scenario chain, a conditional probability	posterior	21
18	**Design for maintainability.** Can any one unit be taken out for service in any month of the year without a violation?	(2.12), the minimum over scenarios	posterior	22

No.	Planner's question	General form	Computation	Chapter
Expected cost and risk				
19	**Expected shortfall.** How many kWh of IT load would be shed per year, in expectation, under this design?	(2.23) over the enumeration	expected consequence	23
20	**Risk-cost tradeoff.** Where on the cost-versus-expected-shortfall curve does the design sit, and what is the band on that position?	(2.23) as a function of the design variable	expected consequence	24

Three readings of the table carry through the book. Questions 1 to 4 and 16 to 18 are the what-ifs of operations asked before the plant exists: each is a scenario read of one calibration, the scenario node of Chapter 4 pinned at a load, an ambient or a unit state not yet seen, on the fit that calibrated the existing plant or, for an element not yet built, on its handbook prior. Questions 5, 6, 11 and 19 are the reliability enumeration, the states of the units up or down with their rates, asked with a cost attached; there capacity planning and reliability engineering are one discipline with two readers, and each state of the enumeration carries availability and capacity as two columns, never one for the other. Questions 13 to 15 are about the uncertainty of the design inputs rather than the inputs themselves, and a spreadsheet cannot ask them at all, because it carries each input as one number; they are the questions that change what gets built.

2.4 Worked examples of estimation, prediction and joint events

Three pairs of questions are often confused in a design review: the estimate of a quantity and the probability of a threshold event, the estimate of an exceedance rate and the prediction of a design day within the committed load, a joint event in two halls and the product of two marginal probabilities. In each pair the two questions are put to the same model and differ in the form put to it.

2.4.1 Estimation of the conductance of the hall

The planner has the hall temperature measured at today's load of 30 kW and wants to know what the conductance is (question A) and what will happen at the design load of $Q_2 = 60\,\mathrm{kW}$ (question B). Step 2 of the workflow has extended the graph with the scenario node of Figure 2.1, the hall at the design load, whose temperature T_2 obeys the balance of equation (2.2),

$$T_2 = T_c + \frac{Q_2}{G}, \qquad T_c = 18^\circ\mathrm{C}, \quad Q_2 = 60\,\mathrm{kW}. \tag{2.28}$$

The conductance enters the state as the log-conductance $g = \ln(G/G_{\mathrm{ref}})$ with $G_{\mathrm{ref}} = 1\,\mathrm{kW/K}$ in this chapter's examples (Chapter 1 takes 6 kW/K; the reference only shifts g by a constant). For this example, suppose that a longer record of the hall has been summarized as one row on the log-conductance, the value $\hat{g}$ with the standard deviation σ_g,

$$r_g(\mathbf{z}) = \frac{g - \hat{g}}{\sigma_g}, \qquad \hat{g} = \ln 6.5 = 1.871802, \quad \sigma_g = 0.05. \tag{2.29}$$

These values are illustrative. The row stands for that longer record and is not the fit of Chapter 1, whose one reading of 0.5 K cannot give a spread of 5 %: there the MAP conductance is 5.917 kW/K and the probability of staying below 27.5°C at the design load is 0.246, the same question under different evidence. How readings are reduced to a row of this kind is the subject of later chapters. The row is the only one that acts on g alone, and the fit below returns the posterior of the third entry with center $\hat{g}$ and spread $\sqrt{\Sigma_{33}} = \sigma_g$.

Question A is an estimation question, of the general form (2.9). The quantity asked for is the conductance, a derived quantity of the log-conductance, $f(\mathbf{z}) = G = G_{\mathrm{ref}}\,\mathrm{e}^{g}$, with g the third entry of the state; it is increasing in that one entry. The answer is therefore read from three objects of the fit of the hall: the MAP state $\hat{\mathbf{z}}$, the Laplace covariance $\boldsymbol{\Sigma}$ and the gradient of f. The gradient has a single nonzero entry, $\mathbf{g} = (0,\ 0,\ G)^{\mathsf{T}}$, so the band $\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g} = G^2\,\Sigma_{33}$ uses only the third diagonal entry of $\boldsymbol{\Sigma}$; and because f is increasing in one entry, its quantiles are the images of that entry's quantiles under the exponential. When the third entry is exactly normal, as it is here, G is exactly lognormal, with median $G_{\mathrm{ref}}\,\mathrm{e}^{\hat{z}_3}$, mean $G_{\mathrm{ref}}\,\mathrm{e}^{\hat{z}_3+\Sigma_{33}/2}$ and central 95 % interval $G_{\mathrm{ref}}\,\mathrm{e}^{\hat{z}_3 \pm 1.960\sqrt{\Sigma_{33}}}$, and these exact summaries are reported beside the Laplace band. The fit of the state of the extended graph is the building block.

The posterior of the whole state of the extended graph follows from the row (2.29) and the two balance rows as a fit. The state holds the hall temperature T_1 at the present load $Q_1 = 30\,\mathrm{kW}$, the design-load temperature T_2 and the log-conductance,

$$\mathbf{z} = \begin{bmatrix} T_1 \\ T_2 \\ g \end{bmatrix}. \tag{2.30}$$

The rows are the energy balances of the two hall nodes, written in temperature form as in Chapter 1 and divided by a model tolerance σ_p in kelvin, and the earlier inference, which enters as a row on g of the same form as a prior row:

$$r(\mathbf{z}) = \begin{bmatrix} (T_1 - T_c - (Q_1/G_{\mathrm{ref}})\,\mathrm{e}^{-g})/\sigma_p \\ (T_2 - T_c - (Q_2/G_{\mathrm{ref}})\,\mathrm{e}^{-g})/\sigma_p \\ (g - \hat{g})/\sigma_g \end{bmatrix}, \qquad \mathcal{L}(\mathbf{z}) = \tfrac{1}{2}\,r(\mathbf{z})^{\mathsf{T}} r(\mathbf{z}). \tag{2.31}$$

The exact balance of equation (2.28) is the limit $\sigma_p \to 0$. The Jacobian has one row per row of r and one column per entry of $\mathbf{z}$, so it is 3×3; each entry is the sensitivity of a row to an entry of the state, in standard deviations of the row per unit of the entry:

$$\mathbf{J} = \frac{\partial r}{\partial \mathbf{z}} = \begin{bmatrix} 1/\sigma_p & 0 & (Q_1/G_{\mathrm{ref}})\,\mathrm{e}^{-g}/\sigma_p \\ 0 & 1/\sigma_p & (Q_2/G_{\mathrm{ref}})\,\mathrm{e}^{-g}/\sigma_p \\ 0 & 0 & 1/\sigma_g \end{bmatrix} = \begin{bmatrix} 100 & 0 & 461.5 \\ 0 & 100 & 923.1 \\ 0 & 0 & 20 \end{bmatrix}$$

at the most probable state with $\sigma_p = 0.01\,\mathrm{K}$, in K^{-1} in the first two columns. The third column says that a unit of g moves the design-load balance twice as far as the present-load balance, because the design load is twice the present load. The normal equations $(\mathbf{J}^{\mathsf{T}}\mathbf{J})\,\Delta\mathbf{z} = -\mathbf{J}^{\mathsf{T}} r$ give the Gauss–Newton step. The table below lists the iterations from the starting state $(23^{\circ}\mathrm{C},\ 28^{\circ}\mathrm{C},\ \ln 6)$, the forward solution of both balances at $G = 6\,\mathrm{kW/K}$: the columns are the iteration, the three entries and the objective.

iteration [–]	T_1 [°C]	T_2 [°C]	g [–]	$\mathcal{L}$ [–]
0	23.0000	28.0000	1.7918	1.281
1	22.5998	27.1996	1.8718	6.083
2	22.6154	27.2308	1.8718	6×10^{-18}
3	22.6154	27.2308	1.8718	2×10^{-26}

At the start both balances hold, so the objective 1.281 is the prior row's alone, $\frac{1}{2}\big((\ln 6 - \ln 6.5)/0.05\big)^2$. The row on g is linear, so the first step puts g at $\hat{g}$; the objective rises at that step because the step is taken on the linearization of e^{-g}, and the balance residuals it leaves, -0.016 and $-0.031\,\mathrm{K}$, are magnified a hundredfold by the division by σ_p. The second step closes both balances. At the most probable state

$$\hat{\mathbf{z}} = \begin{bmatrix} 22.615^{\circ}\mathrm{C} \\ 27.231^{\circ}\mathrm{C} \\ 1.871802 \end{bmatrix}$$

every row is zero, $T_1 = 18 + 30/6.5$ and $T_2 = 18 + 60/6.5$.

For rows divided by their standard deviations, as here, the information matrix is $\mathbf{\Lambda} = \mathbf{J}^\mathsf{T}\mathbf{J}$. When the rows are left in their physical units, with Jacobian $\mathbf{J}_u$ and weights $\mathbf{W} = \mathrm{diag}(1/\sigma_k^2)$, it is $\mathbf{J}_u^\mathsf{T}\mathbf{W}\mathbf{J}_u$, the same matrix, since $\mathbf{J} = \mathbf{W}^{1/2}\mathbf{J}_u$. Its inverse is the Laplace covariance of the state. The curvature term $\sum_k r_k\, \partial^2 r_k/\partial \mathbf{z}^2$ that Newton's method would add vanishes here, because every row is zero at $\hat{\mathbf{z}}$, so

$$\mathbf{\Sigma} = (\mathbf{J}^\mathsf{T}\mathbf{J})^{-1} = \begin{bmatrix} 0.05335 & 0.10651 & -0.01154 \\ 0.10651 & 0.21312 & -0.02308 \\ -0.01154 & -0.02308 & 0.00250 \end{bmatrix} \qquad (\sigma_p = 0.01\,\mathrm{K}). \tag{2.32}$$

The ridge of the posterior. The eigenvalues of this $\mathbf{\Sigma}$ are 0.2689, 1.0×10^{-4} and 9.3×10^{-7}: one direction carries almost all of the uncertainty, and the other two shrink with σ_p^2. In the limit $\sigma_p \to 0$ the covariance is a single outer product of the vector $\mathbf{v} = (-Q_1/G,\ -Q_2/G,\ 1)^\mathsf{T}$,

$$\mathbf{\Sigma} \to \sigma_g^2\, \mathbf{v}\mathbf{v}^\mathsf{T} = \begin{bmatrix} 0.05325 & 0.10651 & -0.01154 \\ 0.10651 & 0.21302 & -0.02308 \\ -0.01154 & -0.02308 & 0.00250 \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} -4.615 \\ -9.231 \\ 1 \end{bmatrix}, \tag{2.33}$$

where $\mathbf{v}$ is the direction along which both balance rows stay zero (it is orthogonal to the first two rows of $\mathbf{J}$) and σ_g^2 the variance the third row leaves along it. The three entries move as one along $\mathbf{v}$: a unit rise of g cools the hall by 4.615 K at the present load and 9.231 K at the design load. The limit is reached quickly: $\sqrt{\Sigma_{22}}$ is 1.1014 K at $\sigma_p = 1$ K, 0.4722 K at 0.1 K, 0.46165 K at 0.01 K, 0.46154 K at 0.001 K and 0.461538 K at 10^{-4} K, and $\sqrt{\Sigma_{33}} = 0.05$ at every σ_p. The answers below are read off the limit.

The fit supplies the objects that question A was reduced to. The log-conductance is the third entry of the state, $\hat{z}_3 = 1.871802$ with $\sqrt{\Sigma_{33}} = 0.05$, and the conductance has the gradient and band

$$\mathbf{g} = \frac{\partial G}{\partial \mathbf{z}} = \begin{bmatrix} 0 \\ 0 \\ G \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 6.5 \end{bmatrix} \mathrm{kW/K}, \qquad \sqrt{\mathbf{g}^\mathsf{T}\mathbf{\Sigma}\mathbf{g}} = 6.5 \times 0.05 = 0.325\,\mathrm{kW/K},$$

so that the reported answer is $G = 6.5 \pm 0.325$ kW/K, with the central 95 % band $6.5 \pm 1.960 \times 0.325$, from 5.863 to 7.137 kW/K. The exact lognormal summaries follow from the normal of the third entry and stand beside it: the median is $G_{\mathrm{ref}}\, \mathrm{e}^{\hat{z}_3} = 6.5$ kW/K, the location above; the mean, by the moment-generating function of the normal, is $G_{\mathrm{ref}}\, \mathrm{e}^{\hat{z}_3 + \Sigma_{33}/2} = 6.5\, \mathrm{e}^{0.00125} = 6.508$ kW/K; the standard deviation is 0.3256 kW/K, a gap of 0.0006 from the band; and the central 95 % credible interval is $6.5\, \mathrm{e}^{\pm 1.960 \times 0.05}$, from 5.893 to 7.169 kW/K, whose ends lie 0.030 and 0.032 kW/K above those of the band. Each of these is a summary of a distribution; the answer to question A is the distribution itself, equation (2.9) with $f(\mathbf{z}) = G$, and any single number reported from it must say which summary it is.

2.4.2 The threshold probability of the hall at the design load

Question B asks whether the hall will remain below $T_{\max} = 27.5$°C at the design load, that is, for

$$\Pr(T_2 < 27.5^\circ\mathrm{C} \mid \mathbf{y}).$$

It is a prediction under a threshold, of the general form (2.12). The requirement is the upper limit $T_2 < T_{\max}$ on the second entry of the state, so the form applies with $f(\mathbf{z}) = -T_2 = -z_2$ and $f^* = -T_{\max}$; equivalently, the probability of the requirement is one minus that of the exceedance, $1 - \Pr(T_2 \geq T_{\max} \mid \mathbf{y})$, and the assurance rule is applied to the requirement. The

plant enters through the scenario node: the event is a bound on an entry that no sensor reads, tied to the third entry by the design-load balance row. The tail is read from the fit. Under the Laplace approximation the second entry is normal with center $\hat{z}_2$ and variance Σ_{22}, the second diagonal entry of $\mathbf{\Sigma}$, so the probability is a normal tail in those two numbers, read off $\hat{\mathbf{z}}$ and the limit of equation (2.33), and this is the value reported:

$$\Pr(T_2 < T_{\max} \mid \mathbf{y}) \approx \Phi\left(\frac{T_{\max} - \hat{z}_2}{\sqrt{\Sigma_{22}}}\right) = \Phi\left(\frac{27.5 - 27.231}{0.4615}\right) = \Phi(0.5833) = 0.720. \tag{2.34}$$

The exact value along the ridge. An exact value stands beside it, through the ridge. Along $\mathbf{v}$ the design-load temperature is a decreasing function of the third entry, $T_2 = T_c + (Q_2/G_{\text{ref}})\,\mathrm{e}^{-g}$, so the bound on T_2 is a bound on g, at the conductance where the second row meets the limit,

$$\Pr(T_2 < 27.5^{\circ}\mathrm{C} \mid \mathbf{y}) = \Pr(G > G^* \mid \mathbf{y}), \qquad G^* = \frac{Q_2}{T_{\max} - T_c} = 6.3158\,\mathrm{kW/K}, \tag{2.35}$$

that is $g > \ln(G^*/G_{\text{ref}}) = 1.843053$; and the third entry is exactly normal, so

$$\Pr(T_2 < 27.5^{\circ}\mathrm{C} \mid \mathbf{y}) = \Phi\left(\frac{\hat{z}_3 - \ln(G^*/G_{\text{ref}})}{\sqrt{\Sigma_{33}}}\right) = \Phi\left(\frac{1.871802 - 1.843053}{0.05}\right), \tag{2.36}$$

$$\Pr(T_2 < 27.5^{\circ}\mathrm{C} \mid \mathbf{y}) = \Phi(0.5750) = 0.717. \tag{2.37}$$

The normal tail of the entry and the exact value differ by 0.003, because T_2 is not linear in g along the ridge. The probability that the hall stays below the limit at the design load is 72.0 %, with 71.7 % exact beside it, and the probability that it exceeds the limit is 28.0 %, with 28.3 % exact, conditional on the row of equation (2.29), the design conditions and the steady-state thermal model. Whether a 28 % probability of exceeding the limit is acceptable is not a probability question; it depends on the consequences of exceedance and the design criteria that apply. The assurance rule of equation (2.12) states such a criterion as a number, applied to the probability that the requirement holds: at an assurance level $\kappa = 0.95$ the requirement $T_2 < 27.5^{\circ}\mathrm{C}$ is not accepted by either value, since 0.720 and 0.717 both lie below 0.95.

Figure 2.2 shows the two questions side by side. In panel (a) the horizontal axis is the conductance G and the vertical axis its posterior density; the shaded area to the right of the dashed line at G^* is the exact 0.717. In panel (b) the horizontal axis is the design-load temperature T_2 and the vertical axis its posterior density; the solid curve is the exact density implied by the lognormal G, the dashed curve the normal of the entry of equation (2.34), whose area to the left of the limit is the reported 0.720, marked on the panel, and the shaded area under the solid curve to the left of the limit is the same exact 0.717. The two shaded areas are one probability, seen once on the conductance and once on the temperature: the event $T_2 < 27.5^{\circ}\mathrm{C}$ and the event $G > G^*$ are the same set of states.

The check against the exact value and against draws. Equation (2.37) is exact under the normal posterior of g, because the map from g to T_2 is monotone along the ridge; equation (2.34) is the route that every derived quantity has. Had T_2 not been carried in the state, the same band would come from the gradient of $f(\mathbf{z}) = T_c + (Q_2/G_{\text{ref}})\,\mathrm{e}^{-g}$, $\mathbf{g} = (0,\ 0,\ -9.231)^{\mathsf{T}}\,\mathrm{K}$, and it agrees with Σ_{22} because in the limit of equation (2.33) $\mathbf{g}^{\mathsf{T}}\mathbf{\Sigma}\mathbf{g} = \sigma_g^2(\mathbf{g}^{\mathsf{T}}\mathbf{v})^2$ and $\mathbf{g}$ has the same projection on $\mathbf{v}$, -9.231, as the unit vector on T_2. Table 2.6 compares the reported and the exact probabilities for the limit of the example and three others. The first column is the limit, the second the equivalent conductance threshold of equation (2.35), the third the standardized distance in g, and the last two the reported normal of the entry, with mean $\hat{z}_2$ and variance Σ_{22}, and the exact probability. The two differ by at most 0.007 over these limits and by 0.003 at 27.5°C, because the curvature of $60/G$ across a posterior of 5 % spread is small; panel (b)

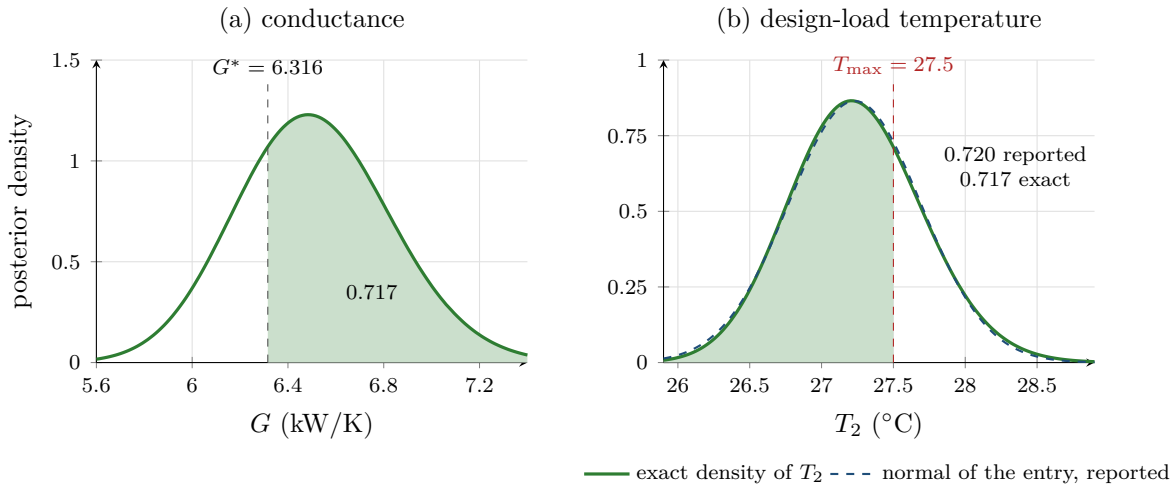


Figure 2.2: One posterior, two questions. (a) Lognormal posterior of the conductance, density $(G\,s\sqrt{2\pi})^{-1}\exp[-(\ln G - \ln 6.5)^2/(2s^2)]$ with $s = \sqrt{\Sigma_{33}} = 0.05$; shaded, $\Pr(G > 6.316\,\text{kW/K} \mid \mathbf{y}) = 0.717$. (b) Posterior of the design-load temperature $T_2 = 18 + 60/G$, lognormal in $T_2 - 18$ with median 9.231 K and log-spread 0.05 (solid), and the normal of the entry, $\mathcal{N}(\hat{z}_2, \Sigma_{22}) = \mathcal{N}(27.231, 0.4615^2)$ (dashed), whose tail is the reported 0.720; shaded, the exact $\Pr(T_2 < 27.5°\text{C} \mid \mathbf{y}) = 0.717$.

of Figure 2.2 shows the slight right skew the normal misses. Small as it is, the gap can cross a decision level: at $T_{\max} = 28°\text{C}$ and $\kappa = 0.95$ the normal gives 0.9522, which accepts the requirement, and the exact value 0.9453, which does not. Both values are then stated, and the planner sees that the requirement sits on the level. A second check draws 200 000 values of g from its posterior, computes T_2 for each, and counts the fraction below the limit: 0.7175, against the exact 0.7174, within the sampling error of 0.001.

Table 2.6: The threshold probability $\Pr(T_2 < T_{\max} \mid \mathbf{y})$ at $Q_2 = 60\,\text{kW}$, $T_c = 18°\text{C}$, $g \mid \mathbf{y} \sim \mathcal{N}(\ln 6.5, 0.05^2)$: reported, by the normal $\mathcal{N}(\hat{z}_2, \Sigma_{22})$ of the entry T_2 (equation (2.34)), and exact (equation (2.36)).

$T_{\max}$ [°C]	$G^* = Q_2/(T_{\max} - T_c)$ [kW/K]	$(\hat{z}_3 - \ln G^*)/\sqrt{\Sigma_{33}}$ [–]	reported, normal of T_2 [–]	exact [–]
27.0	6.6667	−0.5064	0.3085	0.3063
27.5	6.3158	0.5750	0.7202	0.7174
28.0	6.0000	1.6009	0.9522	0.9453
28.5	5.7143	2.5767	0.9970	0.9950

What the two questions share. Both questions used the same posterior. Estimation, $p(G \mid \mathbf{y})$, describes the uncertainty about the conductance; the threshold probability, $\Pr(T_2 < T_{\max} \mid \mathbf{y})$, integrates that uncertainty over the states that satisfy the temperature limit. Neither required a new posterior; each asked a different question of the same one. That reuse is one of the principal advantages of a probabilistic model of the plant: the expensive step, the posterior, is paid once, and each new question is an integral over it.

2.4.3 The posterior of the hall's committed-load exceedance rate

A different record of the same hall: the log of its committed-load exceedances, the hours in which a workload peak, a batch run or a test of new racks, lifted the hall's metered heat load above its committed load. Over the last 1000 h of logged operation, about six weeks, the load exceeded the committed load twice, and the planner asks for the probability that the twelve peak hours of a design day pass within the committed load. This is a prediction question, not only a question about the exceedance rate. Let λ be the unknown exceedance rate, in exceedances per hour, $k = 2$ the observed number of exceedances, $T_{\text{tot}} = 1000\,\text{h}$ the logged operation behind the count, and $\tau = 12\,\text{h}$ the peak hours of the design day. The exceedances are taken to follow a homogeneous Poisson process, so that the count in the logged time is Poisson with mean λT_{tot},

$$N(T_{\text{tot}}) \mid \lambda \sim \text{Poisson}(\lambda T_{\text{tot}}), \qquad \Pr(k \mid \lambda, T_{\text{tot}}) = \frac{(\lambda T_{\text{tot}})^k \, \mathrm{e}^{-\lambda T_{\text{tot}}}}{k!}.$$

This is a modeling choice. It is what lets the log be summarized by k and T_{tot} alone, and it fails when one peak produces several exceedances (a load that hovers about the committed load) or when the exceedances crowd into the peak hours of the day, so that the rate over the peak hours exceeds the rate over all logged hours; the count here is of peaks, a load hovering about the committed load counted once. As a function of λ, the likelihood is proportional to $\lambda^k \, \mathrm{e}^{-\lambda T_{\text{tot}}}$; the factor $T_{\text{tot}}^k/k!$ does not depend on the rate. The prior is a gamma distribution in the shape–rate parameterization, $\lambda \sim \text{Gamma}(a, b)$, with density proportional to $\lambda^{a-1} \, \mathrm{e}^{-b\lambda}$ and, here,

$$a = 2, \qquad b = 500\,\text{h},$$

the exceedance logs of the site's other halls, whose prior mean $a/b = 0.004\,\text{h}^{-1}$ is one exceedance every 250 h. Bayes' rule gives the posterior up to a constant as likelihood times prior, $p(\lambda \mid k, T_{\text{tot}}) \propto \Pr(k \mid \lambda, T_{\text{tot}})\, p(\lambda)$.

The rate enters the state, as every rate does in this book, through its log-rate $\omega = \ln(\lambda/\lambda_{\text{ref}})$ with $\lambda_{\text{ref}} = 0.001\,\text{h}^{-1}$, so that the state is the one number

$$\omega = \ln \frac{\lambda}{\lambda_{\text{ref}}}, \qquad \lambda = \lambda_{\text{ref}} \, \mathrm{e}^{\omega}.$$

In the general forms $\mathbf{z}$ is here the one number ω, and every quantity of the fit is a number. The rate is asked for first, an estimation question of the general form (2.9) with $f(\omega) = \lambda = \lambda_{\text{ref}} \, \mathrm{e}^{\omega}$, increasing in ω. It is read from the MAP log-rate $\hat{\omega}$, its variance σ_{ω}^2 and the derivative $\mathrm{d}\lambda/\mathrm{d}\omega = \hat{\lambda}$, whose band is $\hat{\lambda}\,\sigma_{\omega}$; the exact gamma posterior that the objective identifies is set beside it. Neither row is Gaussian, and each enters the objective as its negative log-probability term. The count row is the negative log of the Poisson probability; the prior row is the negative log of the gamma density carried over to ω, where the change of variable multiplies the density by $\mathrm{d}\lambda/\mathrm{d}\omega = \lambda$ and so adds $-\ln\lambda$ (the rule, that a density carried to a new coordinate is multiplied by the derivative of the old coordinate with respect to the new, is derived in Section 4.3.3):

$$\mathcal{L}(\omega) = \underbrace{T_{\text{tot}}\lambda - k\ln\lambda}_{\text{count row, } \mathcal{L}_c} + \underbrace{b\lambda - a\ln\lambda}_{\text{prior row, } \mathcal{L}_0} + \text{const.} \tag{2.38}$$

With primes for derivatives in ω and $\mathrm{d}\lambda/\mathrm{d}\omega = \lambda$, each row has a slope and a curvature, and they add:

$$\mathcal{L}_c' = T_{\text{tot}}\lambda - k, \quad \mathcal{L}_c'' = T_{\text{tot}}\lambda, \qquad \mathcal{L}_0' = b\lambda - a, \quad \mathcal{L}_0'' = b\lambda, \qquad \mathcal{L}' = \mathcal{L}_c' + \mathcal{L}_0', \quad \mathcal{L}'' = \mathcal{L}_c'' + \mathcal{L}_0''.$$

The Newton step $\Delta\omega = -\mathcal{L}'/\mathcal{L}''$ moves to the minimum of the parabola with the current slope and curvature. The table below lists the iterations from the maximum likelihood rate

$k/T_{\text{tot}} = 0.002\,\text{h}^{-1}$, that is $\omega = \ln 2$, where the count row's slope is zero: the columns are the iteration, the log-rate, the rate, the slopes of the two rows, the slope and curvature of the objective, and the step.

iteration j [–]	ω_j [–]	λ [h^{-1}]	$\mathcal{L}'_c$ [–]	$\mathcal{L}'_0$ [–]	$\mathcal{L}'$ [–]	$\mathcal{L}''$ [–]	$\Delta\omega$ [–]
0	0.69315	0.002000	0.0000	−1.0000	−1.0000	3.0000	0.33333
1	1.02648	0.002791	0.7912	−0.6044	0.1868	4.1868	−0.04462
2	0.98186	0.002669	0.6694	−0.6653	0.0041	4.0041	−0.00103
3	0.98083	0.002667	0.6667	−0.6667	2×10^{-6}	4.0000	-5×10^{-7}

The slope vanishes at $\lambda = (a + k)/(b + T_{\text{tot}})$, so the mode is $\hat{\omega} = \ln(8/3) = 0.98083$ and $\lambda_{\text{ref}}\, \text{e}^{\hat{\omega}} = 4/1500\,\text{h} = 0.002667\,\text{h}^{-1}$. The curvature there is the information, $\Lambda = \mathcal{L}''(\hat{\omega}) = T_{\text{tot}}\hat{\lambda} + b\hat{\lambda} = 2.667 + 1.333 = 4 = a + k$: the count row contributes the expected number of exceedances in the logged time, the prior row $b\hat{\lambda}$, and the two add as the information of independent rows does. Then $\sigma_\omega^2 = 1/\Lambda = 0.25$, the log-rate has $\sigma_\omega = 0.5$, and the rate, a derived quantity with derivative $\text{d}\lambda/\text{d}\omega = \hat{\lambda}$, has $\hat{\lambda}\,\sigma_\omega = 0.002667 \times 0.5 = 0.001333\,\text{h}^{-1}$. The reported answer is the exceedance rate $0.002667 \pm 0.001333\,\text{h}^{-1}$.

The objective also identifies the exact posterior. Back in λ, without the $-\ln\lambda$ of the change of variable, the negative log posterior is $(b + T_{\text{tot}})\lambda - (a + k - 1)\ln\lambda$, the negative log of the kernel $\lambda^{a+k-1}\,\text{e}^{-(b+T_{\text{tot}})\lambda}$ of a gamma density. Therefore

$$\lambda \mid k, T_{\text{tot}} \sim \text{Gamma}(a', b'), \qquad a' = a + k, \quad b' = b + T_{\text{tot}}. \tag{2.39}$$

For the hall's exceedances, $a' = 2 + 2 = 4$ and $b' = 500 + 1000 = 1500\,\text{h}$, so $\lambda \mid \mathbf{y} \sim \text{Gamma}(4, 1500\,\text{h})$. The posterior mean and standard deviation are

$$\mathbb{E}[\lambda \mid \mathbf{y}] = \frac{a'}{b'} = \frac{4}{1500\,\text{h}} = 0.002667\,\text{h}^{-1}, \qquad \sqrt{\text{Var}[\lambda \mid \mathbf{y}]} = \frac{\sqrt{a'}}{b'} = 0.001333\,\text{h}^{-1},$$

Set like with like, the reported posterior and the exact one differ, because the Laplace normal of the log-rate makes the rate lognormal: its median is $0.002667\,\text{h}^{-1}$ against the gamma's $0.002448\,\text{h}^{-1}$, a gap of $0.000219\,\text{h}^{-1}$; its mean $\hat{\lambda}\,\text{e}^{\sigma_\omega^2/2} = 0.003022\,\text{h}^{-1}$ against the gamma's $0.002667\,\text{h}^{-1}$, a gap of $0.000355\,\text{h}^{-1}$; and its central 95 % interval runs from 0.001001 to $0.007105\,\text{h}^{-1}$ against the gamma's 0.000727 to $0.005845\,\text{h}^{-1}$. The pair the Newton fit returned coincides with the gamma's mean and standard deviation: the mode of the log-rate maps onto the gamma mean and the linearized band onto the gamma standard deviation, for every gamma posterior, but that compares different summaries, a median of one posterior with a mean of the other. The gamma mean is one exceedance every 375 h. It lies between the prior mean $a/b = 0.004\,\text{h}^{-1}$, one every 250 h, and the maximum likelihood estimate $k/T_{\text{tot}} = 0.002\,\text{h}^{-1}$, one every 500 h, as the weighted form $a'/b' = \frac{b}{b'}\frac{a}{b} + \frac{T_{\text{tot}}}{b'}\frac{k}{T_{\text{tot}}} = \frac{1}{3}(0.004) + \frac{2}{3}(0.002)$ shows. An estimation question has been answered. The planner's question, a design day within the committed load, has not.

2.4.4 The probability of a design day within the committed load

Let T be the time to the next exceedance of the committed load, counted from the start of the design day's peak hours. (In this example T, in hours, is a time; in the hall example, T_1 and T_2 are hall temperatures in °C. The units distinguish them.) Under the Poisson process the time to the next exceedance does not depend on how long ago the last one was, so the question is

$$\Pr(T > \tau \mid \mathbf{y}), \qquad \tau = 12\,\text{h}. \tag{2.40}$$

For a known constant rate the probability of no exceedance is exponential, $\Pr(T > \tau \mid \lambda) = \text{e}^{-\lambda\tau}$. The rate is not known, and the event is a future one that stays random even when the rate is

known, so the question is a prediction of the general form (2.13), in its event form (2.8): y_{new} is the event $T > \tau$, and $p(y_{\text{new}} \mid \mathbf{z}) = \mathrm{e}^{-\lambda\tau}$ with $\lambda = \lambda_{\text{ref}}\,\mathrm{e}^{\omega}$. The integral is a function of the one number ω averaged over its posterior, and under the Laplace approximation that posterior is $\mathcal{N}(\hat{\omega}, \sigma_\omega^2)$, so the value read from the state is a normal expectation in $\hat{\omega}$ and σ_ω^2:

$$\Pr(T > \tau \mid \mathbf{y}) \approx \mathbb{E}_{\omega\sim\mathcal{N}(\hat{\omega},\sigma_\omega^2)}\big[\mathrm{e}^{-\lambda_{\text{ref}}\mathrm{e}^{\omega}\tau}\big] \approx \sum_{j=1}^{5} w_j\,\mathrm{e}^{-\lambda_{\text{ref}}\mathrm{e}^{\omega_j}\tau}, \qquad \omega_j = \hat{\omega} + \sqrt{2}\,\sigma_\omega\,x_j. \tag{2.41}$$

The sum is the five-point Gauss–Hermite rule, which replaces a normal expectation by a weighted sum at fixed nodes x_j with weights w_j and is exact for a polynomial of degree nine. The table below evaluates it for the design day's twelve peak hours, with $\hat{\omega} = 0.98083$ and $\sigma_\omega = 0.5$: the columns are the node, its weight, the log-rate and rate there, the probability of no exceedance at that rate, and the weighted term.

x_j [–]	w_j [–]	ω_j [–]	λ_j [h^{-1}]	$\mathrm{e}^{-12\lambda_j}$ [–]	term [–]
−2.02018	0.01126	−0.44766	0.000639	0.99236	0.01117
−0.95857	0.22208	0.30302	0.001354	0.98388	0.21850
0	0.53333	0.98083	0.002667	0.96851	0.51654
0.95857	0.22208	1.65864	0.005252	0.93892	0.20851
2.02018	0.01126	2.40931	0.011126	0.87501	0.00985

The terms sum to 0.9646, and a direct numerical integration agrees to six decimals. The predicted probability that the design day's peak hours pass within the committed load is 96.46 %, and the probability of at least one exceedance is 0.0354; these are the values reported.

The exact value stands beside them, because the objective identified the posterior as a gamma and the same integral then has a closed form. The conditional probability of no exceedance is averaged over the posterior of the rate,

$$\Pr(T > \tau \mid \mathbf{y}) = \int_0^\infty \mathrm{e}^{-\lambda\tau}\,p(\lambda \mid \mathbf{y})\,\mathrm{d}\lambda.$$

Substitute the density of the gamma posterior of equation (2.39),

$$p(\lambda \mid \mathbf{y}) = \frac{b'^{\,a'}}{\Gamma(a')}\,\lambda^{a'-1}\,\mathrm{e}^{-b'\lambda},$$

and combine the two exponentials:

$$\Pr(T > \tau \mid \mathbf{y}) = \frac{b'^{\,a'}}{\Gamma(a')}\int_0^\infty \lambda^{a'-1}\,\mathrm{e}^{-(b'+\tau)\lambda}\,\mathrm{d}\lambda.$$

The remaining integral is the standard gamma integral, $\int_0^\infty \lambda^{a'-1}\mathrm{e}^{-c\lambda}\,\mathrm{d}\lambda = \Gamma(a')/c^{a'}$, obtained from the definition of the gamma function by the substitution $x = c\lambda$; with $c = b' + \tau$ the gamma functions cancel, and

$$\Pr(T > \tau \mid \mathbf{y}) = \left(\frac{b'}{b'+\tau}\right)^{a'}. \tag{2.42}$$

For the twelve peak hours,

$$\Pr(T > 12\,\mathrm{h} \mid \mathbf{y}) = \left(\frac{1500}{1512}\right)^{4} = 0.9686,$$

and the probability of at least one exceedance is 0.0314. The calculation needed no Gaussian or Laplace approximation: with a gamma posterior and an exponential probability of no

exceedance, the predictive probability has an exact closed form. The gap between the two is 0.00406 for the design day, and over the next 100 h, about eight design days, the state gives 0.748 against the exact 0.7725, a gap of 0.024. The state route errs low because the lognormal it implies for λ has the larger mean $0.00302\,\mathrm{h}^{-1}$ and a heavier upper tail than the gamma. No decision of this chapter sits between the two values; an assurance level of 0.965 on the design day would, and both would then be stated, the one read from the state failing it and the exact one meeting it.

2.4.5 The error of the plug-in exceedance rate

A common shortcut substitutes the posterior mean rate into the exponential:

$$\Pr_{\text{plug}}(T > 12\,\mathrm{h}) = \mathrm{e}^{-\mathbb{E}[\lambda|\mathbf{y}]\times 12\,\mathrm{h}} = \mathrm{e}^{-0.0320} = 0.9685,$$

against the exact predictive 0.9686; over the next 100 h it gives $\mathrm{e}^{-0.2667} = 0.7659$ against 0.7725. The two differ because averaging and exponentiation do not commute,

$$\mathbb{E}\big[\mathrm{e}^{-\lambda\tau} \mid \mathbf{y}\big] \neq \mathrm{e}^{-\mathbb{E}[\lambda|\mathbf{y}]\,\tau} \quad \text{in general,} \tag{2.43}$$

since the probability of no exceedance is a nonlinear function of the rate. The direction of the difference follows from the shape of that function. For $\tau > 0$ the map $\lambda \mapsto \mathrm{e}^{-\lambda\tau}$ has second derivative $\tau^2\mathrm{e}^{-\lambda\tau} > 0$, so it is convex, and Jensen's inequality, the standard result that the expectation of a convex function is at least the function of the expectation, gives $\mathbb{E}[\mathrm{e}^{-\lambda\tau} \mid \mathbf{y}] \geq \mathrm{e}^{-\mathbb{E}[\lambda|\mathbf{y}]\tau}$: the plug-in value understates the exact probability of a design day within the design load. The low rates the posterior admits raise that probability by more than the high rates lower it. In the log-rate the same function is concave while $\lambda\tau < 1$, so the reported average (2.41) starts below the plug-in and crosses it at 356 h. For one design day the plug-in sits 0.0001 from the exact value, closer than the reported 0.9646. The plug-in value is wrong not because it is far from the predictive value, but because it answers a different question: the probability of a design day within the committed load for a hall whose exceedance rate is known to be $0.002667\,\mathrm{h}^{-1}$, which is not the hall in question, and it carries no band.

Table 2.7 shows how the values change with the horizon. The first column is the horizon; the second the reported average of equation (2.41); the third the exact predictive probability of no exceedance, equation (2.42); the fourth their gap; the fifth the plug-in value at the posterior mean; the sixth the exact value minus the plug-in; and the last the plug-in value at the maximum likelihood estimate $0.002\,\mathrm{h}^{-1}$, which ignores the prior as well as the uncertainty. For one design day or one whole day the exact and plug-in values agree to three decimals. Their ratio grows with the horizon, from 1.0005 at 24 h to 1.87 at 1000 h and 7.0 at 2000 h; the absolute difference peaks near 800 h, at 0.0625, and then falls, to 0.0289 at 2000 h, because both values tend to zero; the gap of the reported average peaks in the hundreds of hours, at 0.0387 among the rows of the table. Figure 2.3 draws the ingredients. In panel (a) the horizontal axis is the exceedance rate in exceedances per 1000 h and the vertical axis the probability density per unit of that axis; the blue curve is the prior, the green curve the posterior, and the dotted lines mark the maximum likelihood estimate, the posterior mean and the prior mean. In panel (b) the horizontal axis is the horizon and the vertical axis the probability of no exceedance; the green curve is the exact predictive, the blue dash-dotted curve the reported average over the normal of the log-rate and the dashed red curve the plug-in, and the vertical dotted line marks the 12 h of the design day's peak. The exact curve lies above the plug-in curve at every horizon, and their ratio grows as the horizon lengthens, because the longer the horizon, the more the probability of no exceedance bends across the spread of the posterior; the blue curve falls below both at first and crosses the plug-in at 356 h.

Table 2.7: Probability of no exceedance of the hall's committed load over a horizon τ: the reported average over the normal of the log-rate (equation (2.41), $\hat{\omega} = \ln(8/3)$, $\sigma_\omega = 0.5$); exact, from the posterior Gamma$(4, 1500\,\text{h})$, $(1500/(1500+\tau))^4$; plug-in $\mathrm{e}^{-4\tau/1500}$; plug-in at the maximum likelihood estimate $\mathrm{e}^{-0.002\tau}$. Gaps are computed before rounding.

τ [h]	reported, from the state [–]	exact [–]	gap [–]	plug-in, posterior mean [–]	exact minus plug-in [–]	plug-in, maximum likelihood [–]
12	0.9646	0.9686	0.0041	0.9685	0.0001	0.9763
24	0.9307	0.9385	0.0078	0.9380	0.0005	0.9531
100	0.7481	0.7725	0.0244	0.7659	0.0065	0.8187
500	0.2777	0.3164	0.0387	0.2636	0.0528	0.3679
1000	0.1015	0.1296	0.0281	0.0695	0.0601	0.1353
2000	0.0205	0.0337	0.0132	0.0048	0.0289	0.0183

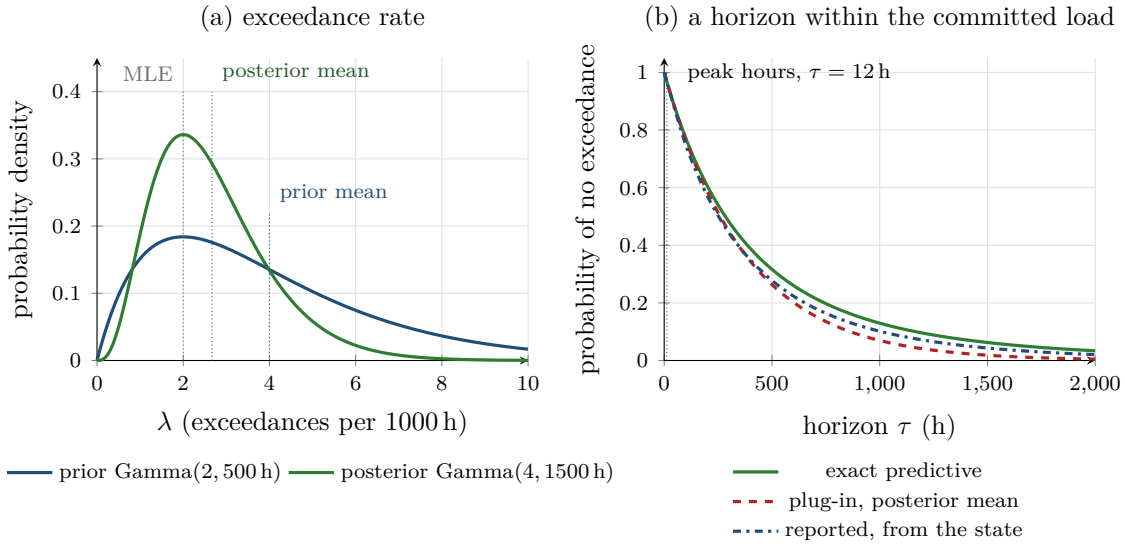


Figure 2.3: The hall's committed-load exceedance rate and the probability of a horizon without exceedance. (a) Prior $0.25\,x\,\mathrm{e}^{-0.5x}$ and posterior $(1.5^4/3!)\,x^3\mathrm{e}^{-1.5x}$ in $x = 1000\lambda$; dotted lines at the maximum likelihood estimate 2.0, the posterior mean 2.67 and the prior mean 4.0 exceedances per 1000 h. (b) Exact predictive probability of no exceedance $(1500/(1500+\tau))^4$ of equation (2.42), plug-in $\mathrm{e}^{-4\tau/1500}$ and the reported average of equation (2.41); at the 12 h of the design day's peak they are 0.9686, 0.9685 and 0.9646, at 100 h 0.7725, 0.7659 and 0.7481.

Principle 2.2. *Predicting a design day within the committed load requires averaging the probability of no exceedance over the uncertainty in the exceedance rate. Substituting a point estimate of the rate answers the question for a different hall, one whose rate is known.*

2.4.6 Two halls within their committed loads through the peak as a joint event

A site has two other halls, A and B, each with its own committed load, and one chilled-water plant that carries the heat of both. Figure 2.4 draws the arrangement: the white box on the left is the site's workload, the IT jobs scheduled across both halls; the two green boxes are the halls, each with its own causes of an exceedance (a rack deployment, a local batch run, a load test); the white box on the right is the chiller plant that receives the heat of both; the solid arrows are the paths along which load and heat pass. Define the events

$$A_A = \{\text{hall A's load exceeds its committed load in the day's peak}\},$$
$$A_B = \{\text{hall B's load exceeds its committed load in the day's peak}\},$$

and write $\bar{A}_A$ and $\bar{A}_B$ for their complements, each hall within its committed load. The planner wants both halls within their committed loads, the intersection $\bar{A}_A \cap \bar{A}_B$, and fears the opposite corner, both halls above their committed loads in the same peak, $A_A \cap A_B$, two excess loads the shared chiller plant must carry at once. The two are tied by inclusion and exclusion, which holds without any independence assumption:

$$\Pr(\bar{A}_A \cap \bar{A}_B) = 1 - \Pr(A_A) - \Pr(A_B) + \Pr(A_A \cap A_B). \tag{2.44}$$

Each is the general form (2.12) with the indicator of an intersection in place of the indicator of one bound, here with the marginal probabilities given directly rather than computed from a posterior, so there is no state route to report and no gap to state; every probability of this example is conditional on the same evidence, and the bar is left out below to keep the products readable. Suppose $\Pr(A_A) = 0.05$ and $\Pr(A_B) = 0.03$, illustrative values of the size the exceedance log of Section 2.4.4 gives one hall, 0.0354 for a design day and 0.0314 exact. If the two halls' exceedances are independent, the probability that both occur is the product of the marginal probabilities,

$$\Pr(A_A \cap A_B) = \Pr(A_A)\ \Pr(A_B) = 0.05 \times 0.03 = 0.0015,$$

and the probability that both halls stay within their committed loads is, by equation (2.44) or as the product of the two complements,

$$\Pr(\bar{A}_A \cap \bar{A}_B) = 0.95 \times 0.97 = 0.9215, \tag{2.45}$$

or 92.15 %.

2.4.7 A site-wide workload peak as the common cause

Both halls run jobs from the same site workload. Then independence may not hold, and the probability of the intersection in the form (2.12) must be written with the product rule, which holds without any independence assumption:

$$\Pr(A_A \cap A_B) = \Pr(A_A)\ \Pr(A_B \mid A_A). \tag{2.46}$$

The conditional probability $\Pr(A_B \mid A_A)$ may be much larger than the marginal $\Pr(A_B)$: an exceedance in hall A raises the probability that a site-wide workload peak is under way, and

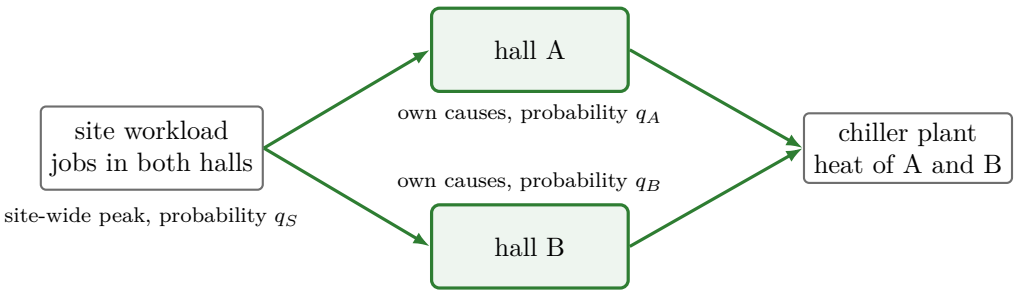


Figure 2.4: Two halls under one site workload. The heat of each hall reaches the shared chiller plant; both halls within their committed loads needs neither load to exceed. A site-wide workload peak (probability q_S in the day), a batch run across both halls, lifts both loads above their committed loads at once; the halls' own causes (probabilities q_A, q_B) are independent outside such a peak. With $q_S = 0$ the two exceedances are independent.

such a peak lifts the load of hall B as well. A shared workload is a common cause that can make the two exceedances strongly dependent. This is the coincidence of the two loads, which the design sheet carries as one coincidence (diversity) factor: the probability that both halls exceed together is not the product of their separate probabilities when a common workload drives both.

An illustrative model shows by how much. Let a site-wide workload peak occur in the day with probability q_S, which lifts both halls above their committed loads, and let each hall otherwise exceed from its own causes, independently, with probabilities q_A and q_B. The law of total probability over the two states of the site workload, the integral of the form over a state with two values, gives the marginal and the joint probabilities, exactly, with no second route beside them:

$$\Pr(A_A) = q_S + (1 - q_S)\, q_A, \qquad \Pr(A_B) = q_S + (1 - q_S)\, q_B,$$
$$\Pr(A_A \cap A_B) = q_S \cdot 1 + (1 - q_S)\, q_A q_B, \qquad \Pr(\bar{A}_A \cap \bar{A}_B) = (1 - q_S)(1 - q_A)(1 - q_B). \quad (2.47)$$

Holding the marginals at 0.05 and 0.03, the own-cause probabilities are $q_A = (0.05 - q_S)/(1 - q_S)$ and $q_B = (0.03 - q_S)/(1 - q_S)$. For $q_S = 0.01$, $q_A = 0.04/0.99 = 0.040404$ and $q_B = 0.02/0.99 = 0.020202$, so

$$\Pr(A_A \cap A_B) = 0.01 + 0.99 \times 0.040404 \times 0.020202 = 0.01 + 0.000808 = 0.010808,$$

and equation (2.46) gives $\Pr(A_B \mid A_A) = 0.010808/0.05 = 0.216$, seven times the marginal 0.03. The probability that both halls stay within their committed loads is $0.99 \times 0.959596 \times 0.979798 = 0.930808$, the same as $1 - 0.08 + 0.010808$ by equation (2.44). Table 2.8 repeats the calculation for five probabilities of a site-wide peak. The first column is q_S; the next two the own-cause probabilities that keep the marginals fixed; then the probability that both halls exceed, the conditional probability of the second exceedance given the first, the ratio of the joint probability to the independent value 0.0015, and the probability that both halls stay within their committed loads. With the marginals unchanged, a site-wide peak probability of 0.01 multiplies the probability of both halls exceeding together by 7.2, and one of 0.02 by 13.5: the site workload, not the halls' own conditions, then governs it. The last two rows, a site-wide peak in one day of a hundred, 3.65 a year, and in one of fifty, are high for a scheduled workload and are there to show the trend.

The shared workload moves the two corners in opposite directions. Both halls within their committed loads becomes slightly more probable, 0.9308 against 0.9215, because when the exceedances come they tend to come together; both halls above their committed loads in the same peak becomes 7.2 times as probable. A product of marginals gets the first nearly right

Table 2.8: Two halls under one site workload through a day's peak, with the marginals held at $\Pr(A_A) = 0.05$ and $\Pr(A_B) = 0.03$ (illustrative model of equation (2.47)).

q_S [–]	q_A [–]	q_B [–]	$\Pr(A_A \cap A_B)$ [–]	$\Pr(A_B \mid A_A)$ [–]	ratio to 0.0015 [–]	$\Pr(\bar{A}_A \cap \bar{A}_B)$ [–]
0	0.05000	0.03000	0.001500	0.0300	1.00	0.92150
0.001	0.04905	0.02903	0.002422	0.0484	1.61	0.92242
0.005	0.04523	0.02513	0.006131	0.1226	4.09	0.92613
0.01	0.04040	0.02020	0.010808	0.2162	7.21	0.93081
0.02	0.03061	0.01020	0.020306	0.4061	13.54	0.94031

and the second badly wrong, and the second is the one that sizes the shared chiller plant. The wording "both halls" names an intersection of events, and an intersection directs the formulation to joint probability and to the modeling of dependence, not to a product of marginals. In a larger plant the event of concern is often a union of many overlapping events, some inlet somewhere above its limit, and Chapter 7 derives its probability as a joint tail over the inlets at the design-day scenario; the coincident peak of correlated rack loads returns in Chapter 8.

2.5 Model comparison, diagnosis and decisions

The questions so far were asked of the state of one model. A question about which model holds or which fault is present is put to several models at once, and its answer is a ratio or a ranking of evidences; a question about which action to take is put to several actions, and its answer is a comparison of expected consequences.

2.5.1 Posterior probabilities of competing models

Consider a hall whose temperature readings are inconsistent with its expected behavior, and two hypotheses about it,

$$M_0 : \text{the air path is as drawn,} \qquad M_1 : \text{the air path is partially obstructed,}$$

the second a clogged filter or a blocked return, say. The question, asked before the hall's design load is approved, is which explanation the evidence supports better. This is not a threshold probability on a fixed, known model: the planner is comparing models. It is a hypothesis comparison, of the general form (2.14) with $M_A = M_1$ and $M_B = M_0$. The plant enters through the two drawings: each is the graph of the air path with its own rows, and M_1 carries an entry for the obstruction, a reduction of the conductance of the path, that M_0 pins at zero. Each evidence is the integral (2.26) over its own model's state, read under the Laplace approximation from that model's own fit, its MAP state $\hat{\mathbf{z}}_j$ and covariance $\boldsymbol{\Sigma}_j$. The extra entry of M_1 is not rewarded merely for fitting the readings better: the evidence multiplies the height of the fit by the posterior volume $\det(\boldsymbol{\Sigma}_j)^{1/2}$, and the prior density $p(\hat{\mathbf{z}}_j \mid M_j)$ in that height is low when the prior spreads the entry over a wide range. Treat the model index as an unknown taking the values $M_0, M_1, \ldots$, with prior probabilities $\Pr(M_j)$. Bayes' rule for a discrete unknown, with the law of total probability in the denominator, gives

$$\Pr(M_j \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid M_j)\, \Pr(M_j)}{\sum_k p(\mathbf{y} \mid M_k)\, \Pr(M_k)}, \tag{2.48}$$

where $\Pr(M_j)$ is the prior probability of model M_j, $\Pr(M_j \mid \mathbf{y})$ its posterior probability, and $p(\mathbf{y} \mid M_j)$ its evidence, equation (2.26). Dividing equation (2.48) for M_1 by the same equation

for M_0 cancels the denominator and gives the odds form,

$$\frac{\Pr(M_1 \mid \mathbf{y})}{\Pr(M_0 \mid \mathbf{y})} = B_{10}\,\frac{\Pr(M_1)}{\Pr(M_0)}, \qquad B_{10} = \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_0)}\,: \tag{2.49}$$

posterior odds equal the Bayes factor (2.14) times the prior odds.

2.5.2 A numerical comparison of two explanations

Suppose the evidences are given, as illustrative values, not read from fits, so there is no Laplace value and exact value to set side by side: $p(\mathbf{y} \mid M_0) = 0.01$ and $p(\mathbf{y} \mid M_1) = 0.04$, and the prior probabilities are $\Pr(M_0) = 0.8$ and $\Pr(M_1) = 0.2$. Equation (2.48) gives

$$\Pr(M_1 \mid \mathbf{y}) = \frac{0.04 \times 0.2}{0.04 \times 0.2 + 0.01 \times 0.8} = \frac{0.008}{0.016} = 0.50.$$

The Bayes factor is

$$B_{10} = \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_0)} = \frac{0.04}{0.01} = 4, \qquad \ln B_{10} = \ln 4 = 1.386 \text{ nats},$$

and the odds form checks the result: posterior odds $4 \times (0.2/0.8) = 1$, a posterior probability of $1/(1+1) = 0.50$. The evidence is four times more probable under the obstruction model, yet its posterior probability is only 50 %, because the prior probability assigned to obstruction was low. The Bayes factor expresses the relative support that the evidence provides; it is not by itself a posterior model probability, and turning it into one requires the prior probabilities of the models.

Figure 2.5 makes the distinction visible. In panel (a) the two pairs of bars are the prior and the posterior probabilities of the two models: the evidence moves probability from M_0 to M_1 until they are equal. In panel (b) the horizontal axis is the prior probability of the obstruction model and the vertical axis its posterior probability; the green curve is $\Pr(M_1 \mid \mathbf{y}) = 4\,\Pr(M_1)/(3\,\Pr(M_1) + 1)$, which equation (2.49) gives for $B_{10} = 4$, and the dashed diagonal is the case $B_{10} = 1$, evidence that does not discriminate. The same Bayes factor of 4 turns a prior of 0.05 into 0.174, a prior of 0.2 into 0.50 (the marked point) and a prior of 0.5 into 0.80: the factor is fixed by the evidence, the conclusion is not.

2.5.3 Diagnosis over a family of candidate faults

Fault detection and isolation usually involves many competing explanations, not two. For an existing hall that runs warmer than its design expects, before more load is planned onto it, they might be a fouled coil, a fan running slow, a biased hall sensor, a load meter reading low or a floor tile left out of place. The mathematical target is the general form (2.15), normalized as in equation (2.48): the posterior probability of each candidate,

$$\Pr(F_j \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid F_j)\,\Pr(F_j)}{\sum_k p(\mathbf{y} \mid F_k)\,\Pr(F_k)},$$

where F_j is a candidate fault condition, which is the drawing M_j of equation (2.15): the plant graph with that fault written into its rows, so that $p(\mathbf{y} \mid F_j) = p(\mathbf{y} \mid M_j)$. The output is the ranking of the candidates by it, not a single verdict. A fault condition can be represented in two ways: as a separate model, whose evidence enters equation (2.48), or as a latent quantity within one common model, such as a bias entry of the state that is zero when the sensor is sound. The second representation means that diagnosis does not always require fitting a separate model for every possible fault. Diagnosis earns its own category for two reasons: the candidates are generated systematically from the plant graph, one for each component or reading that could

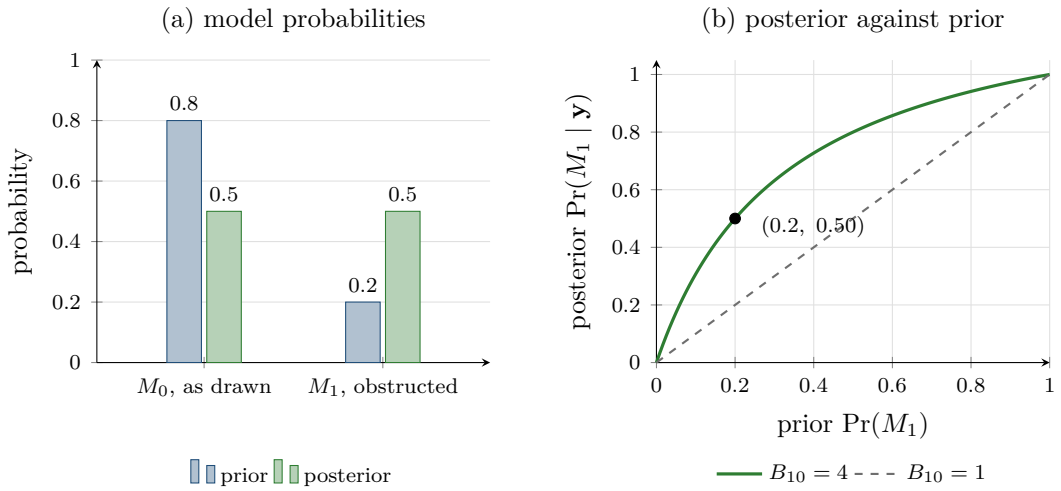


Figure 2.5: Model comparison for the air path of the hall. (a) Prior and posterior probabilities of the as-drawn (M_0) and obstructed (M_1) models, with evidences 0.01 and 0.04. (b) Posterior probability of M_1 as a function of its prior, $4x/(3x+1)$ for a Bayes factor of 4 (equation (2.49)); the dashed diagonal is a Bayes factor of 1; the point is the example.

be at fault, rather than written by hand; and the answer is a ranking. The critical step in either representation is to define the candidate fault events or hypotheses clearly enough that their probabilities can be calculated and compared.

For the hall of Figure 2.1 the family is built on the graph with its reading row kept as a row of its own, $(T_1 - y_T)/\sigma_T$, rather than folded, as in the fit of Section 2.4.1, into the summary row on g; a fault must have a row to act on. One candidate is generated for each element that could be at fault, each the extended graph with one entry added: M_{bias}, a bias entry Δ_T on the reading row, which becomes $(T_1 + \Delta_T - y_T)/\sigma_T$; M_{path}, a shift Δ_g of the log-conductance away from its prior value, an entry added to both balance rows; and M_{load}, an offset Δ_Q of the present load Q_1 from its nominal value, an entry added to the present-load balance row. Each candidate is fitted to the same readings, its evidence is read from its own $\hat{\mathbf{z}}_j$ and $\mathbf{\Sigma}_j$, and the ranking orders them. Which candidates the readings can tell apart is read from the Jacobian before any fit. With the one reading of the hall, each added entry can absorb the misfit of that reading alone, so the three candidates fit it equally well and their ranking rests on their prior rows and prior probabilities. A second reading at another load separates them, because the three added entries then move the two readings in different proportions, which are different columns of $\mathbf{J}$: a bias by equal amounts, a shift of the path in proportion to the loads, and an offset of the present load at the present load only.

2.5.4 Expected loss of renting a temporary chiller for the peak season

Questions about a plant being planned also lead to decisions. Before the peak season the planner can rent a temporary chiller for the hall, a unit outside the one-path model of Chapter 1, or wait, and the probability of an excursion is not enough to decide: the consequence of an excursion and the cost of the action enter as well. The costs of this section and of Section 2.6.2 are stated in thousands of reais (R$ thousand), so that R$ 80 below is R$80 thousand and R$ 100 000 is R$100 million. Let e be the event that the hall air stays above its limit at the season's peak long enough for the IT equipment's protection to act, an excursion that sheds the hall's IT load, with consequence $C_e = \text{R\$ } 100\,000$ in service credits and recovery, and suppose that if the planner waits its probability over the season is $\Pr(e \mid \mathbf{y}, \text{wait}) = 0.0015$, an illustrative value. An excursion in this sense, the hall air held above the limit long enough for the IT protection to

act, is far rarer than the hall crossing the 27.5°C envelope of Question B, whose probability at 60 kW, under the row of equation (2.29), is 0.280, with 0.283 exact. The excursion is a future event that stays random when the state is known, so the loss of waiting given the state is itself an expectation, the derived quantity

$$L(\text{wait}, \mathbf{z}) = C_e \Pr(e \mid \mathbf{z}, \text{wait}) + 0 \cdot \Pr(\bar{e} \mid \mathbf{z}, \text{wait}) = C_e \Pr(e \mid \mathbf{z}, \text{wait}),$$

with no loss when no excursion occurs, and the expected loss is its location (2.10), an estimation question about a consequence, answered by the expected consequence (2.23) of waiting. Averaging over the posterior with equation (2.8) turns the conditional probability into its predictive one, and the expected loss is the probability-weighted average of the two outcomes:

$$\mathcal{R}(\text{wait}) = \mathbb{E}[L(\text{wait}, \mathbf{z}) \mid \mathbf{y}] = \Pr(e \mid \mathbf{y}, \text{wait})\, C_e + \Pr(\bar{e} \mid \mathbf{y}, \text{wait}) \cdot 0, \tag{2.50}$$

the second line of equation (2.23) with one event, consequence C_e, and no cost of action: probability times consequence. Substituting,

$$\mathcal{R}(\text{wait}) = 0.0015 \times \text{R\$}\,100\,000 + 0.9985 \times 0 = \text{R\$}\,150$$

per season, under these assumptions. It does not mean that the planner's site will lose exactly R$ 150; the actual loss is either zero or R$ 100 000. It means that R$ 150 is the probability-weighted average of the possible losses, the amount that would be lost per season on average over many seasons like this one.

Now suppose renting the temporary chiller costs $c_{\text{rent}} = \text{R\$}\,80$ for the season, in rent, connection and energy (a cost, the c_a of equation (2.23)), and lowers the probability of an excursion from 0.0015 to 0.0005. Its loss given the state is $L(\text{rent}, \mathbf{z}) = c_{\text{rent}} + C_e \Pr(e \mid \mathbf{z}, \text{rent})$, and its expected loss is the certain cost of the action plus the expected loss that remains,

$$\mathcal{R}(\text{rent}) = \text{R\$}\,80 + 0.0005 \times \text{R\$}\,100\,000 = \text{R\$}\,130,$$

against $\mathcal{R}(\text{wait}) = \text{R\$}\,150$. The expected net benefit of renting the chiller is $\text{R\$}\,150 - \text{R\$}\,130 = \text{R\$}\,20$ per season, and under these assumptions renting it is the preferred decision. The comparison rests on three numbers, two probabilities and one consequence, and its conclusion changes if any of them moves: the break-even cost of the action is $(0.0015 - 0.0005) \times \text{R\$}\,100\,000 = \text{R\$}\,100$, and, read the other way, acting and waiting cost the same when the probability of an excursion if the planner waits is $0.0005 + 80/100\,000 = 0.0013$, the break-even probability, the same in any unit of cost because it is a ratio. A season whose excursion probability the posterior puts below 0.0013 is a season to wait. The probabilities here stand for the predictives a posterior would supply; they are given, so the expected losses carry no gap of their own, and a gap in the excursion probability that straddled 0.0013 would be stated with both values and both actions.

2.5.5 The general decision formulation

Let a be a design action from a set of candidates, and $L(a, \mathbf{z})$ the loss of taking it when the state is $\mathbf{z}$. The expected loss of the action is its expected consequence (2.23), $\mathcal{R}(a) = \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$, the location (2.10) of the derived quantity $L(a, \mathbf{z})$, one for each candidate, and the preferred action is the one that minimizes it:

$$a^* = \arg\min_a \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]. \tag{2.51}$$

The temporary-chiller example of Section 2.5.4 is this rule with one event, the excursion, and two actions. The change is an important one. The questions of the earlier sections asked what will happen; equation (2.51) asks what should be done, given what might happen. The transition from inference to decisions is what the design review is for.

2.6 Measurements, the classification of questions and the derivation workflow

A design question asks whether more information is worth obtaining, and it is answered before the information exists.

2.6.1 Uncertainty reduction by a measurement

Suppose the planner is uncertain about a quantity because the available record is short, and an additional sensor or a capacity test could reduce that uncertainty. The question "is the information worth its cost?" is a design question, of the general form (2.16). Its first part is how much the measurement would narrow the posterior: the form with u the posterior covariance, which for linear Gaussian rows is the preposterior covariance (2.17), with the information gain (2.18) as its summary. The plant enters through two objects: the Jacobian $\mathbf{J}_a$ of the candidate rows, which says which entries of the state the measurement acts on and how strongly, and the current covariance $\mathbf{\Sigma}$, which says along which directions the posterior can still shrink.

For the hall, consider n readings of the hall temperature at the present load $Q_1 = 30\,\mathrm{kW}$, each with an error of standard deviation $\sigma_T = 0.5\,\mathrm{K}$, that of the hall sensor of Chapter 1 (independent readings, from sensors at different places of the hall). Each reading is a row on the first entry of the state of equation (2.30),

$$r_{a,j}(\mathbf{z}) = \frac{T_1 - \tilde{y}_j}{\sigma_T}, \qquad \mathbf{J}_{a,j} = \frac{\partial r_{a,j}}{\partial \mathbf{z}} = \begin{bmatrix} 1/\sigma_T & 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \end{bmatrix} \mathrm{K}^{-1},$$

so that n readings stack to an $n \times 3$ Jacobian $\mathbf{J}_a$ with $\mathbf{\Lambda}_a = \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a = \mathrm{diag}(4n, 0, 0)$ in K^{-2}. The current $\mathbf{\Sigma}$ is the ridge of equation (2.33), $\sigma_g^2 \mathbf{v}\mathbf{v}^{\mathsf{T}}$, so in the update form (2.27) every product runs through $\mathbf{v}$: $\mathbf{\Sigma}\mathbf{J}_{a,j}^{\mathsf{T}} = \sigma_g^2\,\mathbf{v}\,(2v_1)$ and $\mathbf{J}_{a,j}\mathbf{\Sigma}\mathbf{J}_{a,k}^{\mathsf{T}} = 4\sigma_g^2 v_1^2 = 4\Sigma_{11} = 0.21302$ for every pair of readings. The update therefore removes a multiple of $\mathbf{v}\mathbf{v}^{\mathsf{T}}$ itself, and

$$\mathbf{\Sigma}_a = \frac{\mathbf{\Sigma}}{1 + 0.21302\,n}, \qquad \mathbf{\Sigma}_a = \frac{\mathbf{\Sigma}}{1.21302} = \begin{bmatrix} 0.04390 & 0.08780 & -0.00951 \\ 0.08780 & 0.17561 & -0.01902 \\ -0.00951 & -0.01902 & 0.00206 \end{bmatrix} \quad (n = 1). \tag{2.52}$$

The reading sees only T_1, yet every entry of $\mathbf{\Sigma}$ shrinks by the same factor, because the posterior has only the one direction $\mathbf{v}$ to shrink along: one reading removes $1 - 1.21302^{-1/2} = 9.2\,\%$ of the spread of T_1, T_2 and g alike, so that $\sqrt{\Sigma_{a,33}}$ falls from 0.05 to 0.0454 and $\sqrt{\Sigma_{a,22}}$ from 0.462 to 0.419 K. Table 2.9 gives the result for 0 to 64 readings. The columns are the number of readings, the precision $1/\Sigma_{a,33}$ of g, its posterior standard deviation $\sqrt{\Sigma_{a,33}}$, the standard deviation $\sqrt{\Sigma_{a,22}}$ of T_2, the fraction of that spread removed, and the expected information gain. The precision of g is $400\,(1 + 0.21302\,n) = 400 + 85.21\,n$ by equation (2.52). The expected information gain is equation (2.18); in the limit both determinants are zero, and the determinant identity $\det\mathbf{\Sigma}/\det\mathbf{\Sigma}_a = \det(\mathbf{I} + \mathbf{J}_a\mathbf{\Sigma}\mathbf{J}_a^{\mathsf{T}})$, with every entry of the $n \times n$ matrix $\mathbf{J}_a\mathbf{\Sigma}\mathbf{J}_a^{\mathsf{T}}$ equal to 0.21302, gives $\mathrm{EIG} = \frac{1}{2}\ln(1 + 0.21302\,n)$. Equation (2.17) evaluated directly at $\sigma_p = 10^{-4}\,\mathrm{K}$, where $\mathbf{\Sigma}^{-1}$ exists, gives the same entries, the precision of 64 readings to within 0.02. The reduction is steep at first and then slows, because the precision grows linearly in n and the spread only as its inverse square root: halving the spread requires a precision of 1600, about fourteen readings, and halving it again a precision of 6400, about seventy.

2.6.2 The value of a capacity test before deciding

Reducing uncertainty is not necessarily the same as making a better decision. Suppose there are two possible actions, rent the temporary chiller for the season and wait. The best action

Table 2.9: Preposterior spread of the hall after n readings at $Q_1 = 30\,\mathrm{kW}$ with $\sigma_T = 0.5\,\mathrm{K}$ (illustrative), from the posterior $\boldsymbol{\Sigma}$ of equation (2.33) and equation (2.52).

n [–]	precision $1/\Sigma_{a,33}$ [–]	$\sqrt{\Sigma_{a,33}}$ [–]	$\sqrt{\Sigma_{a,22}}$ [K]	spread removed [–]	EIG [nat]
0	400.0	0.0500	0.462	0	0
1	485.2	0.0454	0.419	9.2 %	0.097
4	740.8	0.0367	0.339	26.5 %	0.308
16	1763.3	0.0238	0.220	52.4 %	0.742
64	5853.3	0.0131	0.121	73.9 %	1.342

now minimizes the expected loss, and the expected loss of acting now is

$$\bar{L}_{\text{now}} = \min_a \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}].$$

(The book writes $\bar{L}$ for these minimum expected losses because C names a consequence.) If a candidate measurement is taken first, written m here because a ranges over the actions, with outcome $\tilde{y}_m$, the best action can be chosen again after its value is known, with expected loss $\min_a \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}, \tilde{y}_m]$. Before the measurement is taken its result is unknown, so that expected loss is averaged over the possible results, weighted by their predictive distribution given the present evidence:

$$\bar{L}_{\text{after}} = \mathbb{E}_{\tilde{y}_m|\mathbf{y}}\Big[\min_a \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}, \tilde{y}_m]\Big].$$

This is the general form (2.16) for the candidate m, with u the minimum expected loss under the updated posterior and the candidate ranked smallest first, and here, unlike for the spread, u depends on the outcome itself and not only on its covariance. The expected value of sample information is the difference,

$$\text{EVSI} = \bar{L}_{\text{now}} - \bar{L}_{\text{after}}, \tag{2.53}$$

the expected improvement in the decision, measured as a reduction in expected loss. It is never negative. For each fixed action, averaging the post-measurement expected loss over the results returns the present one, $\mathbb{E}_{\tilde{y}_m|\mathbf{y}}[\mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}, \tilde{y}_m]] = \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$, by the law of total expectation; and the minimum over actions inside the average is at most the value of any one action, so $\bar{L}_{\text{after}} \le \min_a \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}] = \bar{L}_{\text{now}}$. Equality holds when no possible result would change the action: a measurement that cannot change the decision has no value for it, however much it narrows the posterior. If the measurement costs more than its EVSI, commissioning it is not justified on this criterion.

An illustrative case shows the calculation, with the state here the capacity of the hall unit's coil, two values, so that every integral is a finite sum, exact, and no second route stands beside it. Suppose the excursion probability 0.0015 of Section 2.5.4 is an average over two conditions of the coil: below its rating, with probability 0.25 and excursion probability 0.0045 if the planner waits, and at its rating, with probability 0.75 and excursion probability 0.0005; indeed $0.25 \times 0.0045 + 0.75 \times 0.0005 = 0.0015$. Renting the temporary chiller costs R$ 80 (R$ thousand, as in Section 2.5.4) and leaves either condition at 0.0005. A capacity test of the coil, its heat removal measured at a stated load and set against its rating, would reveal the condition exactly. Table 2.10 lists, for each condition and for the present state of knowledge, the expected loss of each action and the better one. Its columns are the condition, its probability, the excursion probability if the planner waits, the expected loss of waiting, that of renting the chiller, and the better action. Without the test the best action is to rent the chiller, at R$ 130. With it, a coil below its rating leads to renting at R$ 130 and one at its rating to waiting at R$ 50, so

$$\bar{L}_{\text{after}} = 0.25 \times \text{R\$}\,130 + 0.75 \times \text{R\$}\,50 = \text{R\$}\,70, \qquad \text{EVSI} = \text{R\$}\,130 - \text{R\$}\,70 = \text{R\$}\,60.$$

A capacity test that reveals the condition is worth up to R$ 60 per season; because it reveals the condition exactly, this is the expected value of perfect information, an upper bound on what any imperfect test could be worth.

Table 2.10: Expected loss per season of the two actions for a coil of unknown capacity (illustrative, in R$ thousand): temporary chiller R$ 80 for the season, excursion consequence R$ 100 000.

Condition	probability [–]	Pr(e), wait [–]	wait [R$ thousand]	rent chiller [R$ thousand]	better action
below its rating	0.25	0.0045	450	130	rent chiller
at its rating	0.75	0.0005	50	130	wait
unknown (present)	1	0.0015	150	130	rent chiller

Three quantities are now on the table, and they are related but not interchangeable. The reduction of the posterior covariance, equation (2.17), measures how much uncertainty disappears. The expected information gain, equation (2.18), measures how much is learned. The value of information, equation (2.53), measures how much better the resulting decision is expected to be. A measurement can score well on the first two and be worth nothing on the third, which is why the choice of the next measurement and the choice of an action both need the preposterior but do not ask the same question of it. The value of information is where two computations meet: the preposterior supplies the outcomes and their weights, and the expected consequence, minimized over the actions, is what it averages. Chapter 18 ranks the tests and surveys on the existing plant by the width each takes off the quantity that decides the build.

2.6.3 Identifying the mathematical form of a question

The examples supply a systematic procedure: before computing anything, identify the mathematical objective of the question. Table 2.11 is the reference for that step. Each row is a form of question in the planner's words, the general form it takes, by equation number, its mathematical form, what it derives to, and its category. These forms recur in the atlas of Appendix A, which collects the questions of each chapter that has a questions table, with the section that answers each.

Many variations of these forms exist. Redundancy levels, load paths over a planning horizon, maintenance windows and coincident loads in several halls each add mathematical structure of their own: joint tails over many entries, integrals of an excursion probability over a horizon, consequences that accrue over a season. The rows of the table, over the general forms of the categories, of a trend and of an expected consequence, already cover a substantial range of inference problems, and a new question is usually classified by recognizing which of them it is, before any computation.

2.6.4 The general derivation workflow

The classification is one step of the derivation workflow. Figure 2.6 draws its eight steps: the top row runs from left to right, the bottom row from right to left, and the arrow beneath the bottom row is the return from step 7 to step 3 when a new measurement is added. As in the preface, at more length:

The plant graph (step 0). The physical system as it is, before any question is asked: its nodes, paths and sources; which quantities are pinned and which are free; the physical relations that must hold among them. The free quantities form the state vector $\mathbf{z}$.

Table 2.11: Classification of questions by their mathematical form, with the general form of each category from Section 2.3.3 and the forms of a trend and of an expected consequence from Sections 2.3.4 and 2.3.5.

Planner's question	General form	Mathematical form	Derives to	Category
What is unknown, and how well is it known?	(2.9), (2.10)	$p(z_i \mid \mathbf{y})$, $\mathbb{E}[z_i \mid \mathbf{y}]$	a distribution, a mean, a credible interval: $\hat{z}_i$ and Σ_{ii}	estimation
Does a requirement hold, with the stated assurance?	(2.12)	$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y}) \geq \kappa$, f^* the requirement	the posterior integral of the indicator of the requirement; a normal tail under linearization	prediction under a threshold
What happens on the design day?	(2.13)	$\int \Pr(A_{\text{future}} \mid \mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}$	the average over the state, equation (2.41), with the closed form (2.42) beside it where the posterior is a gamma	prediction under a threshold
Is it rising, and when does it cross its limit?	(2.21)	$\Pr(\gamma > 0 \mid \mathbf{y})$; $\Pr(t^* \leq s \mid \mathbf{y}) = \Pr(f_s(\mathbf{z}) \geq f^* \mid \mathbf{y})$	the posterior read along time, $\mathbf{g}(t)^{\mathsf{T}}\hat{\mathbf{z}}$ and $\mathbf{g}(t)^{\mathsf{T}}\boldsymbol{\Sigma}\,\mathbf{g}(t)$; a crossing band by inversion	prediction under a threshold for the sign and a crossing by a date; estimation for a forecast and the distribution of a crossing time
Which explanation is correct?	(2.14), (2.15)	$\Pr(M_j \mid \mathbf{y}) \propto p(\mathbf{y} \mid M_j)\, \Pr(M_j)$	the evidences and the prior model probabilities; the Bayes factor	hypothesis comparison; diagnosis over a family of faults
What is the expected consequence?	(2.23)	$\mathcal{R}(a) = \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}] = \int L(a, \mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}$	probability times consequence; $L(a, \hat{\mathbf{z}}) + \frac{1}{2}\operatorname{tr}(\mathbf{H}_a\boldsymbol{\Sigma})$	estimation of a derived quantity
What should be done or measured?	(2.23), (2.16)	$\arg\min_a \mathcal{R}(a)$; EVSI	the expected consequence of each action; the preposterior $\boldsymbol{\Sigma}_a$	design

The question (step 1). A probability or an expectation of a function of the state, with the unknown on the left of the conditioning bar and everything treated as known on the right.

The graph extended for the question (step 2). Scenario nodes for a load, an ambient, a unit out of service or a design alternative not yet built, candidate faults when something may have changed, and the evidence attached as rows: readings, exceedance counts, capacity tests, handbook values as priors, each with the modeling assumption it rests on, what that assumption buys and how it can fail.

The posterior of the state (step 3). The rows combined into one objective; its minimum, the MAP state; its curvature, the information; and the covariance, derived and then computed for the plant, with the exact posterior beside them where the model admits one. It is built once and serves every question.

The answers (step 4). Each question read from that posterior, from the planner's question through its general and specialized forms down to its number: an entry of the state or the gradient of a derived quantity, the quantity at a time asked included; when drawings compete, a ratio of evidences; when actions compete, an expected consequence.

Evaluation and checks (step 5). The gaps between the reported values and the exact values or draws collected in a table, the approximations tested, the answers compared with the truth where the record is constructed.

The report (step 6). The probability, the Bayes factor or the expected consequence with its band, or a crossing time with its band by inversion, with the exact value beside it where one exists.

The next measurement (step 7). If the band is too wide to decide, the tests and surveys on the existing plant ranked by the width each would take off the quantity that decides, and a return to step 3 with the new row.

The examples of this chapter sit in it as follows. Stating precisely what the planner needs to know or decide, and defining the events, uncertain quantities, future outcomes, hypotheses or actions that express it, is step 1; adding the design-load scenario node of the hall example, and attaching the readings and priors as rows with the assumptions behind them, is step 2. Step 3 builds the posterior the questions are read from: a fit for the most probable state and its Laplace covariance, a numerical integration, the evidence of each competing model, or another operation appropriate to the problem; for the hall it is the Gauss–Newton fit of three rows, and for the exceedance log the Newton fit of the log-rate, whose objective is a gamma kernel and so gives the update in closed form as well. Step 4 writes each question in its general form and reduces it, through the specialized form, to posterior marginals, integrals, gradients, evidences or decision objectives, and evaluates them. Step 5 checks assumptions, identifiability, approximations, numerical convergence and the sensitivity of the answer to modeling choices, as the tolerance, exact-value and sampling checks of the hall example did. Step 6 communicates the answer, its uncertainty, its limitations and what it means for the design, and step 7 is the design question of Section 2.6.1.

The workflow is broader than any particular software implementation. The derivation fixes what has to be calculated before a computational tool is chosen, and the tool then calculates it.

Principle 2.3. *The derivation tells the computational tool what must be calculated. A computational tool should not dictate which planning question is asked.*

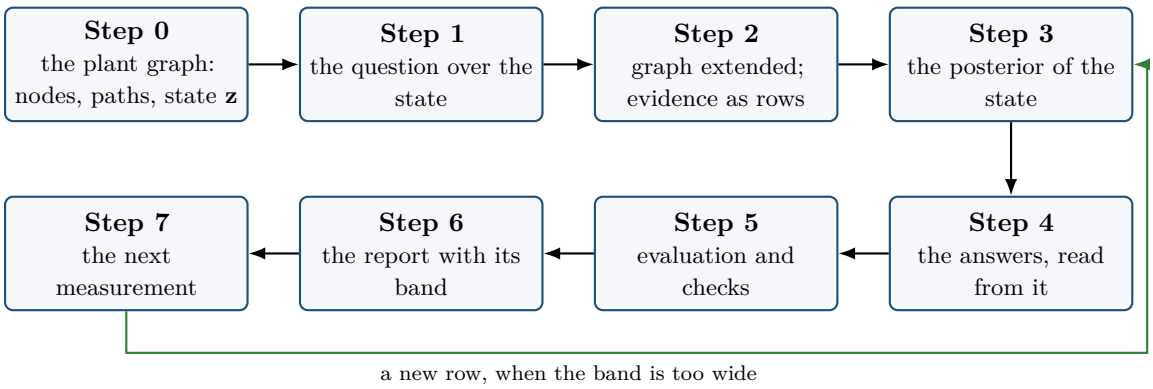


Figure 2.6: The derivation workflow of the book. Step 0 is built once per plant; steps 1 to 7 are one pass; step 3 is shared by every question put to the same plant and evidence; the green return arrow adds the measurement chosen at step 7 as a new row and rebuilds the posterior from step 3.

2.6.5 Common mistakes in formulating capacity questions

Five mistakes of formulation recur in design reviews, and each is a confusion between two questions that look alike.

Common mistake

Confusing uncertainty about an exceedance rate with the probability of a design day within the committed load. For a hall with an unknown exceedance rate λ, the probability of no exceedance over a horizon τ at a given rate is $\mathrm{e}^{-\lambda\tau}$. The question $\mathbb{E}[\mathrm{e}^{-\lambda\tau} \mid \mathbf{y}]$ is the posterior predictive probability of a horizon without exceedance: how likely the next design day is to pass within the committed load. The question $\Pr(\mathrm{e}^{-\lambda\tau} > 0.99 \mid \mathbf{y})$ is the posterior probability that the hall's underlying rate meets a target written as a 99 % chance of a horizon without exceedance: how much support there is for the hall meeting its planning target. They are different numbers and answer different questions.

Common mistake

Confusing prediction with diagnosis. When the hall runs warm, the prediction $\Pr(T > T_{\max} \mid \mathbf{y})$ concerns a temperature event, and the diagnosis $\Pr(M_{\text{fouled}} \mid \mathbf{y})$ concerns an explanation, a fouled coil against a sound one. They can use the same readings but require different calculations: a tail of a posterior in one model against a comparison of evidences across models.

Common mistake

Ignoring the time horizon. $\Pr(T > 12\,\mathrm{h} \mid \mathbf{y})$ and $\Pr(T > 8760\,\mathrm{h} \mid \mathbf{y})$ are not interchangeable statements about the exceedances; for the exceedance log of Section 2.4.4 the first, the peak hours of a design day, is 0.9646 (0.9686 exact) and the second, a year of logged operation, is 0.00004 (exact, $(1500/10\,260)^4 = 0.0005$). A probability of a horizon without exceedance is incomplete without its horizon and its load conditions.

Common mistake

Forgetting dependence. In general $\Pr(A \cap B) \neq \Pr(A)\,\Pr(B)$; the product is justified only under independence. A shared workload, a shared chiller plant, the weather and the site's maintenance work create dependence between halls and units, and Table 2.8 shows that a small common cause can multiply the probability of two halls exceeding their committed loads in the same peak several times over.

Common mistake

Treating the model assumptions as certain. A calculated probability may be numerically precise while depending strongly on uncertain modeling assumptions. A temperature-exceedance probability computed under the steady-state model of equation (2.2) may be unsuitable for the minutes after the cooling unit trips, when the hall air rises through a transient that the steady-state model does not describe and the supply air is no longer held at T_c. Every probability is conditional on its model, $\Pr(A \mid \mathbf{y}, M)$, and must be interpreted in that context.

The questions of this chapter were all written with conditional probabilities and the posterior was taken as given; the next chapter, Chapter 3, derives conditional probability and Bayes' rule, separates aleatory variability from epistemic uncertainty, and shows how readings change the uncertainty about the state of the hall through marginalization, dependence and conditional independence on the plant graph.

2.7 What to retain

Definition 2.1 (Planning question)**.** A planning question about a plant is a probability or an expectation of a function of the state of its plant graph, written with the unknown on the left of the conditioning bar and the evidence and the model on the right, $\Pr(A \mid \mathbf{y}, M) = \int \mathbf{1}_A(\mathbf{z})\, p(\mathbf{z} \mid \mathbf{y}, M)\, \mathrm{d}\mathbf{z}$ for a present event and $\int \Pr(A_{\text{future}} \mid \mathbf{z}, M)\, p(\mathbf{z} \mid \mathbf{y}, M)\, \mathrm{d}\mathbf{z}$ for a future one. Its form decides its category, estimation, prediction under a threshold, hypothesis comparison, diagnosis or design, each with its general forms, equations (2.9) to (2.16); the question decides which of five computations answers it, the posterior of the state, the trend, the evidence, the preposterior or the expected consequence (Table 2.3), two of which carry general forms of their own, the trend form (2.21) and the expected-consequence form (2.23); and the plant enters only in the specialized form, through the entry or function of the state, the event or family of models, and the objects of the fit that carry the answer.

Table 2.12 lists the chapter's terms with their expressions, units and the subsection that defines each, grouped by their role.

Table 2.12: The terms of capacity questions and where each is defined.

Term	Expression	Units	Meaning	Defined in
Event	A, e.g. $T > 28\,^{\circ}\mathrm{C}$	–	a statement true or false	2.2.1
Conditional probability	$\Pr(A \mid \mathbf{y}, M)$	–	probability given evidence and model	2.2.3
Event probability from a density	$\int_5^7 p(G \mid \mathbf{y})\, \mathrm{d}G$	–	probability of the range $5 < G < 7$	2.2.4
Probability over the state	$\int \mathbf{1}_A(\mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}$	–	posterior expectation of the indicator	2.3.1

Term	Expression	Units	Meaning	Defined in
Posterior predictive	$\int \Pr(A_{\text{future}} \mid \mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}$	–	future event averaged over the state	2.3.2
General forms				
Estimation	$\Pr(f(\mathbf{z}) \le f^* \mid \mathbf{y})$; $\mathbb{E}[f \mid \mathbf{y}]$, $\mathrm{Var}(f \mid \mathbf{y})$	–	posterior of a derived quantity	2.3.3
Credible interval	$\Pr(f \le f_L \mid \mathbf{y}) = 0.025$, $\Pr(f \le f_U \mid \mathbf{y}) = 0.975$	–	central 95 % band; quantiles map through monotone f	2.3.3
Threshold and assurance	$\Pr(f(\mathbf{z}) \ge f^* \mid \mathbf{y}) \ge \kappa$	–	probability that the requirement $f \ge f^*$ holds, against κ	2.3.3
Predictive	$\int p(y_{\text{new}} \mid \mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}$	–	future observation averaged over the state	2.3.3
Comparison	$B_{AB} = p(\mathbf{y} \mid M_A)/p(\mathbf{y} \mid M_B)$	–	evidence ratio of two drawings	2.3.3
Diagnosis	$\Pr(M_j \mid \mathbf{y}) \propto p(\mathbf{y} \mid M_j)\, \Pr(M_j)$, ranked	–	comparison over a family of faults	2.3.3
Design	$U(a) = \mathbb{E}_{\tilde{y}_a \mid \mathbf{y}}[u(p(\mathbf{z} \mid \mathbf{y}, \tilde{y}_a))]$	varies	preposterior expectation	2.3.3
Preposterior covariance	$\boldsymbol{\Sigma}_a = (\boldsymbol{\Sigma}^{-1} + \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a)^{-1}$	varies	spread after a future measurement	2.3.3
Expected information gain	$\frac{1}{2}\ln(\det \boldsymbol{\Sigma} / \det \boldsymbol{\Sigma}_a)$	nat	what a measurement teaches	2.3.3
Trend	$\Pr(\gamma > 0 \mid \mathbf{y})$; $\Pr(t^* \le s \mid \mathbf{y}) = \Pr(f_s(\mathbf{z}) \ge f^* \mid \mathbf{y})$; $p(f_t(\mathbf{z}) \mid \mathbf{y})$; read with $\mathbf{g}(t)^{\mathsf{T}}\hat{\mathbf{z}}$, $\mathbf{g}(t)^{\mathsf{T}}\boldsymbol{\Sigma}\,\mathbf{g}(t)$	–	sign, crossing and forecast: the posterior read along time	2.3.4
Rows of a record in windows	$\mathbf{J} = \mathbf{D}\mathbf{X}$, $[\boldsymbol{\Lambda}]_{\gamma\gamma} = \sum_k d_k^2 (t_k - \bar{t})^2$	varies	a window's distance enters squared, windows add linearly	2.3.4
Expected consequence	$\mathcal{R}(a) = \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$, $a^* = \arg\min_a \mathcal{R}(a)$; spread $\sqrt{\mathbf{g}_a^{\mathsf{T}}\boldsymbol{\Sigma}\,\mathbf{g}_a}$	varies	probability times consequence, minimized over actions	2.3.5
Curvature term	$L(a, \hat{\mathbf{z}}) + \frac{1}{2}\,\mathrm{tr}(\mathbf{H}_a\boldsymbol{\Sigma})$	varies	what the plug-in loss misses, in the state's coordinates	2.3.5
Computations				
Posterior	$p(\mathbf{z} \mid \mathbf{y}) \propto \mathrm{e}^{-\mathcal{L}(\mathbf{z})}$; $\mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$	varies	what estimation and prediction integrate	2.3.6
Evidence	$p(\mathbf{y} \mid M) = \int p(\mathbf{y} \mid \mathbf{z}, M)\, p(\mathbf{z} \mid M)\, \mathrm{d}\mathbf{z}$	–	normalizing constant of a model	2.3.6
Preposterior	$p(\mathbf{z} \mid \mathbf{y}, \tilde{y}_a)$ averaged over $p(\tilde{y}_a \mid \mathbf{y})$	varies	the expected posterior under a measurement not yet taken	2.3.6
Information adds	$\mathbf{J}^{\mathsf{T}}\mathbf{J} + \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a$	varies	stacked rows of a candidate	2.3.6
Update form	$\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{J}_a^{\mathsf{T}}(\mathbf{I} + \mathbf{J}_a\boldsymbol{\Sigma}\mathbf{J}_a^{\mathsf{T}})^{-1}\mathbf{J}_a\boldsymbol{\Sigma}$	varies	$\boldsymbol{\Sigma}_a$ when $\boldsymbol{\Sigma}$ is singular	2.3.6
Building blocks of the fit				
State	$\mathbf{z} = (T_1, T_2, g)^{\mathsf{T}}$	varies	every free quantity of the graph at once	2.4.1
Rows and objective	$r_k(\mathbf{z})/\sigma_k$; $\mathcal{L} = \frac{1}{2}r^{\mathsf{T}}r$	–	physics, reading and prior rows in standard deviations	2.4.1
Jacobian	$\mathbf{J} = \partial r/\partial \mathbf{z}$	varies	sensitivity of each row to each entry	2.4.1

Term	Expression	Units	Meaning	Defined in
Normal equations	$(\mathbf{J}^{\mathsf{T}}\mathbf{J})\,\Delta\mathbf{z} = -\mathbf{J}^{\mathsf{T}}r$	varies	Gauss–Newton step to $\hat{\mathbf{z}}$	2.4.1
Laplace covariance	$\boldsymbol{\Sigma} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}$	varies	band of every entry of the state	2.4.1
Ridge of exact rows	$\boldsymbol{\Sigma} \to \sigma_g^2 \mathbf{v}\mathbf{v}^{\mathsf{T}}$ as $\sigma_p \to 0$	varies	entries tied by the physics move as one	2.4.1
Band of a derived quantity	$\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}$, $\mathbf{g} = \partial f/\partial \mathbf{z}$	varies	delta method	2.4.1
Specialized forms and answers				
Requirement probability of the hall	$\Phi\big((T_{\max} - \hat{z}_2)/\sqrt{\Sigma_{22}}\big) = 0.720$	–	reported, the normal tail of the entry	2.4.2
Its exact value	$\Phi\big((\hat{z}_3 - \ln(G^*/G_{\mathrm{ref}}))/\sqrt{\Sigma_{33}}\big) = 0.717$	–	the bound on T_2 mapped along the ridge to the third entry; gap 0.003	2.4.2
Exceedance rate	$\hat{\lambda} \pm \hat{\lambda}\,\sigma_\omega = 0.002667 \pm 0.001333$	h^{-1}	reported, from the log-rate	2.4.3
Exact posterior of the rate	$\mathrm{Gamma}(a + k,\ b + T_{\mathrm{tot}})$	–, h	the same exceedance log, conjugate	2.4.3
Design day within the committed load	$\sum_j w_j\, \mathrm{e}^{-\lambda_{\mathrm{ref}} \mathrm{e}^{\omega_j}\tau} = 0.9646$	–	reported, no exceedance averaged over the normal of the log-rate	2.4.4
Its exact value	$(b'/(b'+\tau))^{a'} = 0.9686$	–	the same average over the gamma posterior; gap 0.00406	2.4.4
Plug-in design day	$\mathrm{e}^{-\mathbb{E}[\lambda\mid\mathbf{y}]\tau}$	–	the same at a point estimate; below the exact value	2.4.5
Both halls within their committed loads	$1 - \Pr(A_A) - \Pr(A_B) + \Pr(A_A \cap A_B)$	–	inclusion and exclusion	2.4.6
Both halls above their committed loads	$\Pr(A_A \mid \mathbf{y})\ \Pr(A_B \mid A_A, \mathbf{y})$	–	intersection; product rule; a site-wide workload peak; coincidence	2.4.7
Model posterior	$\Pr(M_j \mid \mathbf{y}) \propto p(\mathbf{y} \mid M_j)\ \Pr(M_j)$	–	probability of a model	2.5.1
Bayes factor of the air path	$B_{AB} = p(\mathbf{y} \mid M_1)/p(\mathbf{y} \mid M_0)$, $M_A = M_1$, $M_B = M_0$	–	relative support from the evidence	2.5.1
Expected loss of an action	$\mathcal{R}(a) = c_a + C_e\ \Pr(e \mid \mathbf{y}, a)$	R$ thousand	renting a temporary chiller against waiting	2.5.4
Break-even probability	$\Pr(e \mid \mathbf{y}, \mathrm{rent}) + c_{\mathrm{rent}}/C_e$	–	the probability at which acting and waiting cost the same	2.5.4
Optimal action	$a^* = \arg\min_a \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$	–	decision by expected loss	2.5.5
Readings on the hall	$\boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}/(1 + 0.21302\,n)$	varies	every entry shrinks along the ridge	2.6.1
Value of information	$\mathrm{EVSI} = \bar{L}_{\mathrm{now}} - \bar{L}_{\mathrm{after}}$	R$ thousand	what a capacity test improves	2.6.2

Term	Expression	Units	Meaning	Defined in
The book's questions				
Drill questions	twenty planning questions	–	general form, computation and chapter of each	2.3.8

2.8 Exercises

Exercise 2.1 (Formulation)**.** For each statement, name the category of the question, the expression that answers it and the computation beneath it. (a) What is the probability that the hall stays within its committed load through the next 12 h of peak? ($\Pr(T > 12\,\mathrm{h} \mid \mathbf{y})$, a posterior predictive probability over the stated horizon; prediction under a threshold; the posterior.) (b) What is the hall's exceedance rate, given its exceedance log? ($p(\lambda \mid \mathbf{y})$ or a summary of it; estimation; the posterior.) (c) Two halls exceeded their committed loads in the same peak; is a common cause, a site-wide workload peak, more likely than two independent causes? ($\Pr(M_{\text{common}} \mid \mathbf{y})$ against $\Pr(M_{\text{independent}} \mid \mathbf{y})$, or their Bayes factor; hypothesis comparison; the evidence.) (d) In which week before the peak season should the temporary chiller be connected to minimize the expected cost of the season? ($\arg\min_a \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$ over the candidate weeks; a decision, in the design category; the expected consequence.) (e) Is the exceedance rate of a hall rising, and in which period will it pass its planning limit? ($\Pr(\gamma > 0 \mid \mathbf{y})$ and $\Pr(t^* \le s \mid \mathbf{y})$ over a state that carries the log-rate and its trend; prediction under a threshold; the trend.) (f) Of the questions A to D of Table 2.1, which two are answered from the same posterior without a new calculation of it? (A and B: an estimation and a threshold probability read from one posterior.)

Exercise 2.2 (Formulation)**.** A hall has a constant exceedance rate λ with a known posterior, and the site's planning target asks for at least a 99 % probability of a 24 h day within the committed load. (a) Formulate the posterior predictive probability of a day within the committed load and the probability that the hall's underlying rate meets the target. ($\mathbb{E}[\mathrm{e}^{-24\lambda} \mid \mathbf{y}] = \int_0^\infty \mathrm{e}^{-24\lambda} p(\lambda \mid \mathbf{y})\,\mathrm{d}\lambda$; $\Pr(\mathrm{e}^{-24\lambda} \ge 0.99 \mid \mathbf{y}) = \Pr(\lambda \le -\ln 0.99/24 \mid \mathbf{y}) = \Pr(\lambda \le 4.188 \times 10^{-4}\,\mathrm{h}^{-1} \mid \mathbf{y})$.) (b) Evaluate both from the posterior of the state, $\hat{\omega} = 0.98083$ and $\sigma_\omega = 0.5$, and set beside each the exact value from the posterior Gamma(4, 1500 h) of the hall's exceedance rate, using for the second the integer-shape identity $\Pr(\lambda \le x \mid \mathbf{y}) = 1 - \sum_{j=0}^{a'-1} \mathrm{e}^{-b'x}(b'x)^j/j!$. (From the state, equation (2.41) with $\tau = 24$ h gives 0.9307, and $\Phi\big((\ln 0.41876 - 0.98083)/0.5\big) = \Phi(-3.703) = 0.0001$. Exactly, $(1500/1524)^4 = 0.9385$, a gap of 0.0078; $b'x = 0.6281$ and the four Poisson terms 0.5336, 0.3352, 0.1053 and 0.0220 leave 0.0039, a gap of 0.0038.) (c) Explain why they differ. (The first averages the probability of no exceedance over the uncertainty in the rate and asks how likely the next day is to pass within the committed load; the second asks how much posterior probability supports the target on the rate itself. Here a day within the design load is likely, 0.93 and 0.94 exact, while the target is almost certainly not met by either route, because the posterior puts almost no mass below $4.2 \times 10^{-4}\,\mathrm{h}^{-1}$.)

Exercise 2.3 (Derivation)**.** The hall obeys $T = T_c + Q/G$ with $T_c = 18$°C and $Q = 60$ kW. (a) With the state, rows and $\boldsymbol{\Sigma}$ of equations (2.30) to (2.33), find the probability that the design-load temperature remains below 28°C from the normal of the entry T_2, then the exact value from the decreasing map between G and T_2 along the ridge. ($\Pr(T_2 < 28°\mathrm{C} \mid \mathbf{y}) \approx \Phi\big((28 - \hat{z}_2)/\sqrt{\Sigma_{22}}\big) = \Phi\big((28 - 27.231)/0.4615\big) = \Phi(1.667) = 0.952$. Exactly, along $\mathbf{v}$ the bound on T_2 is $G > Q_2/(28 - T_c) = 6$ kW/K, that is $g > \ln 6$ on the third entry, so $\Pr(T_2 < 28°\mathrm{C} \mid \mathbf{y}) = \int_6^\infty p(G \mid \mathbf{y})\,\mathrm{d}G = \Phi\big((\hat{z}_3 - \ln 6)/\sqrt{\Sigma_{33}}\big) = \Phi(1.601) = 0.945$; both are rows of Table 2.6. At $\kappa = 0.95$ the first accepts and the second does not, so both are stated.) (b) Now let the supply air be uncertain as well, as it is when the unit's controller is near saturation,

so that G and T_c have a joint posterior $p(G, T_c \mid \mathbf{y})$, and derive an integral for $\Pr(T > T_{\max} \mid \mathbf{y})$ without assuming that G and T_c are independent. (The event is $T_c + Q/G > T_{\max}$, so $\Pr(T > T_{\max} \mid \mathbf{y}) = \iint \mathbf{1}[T_c + Q/G > T_{\max}]\, p(G, T_c \mid \mathbf{y})\, \mathrm{d}G\, \mathrm{d}T_c$; the supply air becomes an entry of the state, and the joint density is required because the readings can leave it and the conductance correlated, a nonzero off-diagonal entry of $\mathbf{\Sigma}$.)

Exercise 2.4 (Derivation)**.** Two halls on one site have probabilities $\Pr(A_A) = 0.08$ and $\Pr(A_B) = 0.04$ that their loads exceed their committed loads in a day's peak. (a) Assuming independence, derive the probability that both exceed and the probability that both stay within their committed loads. ($\Pr(A_A \cap A_B) = 0.08 \times 0.04 = 0.0032$; $\Pr(\bar{A}_A \cap \bar{A}_B) = 0.92 \times 0.96 = 0.8832$.) (b) Now let both halls share a site workload whose site-wide peak occurs with probability $q_S = 0.01$ in the day, with the marginals unchanged. Derive the joint probability from equation (2.47), the conditional probability $\Pr(A_B \mid A_A)$ and the probability that both stay within their committed loads. ($q_A = 0.07/0.99 = 0.07071$, $q_B = 0.03/0.99 = 0.03030$; $\Pr(A_A \cap A_B) = 0.01 + 0.99 \times 0.07071 \times 0.03030 = 0.01212$; $\Pr(A_B \mid A_A) = 0.01212/0.08 = 0.152$; $\Pr(\bar{A}_A \cap \bar{A}_B) = 1 - 0.12 + 0.01212 = 0.89212$; the probability of both exceeding is 3.8 times the independent value: the coincidence that a product of marginals misses.)

Exercise 2.5 (Derivation)**.** The committed-load exceedances of a sister hall whose load is growing are counted in six periods of forty days, period j centered at $t = j$ so that the log ends at $t = 6.5$, the first period with 36 logged days, and number 5, 6, 7, 9, 11 and 14. The log exceedance rate, in exceedances per logged day, is carried as a straight line in time, $f_t(\mathbf{z}) = \alpha + \gamma(t - \bar{t})$ with t in periods and $\bar{t} = 3.5$, and the fit of the six count rows with flat priors gives $\hat{\mathbf{z}} = (-1.5768,\ 0.1989)^{\mathsf{T}}$ and $\mathbf{\Sigma}$ with entries $\Sigma_{\alpha\alpha} = 0.021811$, $\Sigma_{\alpha\gamma} = -0.004328$ and $\Sigma_{\gamma\gamma} = 0.007259$. The planning limit is 0.5 exceedances per logged day, $f^* = \ln 0.5$. (a) From the trend form (2.21) under the Laplace posterior, find the probability that the rate is rising. ($\Phi(0.1989/\sqrt{0.007259}) = \Phi(2.335) = 0.990$.) (b) Find the crossing time at the MAP state and say why it is the median of the crossing time. ($t^*(\hat{\mathbf{z}}) = 3.5 + (\ln 0.5 + 1.5768)/0.1989 = 7.94$ periods, $(7.94 - 6.5) \times 40 = 58$ days after the log ends; the argument of the normal tail in equation (2.22) is zero at $s = t^*(\hat{\mathbf{z}})$, so the distribution function is one half there.) (c) Find the probability that the rate has crossed the limit by $t = 8$. ($\mathbf{g}(8) = (1,\ 4.5)^{\mathsf{T}}$, $\mathbf{g}(8)^{\mathsf{T}}\hat{\mathbf{z}} = -0.6816$, $\mathbf{g}(8)^{\mathsf{T}}\mathbf{\Sigma}\,\mathbf{g}(8) = 0.021811 - 9 \times 0.004328 + 20.25 \times 0.007259 = 0.1299$, so $\Phi((-0.6816 + 0.6931)/0.3604) = \Phi(0.032) = 0.513$.) (d) Explain why the crossing time has no posterior mean. ($t^* = \bar{t} + (f^* - \alpha)/\gamma$ divides by the trend entry, whose Gaussian posterior has a positive density at $\gamma = 0$, and the integral of $1/\gamma$ against such a density diverges.)

Exercise 2.6 (Application)**.** With costs in R$ thousand, as in Section 2.5.4: the probability of an excursion at the coming season's peak is 0.004 if the planner waits, and an excursion costs R$ 250 000. A temporary chiller sized for the whole hall, connected for the season, costs R$ 400 and lowers the probability to 0.001. (a) Derive the expected cost of each action, and state which action minimizes it. (The expected consequence (2.23) of each action, probability times consequence plus the certain cost. Waiting: $0.004 \times$ R$ 250 000 = R$ 1000; acting: R$ 400 + $0.001 \times$ R$ 250 000 = R$ 650; acting minimizes the expected cost, with an expected saving of R$ 350.) (b) Find the cost of the action at which the decision would reverse. ($(0.004 - 0.001) \times$ R$ 250 000 = R$ 750.)

Exercise 2.7 (Application)**.** Two models are competing explanations of a hall sensor that reads warmer than the design of the hall expects, M_0 a sound sensor and M_1 a biased one, with $\Pr(M_0) = \Pr(M_1) = 0.5$ and evidences $p(\mathbf{y} \mid M_0) = 0.02$ and $p(\mathbf{y} \mid M_1) = 0.08$. (a) Derive the Bayes factor B_{10} and the posterior probability $\Pr(M_1 \mid \mathbf{y})$. ($B_{10} = 0.08/0.02 = 4$, $\ln B_{10} = 1.386$ nats; $\Pr(M_1 \mid \mathbf{y}) = 0.08 \times 0.5/(0.08 \times 0.5 + 0.02 \times 0.5) = 0.80$, assuming that the two models exhaust the possibilities under consideration.) (b) Repeat with $\Pr(M_1) = 0.2$, and say what the comparison of the two answers shows. ($0.08 \times 0.2/(0.08 \times 0.2 + 0.02 \times 0.8) = 0.50$; the Bayes

factor is the same 4 in both cases, because it depends only on the evidence, while the posterior model probability also depends on the prior.)

Chapter 3

Uncertainty and Conditional Probability

3.1 Uncertainty and evidence in design questions

The design review for taking the hall to its 60 kW design load has reached the existing cooling unit, and the design sheet carries one note against it: the unit's capacity-shortfall flag, the monitoring check meant to catch a unit that delivers less than its rating, came up on yesterday's peak day. The check compares the cooling the unit delivers with its rating, a quantity the design sheet does not show. The hall carried its 30 kW, the hall sensor read 23.2°C, and the supply air held at its 18°C set point. The flag's data sheet says that it catches nine shortfalls of the unit in ten. Read from the sheet, the conclusion is quick: the flag is right nine times in ten, so the unit is short with probability 0.9, its rating cannot be counted at the design load, and a second unit belongs in the design. The reviewer also knows that a second hall, hall B, serves the same site's workload, and wonders whether the two halls' peak days can be assumed to fall apart, so that the shared plant need not carry both at once.

The data sheet does not settle either point. It states how often the flag comes up when the unit is short, which is not how often the unit is short when the flag comes up; it says nothing about how often shortfalls occur, nor about the flags that come up on a sound unit; and it says nothing about what the two halls share. Four situations of this kind recur through the chapter. The hall sensor scatters from window to window, and the probability that the next reading crosses the 24°C limit is wanted. The flag has come up, and the probability that the unit is short is wanted. Two halls are driven by one site-wide workload, and the probability that both exceed their committed loads on the same day is wanted. The hall has one reading, and the probability that the hall air exceeds its limit is wanted. Chapter 1 represented the hall as a plant graph whose free quantities form the state vector $\mathbf{z}$, and Chapter 2 wrote the planner's questions as probabilities over that state, with the unknown on the left of the conditioning bar and the evidence on the right. The question of this chapter is the mechanism between the two:

how new evidence changes what is known about an existing plant before a design commits to it, and how dependencies between halls, units and physical quantities affect a probability calculation.

Jargon: aleatory and epistemic uncertainty, conditional probability, prior, likelihood, evidence, posterior, marginalization, nuisance quantity, conditional independence, common cause

Aleatory uncertainty is the variability of outcomes that the model treats as random even when its parameters are known, such as the reading a sensor will show in the next window when the hall air is known. *Epistemic* uncertainty is incomplete knowledge of quantities that have one true value, such as the conductance of the hall's air path or the calibration offset of a meter; more evidence can reduce it. A *conditional probability* $\Pr(A \mid B)$ is the probability of event A among the situations in which event B occurs. In Bayes' theorem the *prior* is the probability of a hypothesis before an observation, the *likelihood* is the probability of the observation under the hypothesis, the *evidence* (also *marginal likelihood*) is the probability of the observation averaged over the hypotheses, and the *posterior* is the probability of the hypothesis after the observation. *Marginalization* removes an unknown from a joint distribution by summing or integrating over it; an unknown removed this way because the question does not ask about it is a *nuisance quantity*. Two quantities are *conditionally independent* given a third when, once the third is known, learning one says nothing more about the other. A *common cause* is a shared condition, such as a site-wide workload that drives the loads of two halls, that raises the probability of an exceedance in several places at once.

Key idea

Conditioning on evidence restricts attention to the situations consistent with it and renormalizes; Bayes' theorem does this for a hypothesis, marginalization averages over every unknown the question does not name, and the dependence among unknowns, through a common cause or a shared physical relation, is what lets one observation inform another quantity and what an independence assumption silently discards.

A design review carries two kinds of not knowing. The first is what the plant does from day to day: the scatter of the load and of the readings, which a design must carry and which no survey removes. The second is what is not known about the plant itself: the conductance of its air path, the offset of a load meter, which a test or a survey removes. Keeping them apart changes the probability that a reading crosses a limit, decides which uncertainty a measurement before committing can buy down, and gives each kind its own distribution.

3.1.1 Aleatory variability from window to window

Suppose the hall air were known exactly: take the estimate of Chapter 1, $T = 23.07\,^{\circ}\mathrm{C}$, as if it were the truth. Even then the reading of the next five-minute window cannot be predicted. Let y_T be the reading of one window. The sensor adds an error that changes from window to window, $y_T \mid T \sim \mathcal{N}(T, \sigma_T^2)$ with $\sigma_T = 0.5\,\mathrm{K}$, so the probability that the reading of a window shows above a level c is the integral of the normal density from c to infinity,

$$\Pr(y_T > c \mid T) = \int_c^{\infty} \frac{1}{\sigma_T}\,\varphi\!\left(\frac{s - T}{\sigma_T}\right) \mathrm{d}s = 1 - \Phi\!\left(\frac{c - T}{\sigma_T}\right), \tag{3.1}$$

where s is the variable of integration and φ and Φ are the standard normal density and cumulative distribution function. For the 24°C limit,

$$\Pr(y_T > 24\,^{\circ}\mathrm{C} \mid T = 23.07\,^{\circ}\mathrm{C}) = 1 - \Phi\!\left(\frac{24 - 23.0701}{0.5}\right) = 1 - \Phi(1.8598) = 0.0315.$$

With the hall air perfectly known there is still a probability 0.0315, one window in 32, that the reading shows the hall above its limit; over the twelve peak hours of a design day, 144 windows,

a design check that counted the readings above 24°C would count about 4.5 of them with the hall inside its limit throughout. This uncertainty arises from the variability that the reading process itself represents, and it is called *aleatory*: it is what the plant does from day to day, the design must carry it, and no survey of the plant removes it. The IT load is a second source of it: server utilization moves the load from window to window, and a scatter of 1.5 kW about 30 kW (5 %, illustrative) moves the hall air itself by $1.5/5.917 = 0.2535$ K at the estimated conductance 5.917 kW/K. The reading then scatters by $\sqrt{0.5^2 + 0.2535^2} = 0.561$ K and crosses 24°C in 0.0486 of the windows. The rest of the chapter holds the load fixed within a window and keeps the sensor's scatter alone.

Figure 3.1(a) shows it. The horizontal axis numbers twelve windows of the same hall with the same known air temperature, and the vertical axis is the reading of each. The dots are the twelve quantiles of the normal distribution $\mathcal{N}(23.07, 0.5^2)$, placed in an arbitrary order; the dashed line marks the true hall air and the red line the limit. Identical windows of an identical, exactly known hall read from 22.20 to 23.94°C: knowing the hall fixes the distribution of the reading, not the reading.

3.1.2 Epistemic uncertainty about the plant

Now suppose the hall air is not known, because what fixes it at a given load, the effective conductance G of the air path to the supply air, is not. The handbook gives 7.5 kW/K, right to a factor of about 1.35. What is known about the conductance is then a probability distribution given the evidence, $p(G \mid \mathbf{y})$, or equivalently $p(g \mid \mathbf{y})$ for the log-conductance $g = \ln(G/G_{\text{ref}})$ with $G_{\text{ref}} = 6$ kW/K. The conductance might plausibly be $7.5/1.35 = 5.556$, 7.5 or $7.5 \times 1.35 = 10.125$ kW/K, with different probabilities, or a density, assigned to those values. This is *epistemic* uncertainty: uncertainty caused by incomplete knowledge of the characteristics of the plant, which additional observations can reduce: an airflow survey or a capacity test of the hall narrows it before a design commits to the handbook value. The calibration of a meter is the other common case. A load meter that reads high by a fixed, unknown offset reads high in every window; the offset has one true value, and the evidence that reduces what is known about it is a comparison with an independent measurement, not more windows of the same meter.

Figure 3.1(b) draws such a distribution. The horizontal axis is the conductance in kW/K and the vertical axis the prior density. The curve is the handbook prior of Chapter 1, $g \sim \mathcal{N}(\ln(7.5/6), 0.3^2)$, drawn as the lognormal density of G; the dashed line is its median, 7.5 kW/K, and the three dots are the three plausible values named above. The conductance is a single number; the spread of the curve is the spread of knowledge about it.

3.1.3 The distinction on the hall

In a plant under design the distinction can depend on what information is available and on how the model is defined. The load on a future design day may be modeled as random, aleatory, while the load at the present moment is a fixed but poorly metered quantity, epistemic. The distinction is therefore a statement about the model, and it changes the calculation.

If the hall air is known, the probability that the next reading shows above the limit is equation (3.1). If the hall air is uncertain, the probability that the next reading y_{new} crosses c is the conditional probability averaged over what is known about the hall, which is the law of total probability derived in Section 3.4.2:

$$\Pr(y_{\text{new}} > c \mid \mathbf{y}) = \int \Pr(y_{\text{new}} > c \mid T)\, p(T \mid \mathbf{y})\, \mathrm{d}T = \int \left[1 - \Phi\left(\frac{c - T}{\sigma_T}\right)\right] p(T \mid \mathbf{y})\, \mathrm{d}T. \quad (3.2)$$

For three plausible values the integral is a sum. Before any reading, the three conductances of Section 3.1.2 put the hall air at $T = 18 + 30/G = 23.40$, 22.00 and 20.96°C. With the illustrative

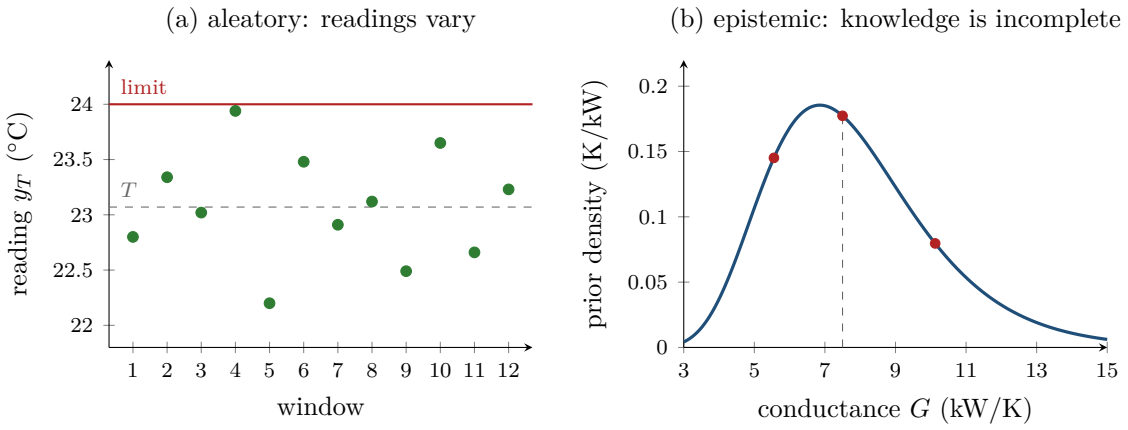


Figure 3.1: The two kinds of uncertainty. (a) Readings of twelve windows of a hall whose air is known, $T = 23.07$°C (dashed), with a sensor of standard deviation 0.5 K: the normal quantiles $T + 0.5\,\Phi^{-1}((i - 0.5)/12)$, $i = 1, \dots, 12$, in an arbitrary order; red: the 24°C limit. (b) The handbook prior of the conductance, $\ln(G/6) \sim \mathcal{N}(\ln 1.25, 0.3^2)$, density $\exp[-\ln^2(G/7.5)/0.18]/(0.3\sqrt{2\pi}\,G)$; dashed: its median 7.5 kW/K; dots: the plausible values 5.556, 7.5 and 10.125 kW/K.

probabilities 0.25, 0.50 and 0.25 on them, the probability that the next reading crosses 24°C is

$$\Pr(y_{\text{new}} > 24°\text{C}) = 0.25 \times 0.1151 + 0.50 \times (3.2 \times 10^{-5}) + 0.25 \times (6 \times 10^{-10}) = 0.0288,$$

against 1.5×10^{-5} at the mean conductance 7.670 kW/K, where $T = 21.91$°C: the average is 1953 times the value at the average, and almost all of it comes from the low conductance. For a Gaussian posterior $T \mid \mathbf{y} \sim \mathcal{N}(\mu, \Sigma_{TT})$ the integral has a closed form. The next reading is $y_{\text{new}} = T + \epsilon$ with the error $\epsilon \sim \mathcal{N}(0, \sigma_T^2)$ independent of T, and the sum of two independent normal quantities is normal with the sum of their variances, a standard result, so that

$$y_{\text{new}} \mid \mathbf{y} \sim \mathcal{N}(\mu,\ \Sigma_{TT} + \sigma_T^2), \qquad \Pr(y_{\text{new}} > c \mid \mathbf{y}) = 1 - \Phi\left(\frac{c - \mu}{\sqrt{\Sigma_{TT} + \sigma_T^2}}\right). \tag{3.3}$$

After the one reading of Chapter 1, $\mu = 23.07$°C and $\sqrt{\Sigma_{TT}} = 0.481$ K, the predictive standard deviation is $\sqrt{0.4806^2 + 0.5^2} = 0.6935$ K and the probability is $1 - \Phi(1.341) = 0.0900$, against 0.0315 for the plug-in that takes 23.07°C as the truth. In the three-value sum the average of the tail exceeds the tail at the average because $1 - \Phi((c - T)/\sigma_T)$ is convex in T below the limit, where all three values lie. In the Gaussian case the reason is that the limit lies above the mean: widening the spread from 0.5 to 0.6935 K moves probability past it.

The scatter of the next reading remains even when the uncertainty about the hall is removed. Table 3.1 repeats the fit of Chapter 1 with n windows at 30 kW whose readings average 23.2°C, so that the state $(T, g)^{\mathsf{T}}$ carries one physics row, n reading rows and the handbook prior row. The first column is the number of windows, the second and third the posterior mean and standard deviation of the hall air, the fourth the predictive standard deviation of equation (3.3), and the last two the probability that the hall air exceeds 24°C and the probability that the next reading does. The epistemic spread halves each time the windows quadruple, and the probability that the hall is above its limit falls from 0.0265 to below 10^{-10}; the predictive spread falls only to the sensor's 0.5 K, and the probability that the next reading crosses the limit tends to $1 - \Phi(1.6) = 0.0548$, not to zero. The sensor's error is taken as independent from window to window, with no fixed offset; an offset would set a floor under $\sqrt{\Sigma_{TT}}$ that more windows cannot remove.

The table reports the Laplace values of the state's posterior. The exact posterior of the same rows, with T integrated out and a quadrature along g, sets the same two events beside them. The probability that the hall air exceeds 24°C is 0.0265 against the exact 0.0225 at $n = 1$, 3.76×10^{-4} against 3.27×10^{-4} at $n = 4$, and 4.43×10^{-11} against 3.88×10^{-11} at $n = 16$: the absolute gap closes while the relative gap barely moves, 14.9, 13.2 and 12.4 %. The probability that the next reading crosses is 0.0900 against 0.0813, 0.0674 against 0.0648, and 0.0583 against 0.0576. The source is the curve of the physics row: the Laplace normal is centered on the mode, 23.0701°C at $n = 1$, and the exact posterior has the mean 23.0300°C, so the exact tail above the limit is the thinner one.

Table 3.1: The hall at 30 kW after n windows whose readings average 23.2°C (the fit of Chapter 1 with n reading rows): the epistemic spread of the hall air shrinks, the scatter of the next reading does not.

n [–]	$\hat{T}$ [°C]	$\sqrt{\Sigma_{TT}}$ [K]	$\sqrt{\Sigma_{TT} + \sigma_T^2}$ [K]	$\Pr(T > 24°\text{C})$ [–]	$\Pr(y_{\text{new}} > 24°\text{C})$ [–]
1	23.070	0.4806	0.6935	0.0265	0.0900
4	23.166	0.2476	0.5580	3.8×10^{-4}	0.0674
16	23.191	0.1247	0.5153	4×10^{-11}	0.0583
64	23.198	0.0625	0.5039	$< 10^{-30}$	0.0557
∞	23.200	0	0.5000	0	0.0548

Principle 3.1. *Collecting more windows can reduce the uncertainty about the hall without eliminating the scatter of the next reading: evidence removes epistemic uncertainty, and the aleatory part remains as long as the reading is modeled as scattering from window to window.*

Table 3.1 is the first argument of the questions on the inputs' uncertainty, questions 13 to 15 of Chapter 2. A measurement before committing to a design can buy down only the epistemic part, the last column's distance above its floor of 0.0548; the scatter at the floor is what the design must carry whatever is measured. Chapters 17 to 19 rank the measurements and the pilots by the epistemic width each would remove from the quantity that decides the build.

3.1.4 Distributions of what is uncertain

Three mathematical objects represent what is uncertain, and Figure 3.2 draws one of each.

A discrete unknown. Suppose a cooling unit (illustrative) either delivers its rating or falls below it, and define the indicator $F = 1$ if it is below its rating and $F = 0$ if it delivers it. The assignment $\Pr(F = 1) = 0.10$, $\Pr(F = 0) = 0.90$ is a discrete probability distribution: each value receives a probability, and the probabilities sum to one. Panel (a) of the figure has the two values on the horizontal axis and their probabilities as bars.

A continuous unknown. Suppose the effectiveness η of a unit's coil, the fraction of the largest possible heat transfer that it achieves, is unknown, and represent it by a density $p(\eta)$. A probability over a range is the integral of the density over that range,

$$\Pr(0.75 < \eta < 0.85) = \int_{0.75}^{0.85} p(\eta)\,\mathrm{d}\eta.$$

For a continuous variable the density itself is not a probability: it is a probability per unit of the variable, and it can exceed one. Panel (b) uses the illustrative density $\mathcal{N}(0.80, 0.04^2)$. The

horizontal axis is the effectiveness and the vertical axis the density, whose peak is $1/(0.04\sqrt{2\pi}) = 9.97$; the shaded area between 0.75 and 0.85 is $\Phi(1.25) - \Phi(-1.25) = 0.789$, the probability of that range.

A joint distribution. Suppose an illustrative hall with a weak air path has two unknown quantities, the conductance G of its air path and the supply-air temperature T_c, as it does when the discharge controller's own sensor is not trusted. Their *joint* distribution $p(G, T_c)$ describes the uncertainty about both together, including any dependence between them. The joint distribution is essential when a design quantity depends on several uncertain inputs, as the hall air does,

$$T = T_c + \frac{Q}{G},$$

with the heat load Q in kW and G in kW/K. The individual distributions $p(G)$ and $p(T_c)$ are not always enough to determine the distribution of T; how G and T_c vary together is needed as well. Panel (c) draws two joint distributions with the same marginals, G with mean 3 kW/K and standard deviation 0.15 kW/K, T_c with mean 20°C and standard deviation 1 K, one with correlation $\mathrm{corr}(G, T_c) = 0$ and one with $\mathrm{corr}(G, T_c) = 0.8$. The horizontal axis is G, the vertical axis T_c, and each ellipse encloses 95 % of its distribution. Both ellipses span the same ranges of G and of T_c, so the marginals cannot tell them apart; Section 3.6.4 shows that at $Q = 30$ kW the hall air has a standard deviation of 1.12 K under the first and 0.67 K under the second.

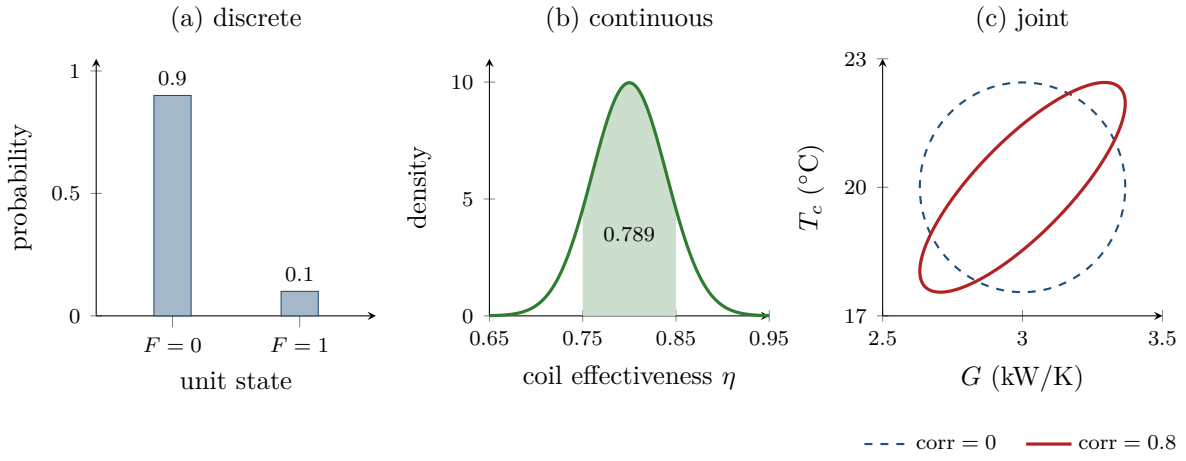


Figure 3.2: Distributions of what is uncertain (illustrative). (a) A discrete unknown: $\Pr(F = 1) = 0.10$ for a cooling unit below its rating. (b) A continuous unknown: the density $\mathcal{N}(0.80, 0.04^2)$ of a coil effectiveness; shaded, $\Pr(0.75 < \eta < 0.85) = 0.789$. (c) Two joint distributions of G and T_c with the same normal marginals (means 3 kW/K and 20°C, standard deviations 0.15 kW/K and 1 K) and correlations 0 and 0.8; the ellipses enclose 95 % of each ($\chi^2_{2,0.95} = 5.991$).

3.2 Conditional probability and Bayes' theorem

Conditional probability is the foundation of every update in this book. A peak-day log, read day by day, shows where it comes from, why the direction of the conditioning matters, and how Bayes' theorem reverses it.

3.2.1 Conditional probability from a peak-day log

A hall's *committed load* is the load allocated to its tenants, 30 kW for this hall today; brief workload peaks lift the metered load above it. Let H be the event that a hall's IT load rises above its committed load during a day, the event of the exceedance log of Chapter 2, and E the event that the inlet air of some rack in the hall exceeds its limit during the same day, an inlet excursion, established afterwards from a review of independent readings. The question *given that the load rose above the committed load, what is the probability that the day brought an excursion* is written $\Pr(E \mid H)$. The phrase "given that" restricts attention to the situations in which H occurred.

Imagine 1000 comparable hall-days of a site's peak-day log, gathered over its halls and its peak seasons and recorded in Table 3.2. The rows say whether the load rose above the committed load, the columns whether an inlet excursion occurred, and the margins are the totals: load above committed and an excursion together on 20 days, the load above committed alone on 60, the excursion alone on 20, and neither on 900. The days are pooled over the site's halls, each against its own committed load, not drawn from this hall's log of Chapter 2. Some of those halls are in their growth years, with tenants' loads pressing on their allocations, and their workload peaks cross the committed load more often than this hall's. The load rises above committed on 0.08 of these days, against a probability of 0.0354 of an exceedance in a design day for this hall, Chapter 2's value (0.0314 exact), 2.26 times as often (2.55 against the exact value).

Table 3.2: Outcomes of 1000 comparable hall-days of a peak-day log (illustrative), in numbers of days.

	inlet excursion [–]	no excursion [–]	total [–]
load above committed	20	60	80
load within committed	20	900	920
total	40	960	1000

From these counts,

$$\Pr(H) = \frac{80}{1000} = 0.08, \qquad \Pr(E) = \frac{40}{1000} = 0.04, \qquad \Pr(H \cap E) = \frac{20}{1000} = 0.02.$$

Among the 80 days on which the load rose above the committed load, an excursion came on 20. Therefore

$$\Pr(E \mid H) = \frac{20}{80} = 0.25.$$

The conditional probability is 25 %, against an unconditional probability of 4 % that a day brings an excursion: in these observations load above committed and excursions are strongly associated, and still on three above-committed days in four the halls carried the excess without an excursion.

3.2.2 The definition

Dividing the numerator and the denominator of 20/80 by the total number of days turns the count ratio into a ratio of probabilities,

$$\Pr(E \mid H) = \frac{20/1000}{80/1000} = \frac{\Pr(H \cap E)}{\Pr(H)}.$$

For any two events A and B with $\Pr(B) > 0$ the definition is

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}. \tag{3.4}$$

Conditioning on B restricts the sample space to the outcomes in B and renormalizes, so that the restricted probabilities again sum to one. The definition requires the conditioning event to have positive probability. Two events are *independent* when $\Pr(A \cap B) = \Pr(A)\,\Pr(B)$, which by equation (3.4) is the same as $\Pr(A \mid B) = \Pr(A)$: learning B does not change the probability of A.

Table 3.3 works the definition for four joint probabilities with the marginals of the log held fixed, $\Pr(H) = 0.08$ and $\Pr(E) = 0.04$. The first column is the probability that a day has both the load above committed and an excursion, the next two are the conditional probabilities in the two directions, and the last says what kind of association the row describes. The third row is the log of Table 3.2: $\Pr(E \mid H) = 0.02/0.08 = 0.25$ and $\Pr(H \mid E) = 0.02/0.04 = 0.50$. The second row is the independent case, in which $\Pr(E \mid H) = 0.0032/0.08 = 0.04 = \Pr(E)$: excursions that come as often on light days as on heavy ones say nothing about the load. The joint probability cannot exceed the smaller marginal, 0.04; at that bound every excursion comes on a day above the committed load.

Table 3.3: Conditional probabilities from equation (3.4) for four joint probabilities of load above committed and an inlet excursion, with $\Pr(H) = 0.08$ and $\Pr(E) = 0.04$ held fixed.

$\Pr(H \cap E)$ [–]	$\Pr(E \mid H)$ [–]	$\Pr(H \mid E)$ [–]	Kind of association
0	0	0	no excursion on an above-committed day
0.0032	0.04	0.08	excursions independent of the load
0.02	0.25	0.50	the logged days: positively associated
0.04	0.50	1.00	every excursion on an above-committed day

3.2.3 The multiplication rule

Multiplying equation (3.4), written for $\Pr(B \mid A)$, by $\Pr(A)$ gives the multiplication rule,

$$\Pr(A \cap B) = \Pr(B \mid A)\,\Pr(A). \tag{3.5}$$

Conditioning in the other direction gives $\Pr(A \cap B) = \Pr(A \mid B)\,\Pr(B)$, and the two expressions for the same intersection are equal:

$$\Pr(B \mid A)\,\Pr(A) = \Pr(A \mid B)\,\Pr(B). \tag{3.6}$$

For the log, $0.25 \times 0.08 = 0.02 = 0.50 \times 0.04$. This relation is the starting point of Bayes' theorem.

Reversing the conditioning. The two conditional probabilities of the log differ: $\Pr(E \mid H) = 0.25$, but $\Pr(H \mid E) = 20/40 = 0.50$. The first asks about the excursion among the days above the committed load, the fraction of heavy days that the halls could not carry cleanly; the second asks about the load among the days with an excursion, the fraction of excursions

that came on days above the committed load. Half the excursions came on days within the committed load: they are not load but airflow, recirculation at an open rack position or a supply path that starves one row, and more cooling capacity would not have prevented them. Equation (3.6) shows that the two are equal only when $\Pr(H) = \Pr(E)$.

Common mistake

Treating $\Pr(A \mid B)$ and $\Pr(B \mid A)$ as the same number. Reversing the conditioning changes the question. In a design review the probability of an excursion given a day above the committed load is what a load test measures, and the probability that an excursion came from load, the one that justifies buying capacity, is what the planner needs; the second follows from the first only through Bayes' theorem and the base rates of the two events.

3.2.4 Bayes' theorem

Dividing both sides of equation (3.6) by $\Pr(B) > 0$ gives Bayes' theorem for discrete events,

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\,\Pr(A)}{\Pr(B)}. \tag{3.7}$$

For the log, the fraction of excursions on above-committed days and the base rates give the fraction of above-committed days that brought an excursion, $\Pr(E \mid H) = 0.50 \times 0.04 / 0.08 = 0.25$, the value read from the counts. In words,

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}},$$

and the four terms have distinct roles when A is a hypothesis about the plant and B an observation. The *prior* $\Pr(A)$ is what was believed about the hypothesis before the observation. The *likelihood* $\Pr(B \mid A)$ is how probable the observation is when the hypothesis holds. The *evidence* $\Pr(B)$ is the overall probability of the observation, averaged over the hypotheses; it is the normalizing constant of the posterior and, for a whole model M, the quantity the notation writes $p(\mathbf{y} \mid M)$. The *posterior* $\Pr(A \mid B)$ is what is believed about the hypothesis after the observation is accounted for. The reviewer's question "the shortfall flag came up on the peak day; what is the probability the unit is really short" is this computation, and Chapter 16 asks it of the site's chillers as the as-built margin.

3.3 Bayesian diagnosis of a capacity shortfall

Bayes' theorem turns the known behavior of a flag into the probability of the shortfall the flag is meant to detect, and the answer of the opening scene is far below the flag's detection probability; a day without a flag is evidence too.

3.3.1 The diagnosis problem

The hall's cooling unit carries a capacity-shortfall flag, the flag of the opening scene, designed to detect a unit that delivers less than its rating on the day. Define the events

$$F = \{\text{the cooling unit is short of its rating}\}, \qquad H = \{\text{the shortfall flag comes up}\}.$$

The test and maintenance record suggests that the unit is short on 5 % of days, $\Pr(F) = 0.05$. The flag comes up on 90 % of short days, $\Pr(H \mid F) = 0.90$, its detection probability; it also comes up falsely, on 10 % of days with a sound unit, $\Pr(H \mid \neg F) = 0.10$, its false-flag probability, the rate of the flag on a sound plant. The flag H of this section is a different event from the

day above the committed load of Section 3.2.1, and its probability, derived below, differs from the log's 0.08. A flag is observed, and the question is the probability that the unit is in fact short, $\Pr(F \mid H)$.

The first step writes Bayes' theorem, equation (3.7), for the hypothesis F and the observation H:

$$\Pr(F \mid H) = \frac{\Pr(H \mid F)\ \Pr(F)}{\Pr(H)}.$$

The numerator is known, $\Pr(H \mid F)\ \Pr(F) = 0.90 \times 0.05 = 0.045$. The denominator, the probability of a flag, is not yet known.

3.3.2 The total probability of a flag

A flag can occur in two mutually exclusive situations: the unit is short and the flag comes up, or the unit is sound and the flag comes up falsely. In this simplified model the two exhaust the cases, so the flag event is the union of two disjoint events and its probability is their sum,

$$\Pr(H) = \Pr(H \cap F) + \Pr(H \cap \neg F).$$

The multiplication rule, equation (3.5), writes each intersection as a likelihood times a prior:

$$\begin{aligned} \Pr(H) &= \Pr(H \mid F)\ \Pr(F) + \Pr(H \mid \neg F)\ \Pr(\neg F) \\ &= 0.90 \times 0.05 + 0.10 \times 0.95 = 0.045 + 0.095 = 0.14. \end{aligned} \tag{3.8}$$

Flags come up on 14 % of days; two thirds of them, 0.095 of the 0.14, come from days with a sound unit.

3.3.3 The posterior shortfall probability

The third step divides,

$$\Pr(F \mid H) = \frac{0.045}{0.14} = 0.3214. \tag{3.9}$$

The second unit of the opening scene therefore does not rest on a probability of 0.9: at 0.32 a sound unit is the likelier explanation, and whether the design should discount the unit's rating depends on what a capacity test of the unit costs against what an undersized design costs. Figure 3.3 compares the probability of a shortfall before and after the observation. Each row is one state of knowledge, the horizontal scale is a probability from 0 to 1, and the green part of each bar is the probability of a shortfall. The flag raises it from 5 % to 32.1 %; it increases suspicion substantially, but it does not establish that the unit is short. The third row is the case without a flag, derived in Section 3.3.4.

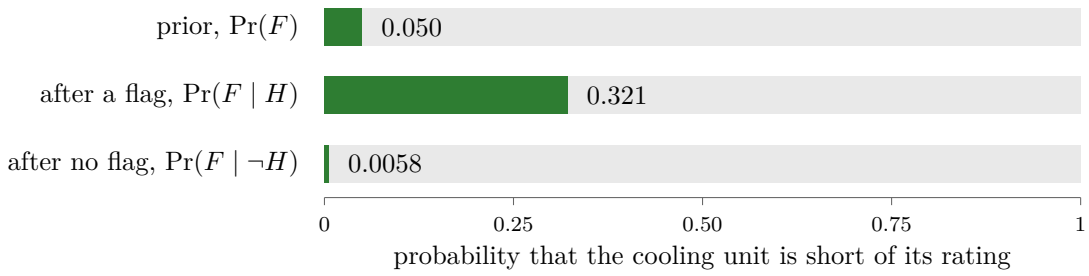


Figure 3.3: What a flag tells about a capacity shortfall. Prior $\Pr(F) = 0.05$ per day; flag with detection probability $\Pr(H \mid F) = 0.90$ and false-flag probability $\Pr(H \mid \neg F) = 0.10$. After a flag, $0.045/0.14 = 0.321$ (equation (3.9)); after no flag, $0.005/0.86 = 0.0058$ (equation (3.11)).

The base rate. The answer is not 90 %. The flag detects 90 % of shortfalls, but shortfalls are uncommon, and many flags come up while the unit is sound. Table 3.4 makes this concrete for 1000 days. The rows are the actual condition of the unit, the columns whether the flag came up, and each entry is the expected number of days: of the 50 short days, 90 %, or 45, raise the flag; of the 950 sound days, 10 %, or 95, raise it falsely. Among the 140 days with a flag only 45 have a short unit, so $\Pr(F \mid H) = 45/140 = 0.3214$. The table and Bayes' theorem give the same result, because the table is equation (3.8) multiplied by 1000.

Table 3.4: Expected outcomes of 1000 days with $\Pr(F) = 0.05$, $\Pr(H \mid F) = 0.90$ and $\Pr(H \mid \neg F) = 0.10$, in numbers of days.

Actual condition	flag [–]	no flag [–]	total [–]
short of its rating	45	5	50
sound	95	855	950
total	140	860	1000

The odds form. Writing Bayes' theorem for F and for $\neg F$ with the same observation and dividing the two, the evidence $\Pr(H)$ cancels:

$$\frac{\Pr(F \mid H)}{\Pr(\neg F \mid H)} = \frac{\Pr(H \mid F)}{\Pr(H \mid \neg F)} \cdot \frac{\Pr(F)}{\Pr(\neg F)}. \tag{3.10}$$

Posterior odds equal the likelihood ratio times the prior odds. Here the likelihood ratio is $0.90/0.10 = 9$ and the prior odds are $0.05/0.95 = 0.05263$, so the posterior odds are 0.4737 and the posterior probability is $0.4737/(1 + 0.4737) = 0.3214$, as before. The flag multiplies the odds of a shortfall by nine whatever the base rate; the base rate decides where that multiplication starts.

3.3.4 The case without a flag

The same derivation applies to the observation that the flag does not come up. Bayes' theorem gives

$$\Pr(F \mid \neg H) = \frac{\Pr(\neg H \mid F)\ \Pr(F)}{\Pr(\neg H)},$$

with $\Pr(\neg H \mid F) = 1 - 0.90 = 0.10$, the probability that a short unit goes undetected, and, by the same total probability as equation (3.8),

$$\Pr(\neg H) = 0.10 \times 0.05 + 0.90 \times 0.95 = 0.005 + 0.855 = 0.86.$$

Therefore

$$\Pr(F \mid \neg H) = \frac{0.005}{0.86} = 0.00581. \tag{3.11}$$

The absence of a flag reduces the shortfall probability from 5 % to about 0.58 %, the third row of Figure 3.3; in odds, a day without a flag multiplies the prior odds by $0.10/0.90 = 1/9$. In this example a flag substantially raises suspicion, and its absence is considerable evidence that the unit delivers its rating.

Interpretation. Bayesian diagnosis does not only ask whether a reading crossed a threshold. It evaluates how strongly the observation supports competing explanations, weighted by their prior plausibility. The distinction is central in a design review because many shortfalls seen on a peak day are not uniquely associated with one physical cause: a high hall temperature can be caused by a cooling unit below its rating, a load above the committed load, a warmer supply air, an air path that starves part of the hall, or a biased sensor. Each of these is a hypothesis with its own prior and its own likelihood for the same reading, and the inference becomes a comparison among them, the diagnosis category of Chapter 2; only the first two are answered by buying capacity.

3.4 Marginalization and dependence

The probability of a flag in Section 3.3.2 was computed without knowing whether the unit was short. The same operation, for discrete and continuous unknowns and for nuisance quantities, is what lets a shared workload make two halls dependent, and what an independence assumption costs.

3.4.1 Discrete marginalization

Suppose a cooling unit (illustrative) can be in one of three conditions, at its rating, derated, or out of service, and let S denote the condition present. Table 3.5 gives, for each state, its probability and the probability that the shortfall flag H comes up in it; the last two columns are computed below.

Table 3.5: Three conditions of a cooling unit: probabilities, flag likelihoods, the joint probabilities of state and flag, and the probabilities of the states given a flag.

State s	$\Pr(S = s)$ [–]	$\Pr(H \mid S = s)$ [–]	$\Pr(H \cap S = s)$ [–]	$\Pr(S = s \mid H)$ [–]
at its rating	0.85	0.05	0.0425	0.272
derated	0.12	0.70	0.0840	0.538
out of service	0.03	0.99	0.0297	0.190
total	1.00		0.1562	1.000

A flag can arise under any of the three states. Because the states are mutually exclusive and exhaustive, the events $H \cap \{S = s\}$ are disjoint and their union is H, so $\Pr(H) = \sum_s \Pr(H \cap S = s)$. The multiplication rule writes each term as a likelihood times a prior:

$$\Pr(H) = \sum_s \Pr(H \mid S = s)\, \Pr(S = s). \tag{3.12}$$

This is the *law of total probability*. Substituting,

$$\Pr(H) = 0.05 \times 0.85 + 0.70 \times 0.12 + 0.99 \times 0.03 = 0.0425 + 0.0840 + 0.0297 = 0.1562.$$

The probability of a flag has been computed without knowing which state is present. That is what marginalization does: it accounts for the uncertainty in variables that are not part of the final question. The same sum is the denominator of Bayes' theorem over the three states, and dividing each joint probability by it gives the last column of the table: given a flag, the unit is derated with probability $0.0840/0.1562 = 0.538$, at its rating with 0.272 and out of service with 0.190.

3.4.2 Continuous marginalization

When the unknown is continuous, the sum becomes an integral. Let the rate λ at which the hall's load exceeds its committed load be uncertain and let $\mathbf{y}$ be the observed evidence; their joint distribution is $p(\lambda, \mathbf{y}) = p(\mathbf{y} \mid \lambda)\, p(\lambda)$. The probability density of the observations is obtained by integrating over every possible value of the rate,

$$p(\mathbf{y}) = \int p(\mathbf{y} \mid \lambda)\, p(\lambda)\, \mathrm{d}\lambda, \tag{3.13}$$

the *marginal likelihood* of the observations. Bayes' theorem for a continuous unknown then reads

$$p(\lambda \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \lambda)\, p(\lambda)}{\int p(\mathbf{y} \mid \lambda')\, p(\lambda')\, \mathrm{d}\lambda'}, \tag{3.14}$$

where λ' is a dummy variable of integration. The numerator evaluates the relative plausibility of one particular rate; the denominator ensures that the posterior integrates to one.

For the hall's exceedance log of Chapter 2, $k = 2$ exceedances of the committed load in $T_{\mathrm{tot}} = 1000\,\mathrm{h}$ of logged operation under the Poisson model $\Pr(k \mid \lambda) = (\lambda T_{\mathrm{tot}})^k \mathrm{e}^{-\lambda T_{\mathrm{tot}}}/k!$ and the prior $\mathrm{Gamma}(a, b) = \mathrm{Gamma}(2, 500\,\mathrm{h})$, with density $b^a \lambda^{a-1} \mathrm{e}^{-b\lambda}/\Gamma(a)$, the standard gamma integral $\int_0^\infty \lambda^{a-1} \mathrm{e}^{-s\lambda}\, \mathrm{d}\lambda = \Gamma(a)/s^a$ with $s = b + T_{\mathrm{tot}}$ evaluates equation (3.13) in closed form:

$$\Pr(k \mid T_{\mathrm{tot}}) = \frac{T_{\mathrm{tot}}^k}{k!} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+k)}{(b + T_{\mathrm{tot}})^{a+k}} = \frac{1000^2}{2} \cdot \frac{500^2}{1} \cdot \frac{3!}{1500^4} = 0.1481.$$

Before the log, the model gave a probability of 0.148 to exactly two exceedances in 1000 h; dividing the numerator of equation (3.14) by it gives the posterior $\mathrm{Gamma}(4, 1500\,\mathrm{h})$. The probability of a design day within the committed load, no exceedance in its $\tau = 12$ peak hours, is the same operation applied to the event of no exceedance: its conditional probability $\mathrm{e}^{-\lambda\tau}$ averaged over the posterior of the rate. Combining the two exponentials and using the same gamma integral with $s = b + \tau$, for the posterior $\mathrm{Gamma}(a, b)$,

$$\Pr(\text{no exceedance in } \tau \mid \mathbf{y}) = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^{a-1}\, \mathrm{e}^{-(b+\tau)\lambda}\, \mathrm{d}\lambda = \frac{b^a}{\Gamma(a)} \frac{\Gamma(a)}{(b+\tau)^a} = \left(\frac{b}{b+\tau}\right)^a, \tag{3.15}$$

which for $\mathrm{Gamma}(4, 1500\,\mathrm{h})$ is $(1500/1512)^4 = 0.96863$. Chapter 2 reports this probability from the posterior of the state, the same average over the normal of the log-rate by the five-node sum of equation (2.41): 0.96457, with the gamma integral 0.96863 as its exact value beside it, a gap of 0.0041. An exceedance in the design day is then 0.0354 (0.0314 exact). The plug-in at the posterior mean rate, $\mathrm{e}^{-12\times 4/1500} = 0.9685$, sits 0.0001 from the exact value and answers a different question, an exceedance whose rate is known. The predictive probability of equation (3.2) is the same operation applied to the next reading: the conditional probability averaged over the posterior of the hall air.

3.4.3 Nuisance quantities

Suppose the log-conductance g of the hall is being estimated from the hall reading and the load meter, but the model also contains the uncertain calibration offset b_Q of the load meter, in kW. The full posterior is $p(g, b_Q \mid \mathbf{y})$, and the planner wants the distribution of g alone. It is obtained by integrating b_Q out,

$$p(g \mid \mathbf{y}) = \int p(g, b_Q \mid \mathbf{y})\, \mathrm{d}b_Q. \tag{3.16}$$

The offset does not have to be treated as known; its uncertainty enters the answer through the integration. This matters most in physics-constrained inference, where many unknown state entries influence the quantity studied: the marginal of one entry z_i is the posterior of the whole state $\mathbf{z}$ integrated over all its other entries.

Marginalization against optimization. These two operations are different and must not be confused. Marginalization, equation (3.16), accumulates probability across all values of b_Q. Optimization over the nuisance quantity, $\max_{b_Q} p(g, b_Q \mid \mathbf{y})$, selects the best-fitting offset for each conductance. They generally give different results. A discrete illustration makes the difference plain. Let the conductance be either low or high and the offset take one of three values, with joint posterior probabilities 0.30, 0.05, 0.05 for a low conductance and 0.20, 0.20, 0.20 for a high one (illustrative). Marginalization gives $\Pr(\text{low}) = 0.40$ and $\Pr(\text{high}) = 0.60$; maximization over b_Q gives 0.30 against 0.20 and prefers the low conductance. The low conductance owes its single peak to one value of the offset; the high one is supported by every value. The distinction returns when later chapters compare a maximum a posteriori estimate with the marginals of the posterior.

3.4.4 Two halls driven by one site-wide workload

Many capacity calculations assume that the halls of a site peak independently. Let F_A and F_B be the events that hall A and hall B exceed their committed loads on the same day, the event of the log of Section 3.2.1. If each has probability 0.05, independence gives $\Pr(F_A \cap F_B) = 0.05^2 = 0.0025$ for the probability that both are above their committed loads at once. The conclusion is invalid if the halls share an important cause. Consider two halls that serve one site's workload: a batch campaign, a month-end close or a seasonal peak in demand raises the IT load in both at once. Let C be the event that the site-wide workload runs high during a day, with $\Pr(C) = 0.05$. The event C counts every day on which the site's workload runs high enough to push the halls toward their committed loads, not only the full site-wide peaks q_S of the two-halls example of Chapter 2, so it is more frequent than those; and it raises each hall's exceedance probability, to 0.60 and 0.50 below, rather than forcing both. Figure 3.4 draws the dependency: the box at the top is the shared condition, the two boxes below are the exceedance events of the halls, and each arrow says that the probability of the event at its head depends on the condition at its tail. There is no arrow between the halls.

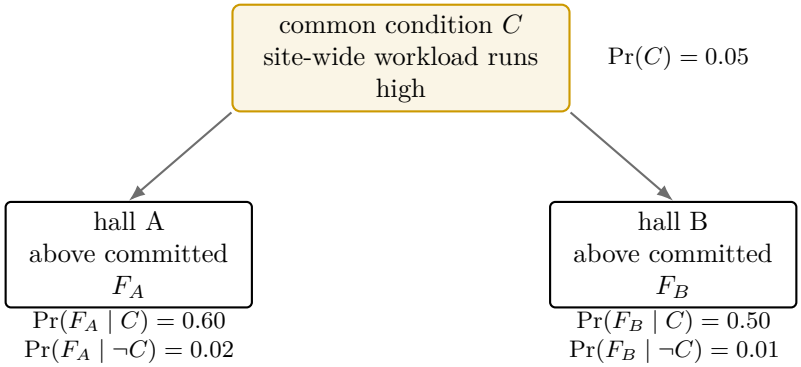


Figure 3.4: A probabilistic dependency graph for two halls driven by one site-wide workload. The shared condition C, a high site-wide workload and with it a high IT load in both halls, influences both exceedance events; the absence of an arrow between the halls states that the exceedances are conditionally independent given C. The probabilities are those of Table 3.6.

The exceedance probabilities of the halls depend on whether the site-wide workload runs high. Table 3.6 lists them: the rows are the two conditions and the columns the probability that each hall exceeds its committed load on a day under that condition.

Assume that the two exceedances are *conditionally independent given* C:

$$\Pr(F_A \cap F_B \mid C) = \Pr(F_A \mid C)\,\Pr(F_B \mid C), \qquad \Pr(F_A \cap F_B \mid \neg C) = \Pr(F_A \mid \neg C)\,\Pr(F_B \mid \neg C).$$

Table 3.6: Probabilities that each hall exceeds its committed load on a day, with and without a high site-wide workload (illustrative); $\Pr(C) = 0.05$.

Condition	$\Pr(F_A \mid \cdot)$ [–]	$\Pr(F_B \mid \cdot)$ [–]
high site-wide workload, C	0.60	0.50
ordinary workload, $\neg C$	0.02	0.01

Once the condition of the site-wide workload is known, an exceedance in one hall says nothing more about the other. This does not mean that the exceedances are unconditionally independent, as the next subsection shows.

3.4.5 The joint exceedance probability and the independence error

The law of total probability, equation (3.12), applied to the event $F_A \cap F_B$ with the two conditions C and $\neg C$ as the states, gives

$$\Pr(F_A \cap F_B) = \Pr(F_A \cap F_B \mid C)\,\Pr(C) + \Pr(F_A \cap F_B \mid \neg C)\,\Pr(\neg C).$$

Under conditional independence each conditional joint probability is a product,

$$\Pr(F_A \cap F_B) = 0.60 \times 0.50 \times 0.05 + 0.02 \times 0.01 \times 0.95 = 0.015 + 0.00019,$$

so that

$$\Pr(F_A \cap F_B) = 0.01519. \tag{3.17}$$

The probability that both halls exceed their committed loads on the same day is about 1.52 %, and almost all of it, 0.015 of 0.01519, comes from the shared workload.

The marginal exceedance probabilities follow from the same law,

$$\Pr(F_A) = 0.60 \times 0.05 + 0.02 \times 0.95 = 0.049, \qquad \Pr(F_B) = 0.50 \times 0.05 + 0.01 \times 0.95 = 0.0345.$$

If the exceedances are incorrectly treated as independent, the joint probability is the product of the marginals,

$$\Pr(F_A \cap F_B) \overset{\text{incorrect}}{=} 0.049 \times 0.0345 = 0.00169,$$

or 0.169 %. The correct value, 1.519 %, is 8.99 times larger: ignoring the shared workload underestimates the probability that both halls are above their committed loads at once by a factor of about nine in this illustrative site, and the independent product is only 11 % of the correct value. With marginals of the same size, 0.049 and 0.0345 here against 0.05 and 0.03 there, this factor of nine sits beside the factor of 7.2 that Chapter 2 finds for two halls both exceeding at a site-wide peak probability $q_S = 0.01$.

This joint probability is what lies beneath the coincidence (diversity) factor that Chapter 8 computes for the site's halls: the ratio of the peak the shared plant sees to the sum of the halls' own peaks. Treating the halls as independent makes their coincident peak look rarer than it is and the coincidence factor smaller than it is, and a shared plant sized on that factor falls short of the load it must carry on the day both halls peak.

Common mistake

Multiplying the exceedance probabilities of two halls to obtain the probability that both are above their committed loads at once. The product is the joint probability only when the exceedances are independent; a shared workload, a shared outdoor condition or a shared control sequence makes them dependent, and the day on which both halls draw more than their committed loads from the shared plant is much more likely than the product suggests.

Principle 3.2. *Understanding what the units and halls of a plant share can matter more than estimating the peak of each one with greater precision.*

3.4.6 Conditional independence given the workload

Two random variables X_1 and X_2 are conditionally independent given C when, for every value c of C,

$$p(x_1, x_2 \mid c) = p(x_1 \mid c)\, p(x_2 \mid c), \tag{3.18}$$

written $X_1 \perp X_2 \mid C$. Dividing both sides by $p(x_1 \mid c)$ and using the definition of conditional probability on the left gives the equivalent statement $p(x_2 \mid x_1, c) = p(x_2 \mid c)$: after conditioning on C, learning X_1 gives no additional information about X_2 under the assumed model. The two halls in physical form are the case: with the site-wide workload W split between them, the hall loads $Q_A = s_A W + u_A$ and $Q_B = s_B W + u_B$, each hall's share of the workload plus a variation u of its own, depend on separate variations, so that given W they are independent; with W unknown, a load reading in hall A informs W and through it hall B.

When C is unknown, observing X_1 informs C, and that changes what is believed about X_2. If hall A exceeds its committed load, a high site-wide workload becomes suspect: by Bayes' theorem,

$$\Pr(C \mid F_A) = \frac{\Pr(F_A \mid C)\ \Pr(C)}{\Pr(F_A)} = \frac{0.60 \times 0.05}{0.049} = 0.612,$$

up from 0.05. The probability that hall B also exceeds its committed load is then the law of total probability conditioned on F_A, with conditional independence used to drop F_A from $\Pr(F_B \mid C, F_A)$:

$$\begin{aligned}\Pr(F_B \mid F_A) &= \Pr(F_B \mid C)\ \Pr(C \mid F_A) + \Pr(F_B \mid \neg C)\ \Pr(\neg C \mid F_A) \\ &= 0.50 \times 0.612 + 0.01 \times 0.388 = 0.310,\end{aligned}$$

which equals $\Pr(F_A \cap F_B)/\Pr(F_A) = 0.01519/0.049$, against the unconditional $\Pr(F_B) = 0.0345$. An exceedance in one hall raises the exceedance probability of the other ninefold, through the workload they share: the planner who sees hall A above its committed load on a day has reason to expect hall B there on the same day. The shortfall flag of the opening scene watches one unit's capacity and does not raise C, so it gives no such reason unless the site's workload itself is suspect. Conditional independence is therefore not unconditional independence: the variables are independent only once the common cause is known.

3.5 Conditional independence and a continuous update in a physical model

The same rules apply when the unknowns are physical quantities tied by a conservation relation. On the equipment hall they give the posterior of a state vector under Gaussian rows, the probability that the hall exceeds its limit after one reading, and, in the two-entry hall of Chapter 1, a reading that reaches the conductance through the physics row.

3.5.1 The hall as a probability model

The hall of the earlier chapters satisfies the steady heat balance

$$T = T_c + \frac{Q}{G}, \tag{3.19}$$

with the hall air T and the supply air T_c in °C, the heat load Q in kW and the conductance G of the air path in kW/K. A sensor reads the hall air with an additive error,

$$y_T = T + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_T^2),$$

where ϵ is the measurement error and σ_T its standard deviation in K. Figure 3.5 draws the system twice. Panel (a) is the plant graph: the source supplies the load Q to the hall node, which carries the free entry T; the green path to the supply-air node carries the conductance, held in the state as the log-conductance $g = \ln(G/G_{\text{ref}})$; the supply-air node is pinned at T_c by the unit's discharge controller; and the dashed arrow is the reading y_T. Panel (b) is the dependency graph of the probability model of this section, in which Q, G and T_c are all treated as uncertain: the load, the conductance and the supply air determine the modeled hall air, and the reading depends on that temperature alone.

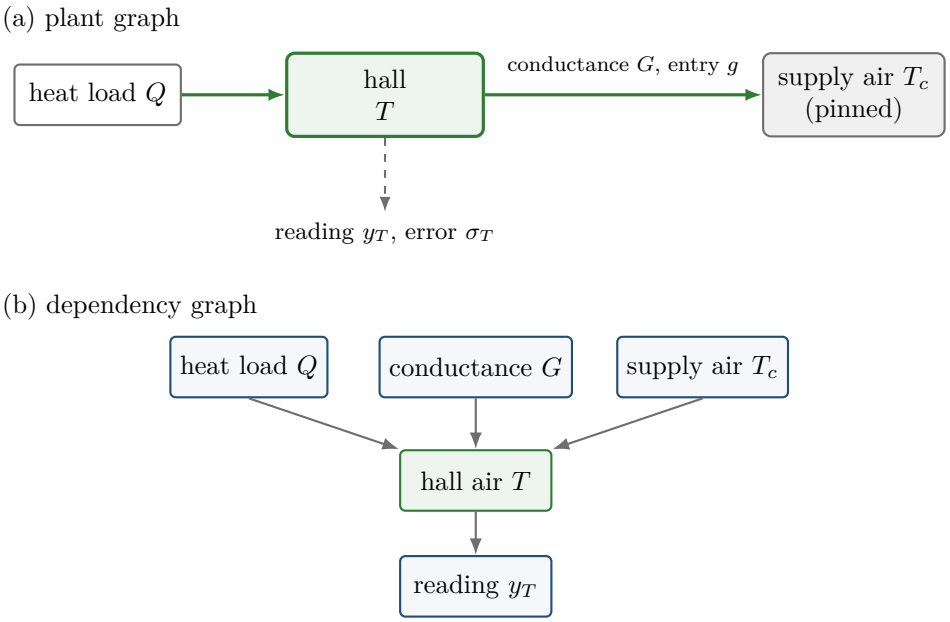


Figure 3.5: The hall in two drawings. (a) The plant graph: free entry T at the hall node and $g = \ln(G/G_{\text{ref}})$ on the air path; supply air T_c pinned by the discharge controller; source Q; reading y_T observed (dashed). (b) The dependency graph of the factorized model of equation (3.21), in which Q, G and T_c carry priors: arrows point from a quantity to what depends on it, and the reading depends on the inputs only through T.

If the measurement error is independent of the physical inputs, the reading is conditionally independent of those inputs once the true temperature is known:

$$p(y_T \mid T, G, Q, T_c) = p(y_T \mid T). \tag{3.20}$$

This does not make the reading uninformative about the conductance. The reading informs T, and the physical relation (3.19) connects T to G: evidence propagates from the sensor reading to the temperature and from the temperature to the conductance. An observation of one quantity updates the beliefs about every quantity connected to it through shared physical constraints, and this is one of the main reasons why inference over a graph is useful.

The factorized probability model. Suppose G, Q and T_c are uncertain inputs with a joint prior $p(G, Q, T_c)$, which relaxes the pins of the plant graph and carries Q and T_c as free entries. The chain rule, the multiplication rule applied twice, and equation (3.20) give the joint model

$$p(G, Q, T_c, T, y_T) = p(G, Q, T_c)\, p(T \mid G, Q, T_c)\, p(y_T \mid T). \tag{3.21}$$

The physics factor is written as a Gaussian around the heat balance, $p(T \mid G, Q, T_c) = \mathcal{N}(T;\ T_c + Q/G,\ \sigma_p^2)$, with a model tolerance σ_p in K: a physical discrepancy is a finite σ_p, and the exact deterministic balance of equation (3.19) is the limit $\sigma_p \to 0$, in which the factor becomes the Dirac delta $\delta(T - T_c - Q/G)$. The joint distribution is the complete probability model, and the posterior wanted is $p(G, Q, T_c, T \mid y_T)$, over the whole state with T kept in it, which answers questions about the conductance, the temperature, the design margins and every other quantity of the model. Each of its three factors is a row over that state: the prior, the physics row with tolerance σ_p and the reading row; the derivations of this section keep all of them as rows and never remove an entry by substitution. In the limit $\sigma_p \to 0$, integrating T out against the delta gives the same posterior of the inputs as $p(G, Q, T_c \mid y_T) \propto p(G, Q, T_c)\, p(y_T \mid T = T_c + Q/G)$: every combination of inputs is weighted by how well the temperature it implies explains the reading.

3.5.2 Prior and reading for the hall air

The simplest continuous case has one unknown, the true hall air temperature T, so that the state is this one number,

$$T \in \mathbb{R}.$$

The general forms of the book write the state as a vector $\mathbf{z}$; here it has the single entry T, so every quantity of the fit below is a number: each row's residual and its slope in T, the objective's slope and curvature, the information and the posterior variance. On an illustrative hall running warmer than the series' hall, a design model gives an uncertain prior estimate,

$$T \sim \mathcal{N}(\mu_0, \sigma_0^2) = \mathcal{N}(26, 2^2), \qquad \mu_0 = 26^\circ\mathrm{C}, \quad \sigma_0 = 2\,\mathrm{K}.$$

A sensor reports $y_T = 27^\circ$C, and its error is Gaussian with $\sigma_T = 0.5\,$K, so that $y_T \mid T \sim \mathcal{N}(T, 0.5^2)$. The question is the posterior distribution of the true temperature after the reading. Bayes' theorem for a continuous unknown, with the evidence dropped because it does not depend on the state, gives

$$p(T \mid y_T) \propto p(y_T \mid T)\, p(T),$$

with the Gaussian prior and the Gaussian likelihood

$$p(T) \propto \exp\left[-\frac{(T-\mu_0)^2}{2\sigma_0^2}\right], \qquad p(y_T \mid T) \propto \exp\left[-\frac{(y_T - T)^2}{2\sigma_T^2}\right].$$

Multiplying,

$$p(T \mid y_T) \propto \exp\left[-\frac{1}{2}\left(\frac{(T-\mu_0)^2}{\sigma_0^2} + \frac{(y_T - T)^2}{\sigma_T^2}\right)\right]. \tag{3.22}$$

Each term of the exponent is the square of a difference divided by its standard deviation. The two differences are the residuals of the problem, one prior row and one reading row, each already divided by its standard deviation,

$$r_1(T) = \frac{T - \mu_0}{\sigma_0}, \qquad r_2(T) = \frac{T - y_T}{\sigma_T}, \tag{3.23}$$

and equation (3.22) reads $p(T \mid y_T) \propto \exp[-\frac{1}{2}(r_1^2 + r_2^2)] = \mathrm{e}^{-\mathcal{L}(T)}$, with the objective $\mathcal{L}(T) = \frac{1}{2}(r_1^2 + r_2^2)$ of the notation chapter. Both rows are linear in T, each with a constant slope $j_k = \mathrm{d}r_k/\mathrm{d}T$: $j_1 = 1/\sigma_0$ and $j_2 = 1/\sigma_T$.

Gaussian rows over any state. The same step holds for a state of any size and any number of rows. The joint model factorizes by the chain rule, as equation (3.21) does for the hall: a prior factor over the inputs, a physics factor for each physical relation, and a reading factor for each reading, which by conditional independence, equation (3.20), depends on the state only through the quantity read. Row k is one such factor written as a residual: the difference between the quantity computed from the state and what was observed, or what the prior or the physics expects, divided by its standard deviation σ_k. A Gaussian factor is $\exp[-\frac{1}{2}r_k(\mathbf{z})^2]$ times a normalizer, $1/(\sqrt{2\pi}\,\sigma_k)$, that is free of the state. Multiplying the factors and dropping the normalizers and the evidence, Bayes' theorem gives

$$p(\mathbf{z} \mid \mathbf{y}) \propto \prod_k \exp\Big[-\tfrac{1}{2}r_k(\mathbf{z})^2\Big] = \exp\Big[-\tfrac{1}{2}r(\mathbf{z})^{\mathsf{T}}r(\mathbf{z})\Big] = \mathrm{e}^{-\mathcal{L}(\mathbf{z})}, \tag{3.24}$$

where $r(\mathbf{z})$ stacks the reading rows, the prior rows and the physics rows with their tolerances σ_p. For the one-entry hall the product has the two factors of equation (3.22) and the sum of squares is the number $r_1^2 + r_2^2$; the vector is needed when the state has several entries, as in Section 3.5.6 and in the hall of Chapter 4. A Gaussian prior over several entries, $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with correlated entries, enters as the rows $\mathbf{L}_0(\mathbf{z} - \boldsymbol{\mu}_0)$ with $\mathbf{L}_0^{\mathsf{T}}\mathbf{L}_0 = \boldsymbol{\Sigma}_0^{-1}$, since then $(\mathbf{z} - \boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}(\mathbf{z} - \boldsymbol{\mu}_0)$ is the sum of their squares.

3.5.3 Completion of the square in the state

Suppose the rows of a state with several entries are linear in it, as both rows of equation (3.23) are in T. Then, for any point $\mathbf{z}_0$,

$$r(\mathbf{z}) = r(\mathbf{z}_0) + \mathbf{J}\,(\mathbf{z} - \mathbf{z}_0)$$

holds exactly, with the Jacobian $\mathbf{J} = \partial r/\partial \mathbf{z}$, a matrix of m rows and n columns for m rows and n entries. Write $\boldsymbol{\Delta} = \mathbf{z} - \mathbf{z}_0$ and $r_0 = r(\mathbf{z}_0)$. The objective expands as

$$\mathcal{L}(\mathbf{z}) = \tfrac{1}{2}(r_0 + \mathbf{J}\boldsymbol{\Delta})^{\mathsf{T}}(r_0 + \mathbf{J}\boldsymbol{\Delta}) = \tfrac{1}{2}r_0^{\mathsf{T}}r_0 + \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{J}^{\mathsf{T}}r_0 + \tfrac{1}{2}\boldsymbol{\Delta}^{\mathsf{T}}\mathbf{J}^{\mathsf{T}}\mathbf{J}\,\boldsymbol{\Delta}.$$

Its gradient, $\mathbf{J}^{\mathsf{T}}r_0 + \mathbf{J}^{\mathsf{T}}\mathbf{J}\,\boldsymbol{\Delta}$, vanishes at the step $\hat{\boldsymbol{\Delta}}$ that solves the *normal equations*,

$$(\mathbf{J}^{\mathsf{T}}\mathbf{J})\,\hat{\boldsymbol{\Delta}} = -\mathbf{J}^{\mathsf{T}}r(\mathbf{z}_0), \qquad \hat{\mathbf{z}} = \mathbf{z}_0 + \hat{\boldsymbol{\Delta}}. \tag{3.25}$$

Writing $\mathbf{J}^{\mathsf{T}}r_0 = -\mathbf{J}^{\mathsf{T}}\mathbf{J}\,\hat{\boldsymbol{\Delta}}$ in the expansion completes the square,

$$\mathcal{L}(\mathbf{z}) = \tfrac{1}{2}(\boldsymbol{\Delta} - \hat{\boldsymbol{\Delta}})^{\mathsf{T}}\mathbf{J}^{\mathsf{T}}\mathbf{J}\,(\boldsymbol{\Delta} - \hat{\boldsymbol{\Delta}}) + \text{constant} = \tfrac{1}{2}(\mathbf{z} - \hat{\mathbf{z}})^{\mathsf{T}}\mathbf{J}^{\mathsf{T}}\mathbf{J}\,(\mathbf{z} - \hat{\mathbf{z}}) + \text{constant},$$

where the constant collects the terms free of $\mathbf{z}$. By equation (3.24) the posterior is the exponential of minus this quadratic form, the density of a multivariate Gaussian with mean $\hat{\mathbf{z}}$ and covariance

$$\boldsymbol{\Sigma} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}. \tag{3.26}$$

The matrix $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ is the information matrix $\boldsymbol{\Lambda}$ of the normalized rows. When the rows are written unnormalized, each in its own unit with the weight $w_k = 1/\sigma_k^2$ and the weights collected in $\mathbf{W} = \mathrm{diag}(w_k)$, the same matrix is $\boldsymbol{\Lambda} = \mathbf{J}^{\mathsf{T}}\mathbf{W}\mathbf{J}$, with $\mathbf{J}$ there the Jacobian of the unnormalized rows. For rows linear in the state the normal equations reach $\hat{\mathbf{z}}$ in one step from any $\mathbf{z}_0$, and equations (3.25) and (3.26) are the exact posterior mean and covariance. For nonlinear rows the same step is the Gauss–Newton step, repeated from the new point until it vanishes, and the result is a local Gaussian approximation, the Laplace approximation, whose derivation and curvature term Chapter 4 gives, with the three-entry hall as its example; Section 3.5.6 reads the two-entry hall of Chapter 1 in the terms of this section.

Precisions add. Split the rows into prior rows $\mathbf{L}_0(\mathbf{z}-\boldsymbol{\mu}_0)$, whose Jacobian $\mathbf{L}_0$ satisfies $\mathbf{L}_0^{\mathsf{T}}\mathbf{L}_0=\boldsymbol{\Sigma}_0^{-1}$, and reading rows $\mathbf{R}^{-1/2}(\mathbf{A}\mathbf{z}-\mathbf{y})$, where the matrix $\mathbf{A}$ maps the state to the quantities read and the diagonal matrix $\mathbf{R}$ holds the variances of the readings. The Jacobian of the reading rows is $\mathbf{J}_y=\mathbf{R}^{-1/2}\mathbf{A}$; stacking the two blocks of rows stacks their Jacobians, and $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ is the sum of the two blocks' products,

$$\boldsymbol{\Sigma}^{-1}=\boldsymbol{\Sigma}_0^{-1}+\mathbf{J}_y^{\mathsf{T}}\mathbf{J}_y=\boldsymbol{\Sigma}_0^{-1}+\mathbf{A}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{A}. \tag{3.27}$$

Starting the normal equations at $\mathbf{z}_0=\boldsymbol{\mu}_0$, where the prior rows vanish, gives $\mathbf{J}^{\mathsf{T}}r(\boldsymbol{\mu}_0)=\mathbf{A}^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{A}\boldsymbol{\mu}_0-\mathbf{y})$, and equation (3.25) gives

$$\hat{\mathbf{z}}=\boldsymbol{\mu}_0+\boldsymbol{\Sigma}\,\mathbf{A}^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}_0). \tag{3.28}$$

The posterior information is the prior information plus the information of the readings, and the posterior mean is moved from the prior mean toward the readings in proportion to their information.

The conditional of a joint Gaussian. The same posterior follows from the joint distribution of the state and the readings. Before the readings, $\mathbf{z}\sim\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$ and $\mathbf{y}=\mathbf{A}\mathbf{z}+\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{R})$ independent of the state, so the stacked vector of $\mathbf{z}$ and $\mathbf{y}$ is Gaussian with means $\boldsymbol{\mu}_0$ and $\mathbf{A}\boldsymbol{\mu}_0$ and covariance blocks

$$\boldsymbol{\Sigma}_{zz}=\boldsymbol{\Sigma}_0,\qquad \boldsymbol{\Sigma}_{zy}=\boldsymbol{\Sigma}_0\mathbf{A}^{\mathsf{T}}=\boldsymbol{\Sigma}_{yz}^{\mathsf{T}},\qquad \boldsymbol{\Sigma}_{yy}=\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}^{\mathsf{T}}+\mathbf{R}.$$

Let $\mathbf{P}$ be the inverse of the joint covariance and $\mathbf{S}=\boldsymbol{\Sigma}_{zz}-\boldsymbol{\Sigma}_{zy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yz}$ the *Schur complement* of $\boldsymbol{\Sigma}_{yy}$. The block-inversion formula, checked by multiplying back, gives $\mathbf{P}_{zz}=\mathbf{S}^{-1}$ and $\mathbf{P}_{zy}=-\mathbf{S}^{-1}\boldsymbol{\Sigma}_{zy}\boldsymbol{\Sigma}_{yy}^{-1}$. With the readings fixed, the exponent of the joint density, times -1, is a quadratic in the state,

$$\tfrac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_0)^{\mathsf{T}}\mathbf{P}_{zz}(\mathbf{z}-\boldsymbol{\mu}_0)+(\mathbf{z}-\boldsymbol{\mu}_0)^{\mathsf{T}}\mathbf{P}_{zy}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}_0)+\text{constant}=\tfrac{1}{2}(\mathbf{z}-\mathbf{m})^{\mathsf{T}}\mathbf{S}^{-1}(\mathbf{z}-\mathbf{m})+\text{constant},$$

with $\mathbf{m}=\boldsymbol{\mu}_0+\boldsymbol{\Sigma}_{zy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}_0)$. The conditional density is the joint density divided by $p(\mathbf{y})$, which is free of the state, so

$$\mathbf{z}\mid\mathbf{y}\sim\mathcal{N}\big(\boldsymbol{\mu}_0+\boldsymbol{\Upsilon}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}_0),\ \boldsymbol{\Sigma}_0-\boldsymbol{\Upsilon}\mathbf{A}\boldsymbol{\Sigma}_0\big),\qquad \boldsymbol{\Upsilon}=\boldsymbol{\Sigma}_0\mathbf{A}^{\mathsf{T}}(\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}^{\mathsf{T}}+\mathbf{R})^{-1}, \tag{3.29}$$

since $\boldsymbol{\Sigma}_{zy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yz}=\boldsymbol{\Upsilon}\mathbf{A}\boldsymbol{\Sigma}_0$. This is the correction form: the prior mean is moved toward the readings by the gain $\boldsymbol{\Upsilon}$ times the innovation $\mathbf{y}-\mathbf{A}\boldsymbol{\mu}_0$, and the covariance is the Schur complement. The Kalman literature writes the gain $\mathbf{K}$; here it is $\boldsymbol{\Upsilon}$, clear of the hat matrix $\mathbf{H}$ and its entries h_{kk} that the series uses for its misfit tests. The correction form gives the same posterior as equations (3.27) and (3.28), by the matrix inversion identity $(\boldsymbol{\Sigma}_0^{-1}+\mathbf{A}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{A})^{-1}=\boldsymbol{\Sigma}_0-\boldsymbol{\Upsilon}\mathbf{A}\boldsymbol{\Sigma}_0$; the information form inverts a matrix of the size of the state, the correction form one of the size of the readings.

Remark 3.1. For one entry and one reading, $\mathbf{A}=1$, $\boldsymbol{\Sigma}_0=\sigma_0^2$ and $\mathbf{R}=\sigma_T^2$, and the three equations reduce to the scalar rules $1/\sigma_{\text{post}}^2=1/\sigma_0^2+1/\sigma_T^2$, $\mu_{\text{post}}=\sigma_{\text{post}}^2(\mu_0/\sigma_0^2+y_T/\sigma_T^2)$ and $\mu_{\text{post}}=\mu_0+\upsilon(y_T-\mu_0)$ with $\upsilon=\sigma_0^2/(\sigma_0^2+\sigma_T^2)$: precisions add, and the posterior mean is the precision-weighted average of the prior mean and the reading.

3.5.4 The numbers and what changed

For the hall the state is the one number T, and the two rows of equation (3.23) have the slopes

$$j_1=\frac{1}{\sigma_0}=0.5\,\mathrm{K}^{-1},\qquad j_2=\frac{1}{\sigma_T}=2\,\mathrm{K}^{-1}.$$

The row terms are $\mathcal{L}_k = \frac{1}{2}r_k^2$, with the slopes $\mathcal{L}'_k = j_k r_k$ and the curvatures $\mathcal{L}''_k = j_k^2$ in T, the primes denoting $\mathrm{d}/\mathrm{d}T$; slopes and curvatures add over the rows,

$$\mathcal{L}' = j_1 r_1 + j_2 r_2, \qquad \mathcal{L}'' = j_1^2 + j_2^2.$$

Because the rows are linear, the curvature has no further term and is the same at every T; it is the information

$$\Lambda = j_1^2 + j_2^2 = 0.5^2 + 2^2 = 0.25 + 4 = 4.25\,\mathrm{K}^{-2},$$

the prior precision $1/\sigma_0^2$ plus the reading precision $1/\sigma_T^2$, as equation (3.27) states in the scalar form of the remark above, with the prior variance $\sigma_0^2 = 4\,\mathrm{K}^2$ and the reading variance $\sigma_T^2 = 0.25\,\mathrm{K}^2$. It is the one-entry case of $\mathbf{J}^\mathsf{T}\mathbf{J}$, and equation (3.26) gives the posterior variance as its reciprocal,

$$\sigma_{\text{post}}^2 = \frac{1}{\Lambda} = \frac{1}{4.25} = 0.2353\,\mathrm{K}^2, \qquad \sigma_{\text{post}} = 0.4851\,\mathrm{K}.$$

With one entry the normal equations (3.25) are the one equation $\Lambda\,\Delta T = -\mathcal{L}'$, the Newton step $\Delta T = -\mathcal{L}'/\mathcal{L}''$. They start at the prior mean, $T_0 = 26°\mathrm{C}$, where the prior row is 0 and the reading row $(26 - 27)/0.5 = -2$, so that $\mathcal{L}' = 0.5 \times 0 + 2 \times (-2) = -4\,\mathrm{K}^{-1}$, and the step is $\Delta T = 4/4.25 = 0.9412\,\mathrm{K}$. The iterations are

k [–]	T [°C]	$\mathcal{L}(T)$ [–]	$\mathcal{L}'_1$ [K^{-1}]	$\mathcal{L}'_2$ [K^{-1}]	$\mathcal{L}'$ [K^{-1}]	$\mathcal{L}''$ [K^{-2}]	ΔT [K]
0	26	2.0000	0	−4	−4	4.25	0.9412
1	26.9412	0.1176	0.2353	−0.2353	0	4.25	0

At the second point the residuals are $r_1 = 0.4706$ and $r_2 = -0.1176$, the slope is $\mathcal{L}' = 0.5 \times 0.4706 + 2 \times (-0.1176) = 0$, and the step vanishes: the rows are linear, and one step reaches the minimum. The correction form (3.29) confirms both moments: the variance of the reading before it arrives is $\sigma_0^2 + \sigma_T^2 = 4 + 0.25 = 4.25\,\mathrm{K}^2$, $\upsilon = 4/4.25 = 0.9412$, the mean $26 + 0.9412 \times (27 - 26) = 26.9412°\mathrm{C}$ and the Schur complement $4 - 0.9412 \times 4 = 0.2353\,\mathrm{K}^2$. The answer is the minimum $\hat{T} = 26.9412°\mathrm{C}$, written μ_{post} below, with the standard deviation σ_{post}:

$$T \mid y_T \sim \mathcal{N}(26.9412, 0.4851^2). \tag{3.30}$$

Figure 3.6(a) draws the update. The horizontal axis is the temperature and the vertical axis the probability density. The gray curve is the prior, the dashed blue curve the likelihood of the reading as a function of T (normalized to integrate to one), and the green curve the posterior. The posterior mean moved from 26 to 26.94°C, and the standard deviation fell from 2 to 0.49 K. The reading was more precise than the prior, so it received the greater weight, $\upsilon = 0.94$; but the posterior mean is not exactly 27°C, because the prior also carries information.

Table 3.7 shows the trend for sensors of four precisions, all reading 27°C against the same prior. The first column is the sensor's standard deviation, the second the weight υ of the reading, and the remaining columns the posterior mean, the posterior standard deviation and the exceedance probability derived in Section 3.5.5. Each row is the same computation with the slopes $j_1 = 1/\sigma_0$ and $j_2 = 1/\sigma_T$, so that $\sigma_{\text{post}}^2 = 1/\Lambda = 1/(0.25 + 1/\sigma_T^2)$. The second row is the case worked above. A sensor as uncertain as the prior, $\sigma_T = 2\,\mathrm{K}$, receives half the weight and leaves the posterior standard deviation at $2/\sqrt{2} = 1.41\,\mathrm{K}$.

3.5.5 The threshold question

Suppose the hall-air limit of this illustration is $T_{\max} = 28°\mathrm{C}$, and the question is the probability that the hall exceeds it, $\Pr(T > T_{\max} \mid y_T)$. Since the posterior is Gaussian, standardizing with

Table 3.7: The hall-air update for four sensor precisions: prior $\mathcal{N}(26, 2^2)$, reading $y_T = 27°\mathrm{C}$, equations (3.27) and (3.29), and the exceedance probability of $T_{\max} = 28°\mathrm{C}$ derived in Section 3.5.5.

σ_T [K]	weight υ [–]	μ_{post} [°C]	σ_{post} [K]	$\Pr(T > T_{\max} \mid y_T)$ [–]
0.25	0.9846	26.9846	0.2481	2×10^{-5}
0.5	0.9412	26.9412	0.4851	0.0145
1.0	0.8000	26.8000	0.8944	0.0899
2.0	0.5000	26.5000	1.4142	0.1444

the mean $\mu_{\text{post}} = \hat{T}$ and the standard deviation $\sigma_{\text{post}} = 1/\sqrt{\Lambda}$ and using the standard normal cumulative distribution function Φ,

$$\begin{aligned} \Pr(T > T_{\max} \mid y_T) &= 1 - \Phi\left(\frac{T_{\max} - \mu_{\text{post}}}{\sigma_{\text{post}}}\right) \\ &= 1 - \Phi\left(\frac{28 - 26.9412}{0.4851}\right) = 1 - \Phi(2.183) = 0.0145, \end{aligned} \tag{3.31}$$

or about 1.45 %. Before the reading the same calculation with the prior gives $\Pr(T > 28°\mathrm{C}) = 1 - \Phi((28-26)/2) = 1 - \Phi(1) = 0.1587$. The estimated exceedance probability fell from 15.9 % to 1.45 %, for the same fixed operating condition and the assumed Gaussian measurement model. The rows are linear, so this normal tail is the exact posterior's: a direct quadrature of the product (3.22) also gives 0.0145, and there is no gap to set beside it, as there is on the curved physics row of the next subsection. Figure 3.6(b) shades the two tails: the axes are those of panel (a), the dashed line is the limit, the gray area to its right under the prior is 0.159, and the green area under the posterior is 0.0145.

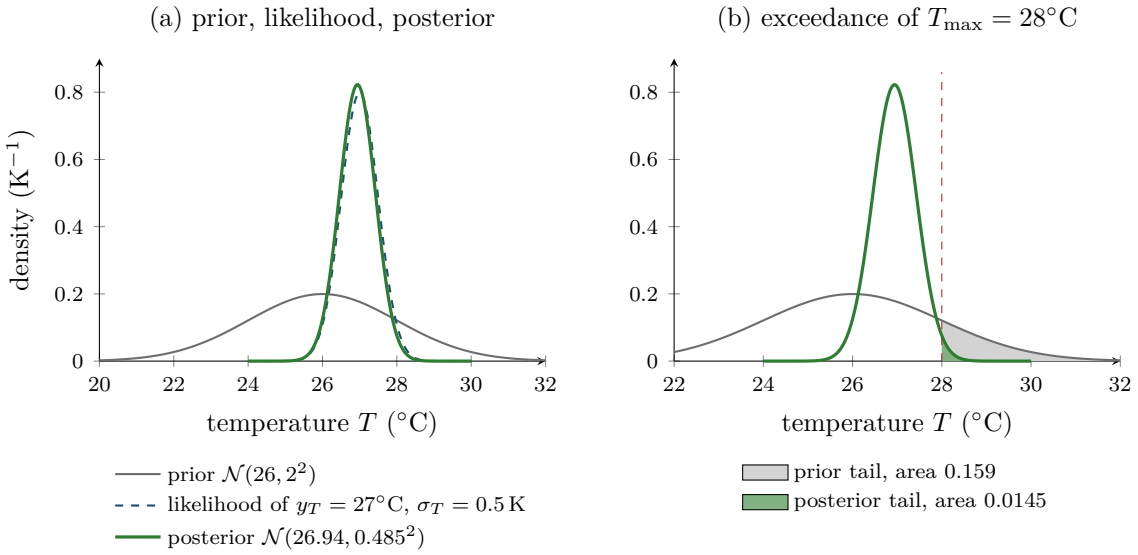


Figure 3.6: The hall-air update of equations (3.27)–(3.30). (a) Prior $\mathcal{N}(26, 2^2)$, likelihood of the reading $y_T = 27°\mathrm{C}$ with $\sigma_T = 0.5\,\mathrm{K}$ drawn as the density $\mathcal{N}(27, 0.5^2)$ in T, and posterior $\mathcal{N}(26.9412, 0.4851^2)$. (b) The probability that the hall exceeds $T_{\max} = 28°\mathrm{C}$ (dashed): $1 - \Phi(1) = 0.159$ under the prior and $1 - \Phi(2.183) = 0.0145$ under the posterior (equation (3.31)).

The physical hall did not necessarily change when the reading arrived; the information about it changed, and that is the mechanism of Bayesian updating. The calculation also walks the

method of *How to read this book* on its smallest example. The plant graph of Figure 3.5(a) is step 0. The question $\Pr(T > T_{\max} \mid y_T)$, unknown on the left and reading on the right, is step 1. Attaching the prior and the reading as the two rows of equation (3.23) is step 2. The posterior, derived through the normal equations and the variance as the reciprocal of the information, equations (3.25) and (3.26) with one entry, and computed by the one Newton step $-\mathcal{L}'/\mathcal{L}''$ with $\Lambda = 4.25\,\mathrm{K}^{-2}$, is step 3. Reading the normal tail from it is step 4, and checking that tail by the correction form and its Schur complement is step 5, and the sentence "the hall exceeds 28°C with probability 1.45 %, posterior 26.94 ± 0.49 K" is step 6. Step 7, the next measurement, reads Table 3.7: if 1.45 % were too high a residual doubt to commit a design on, a sensor with $\sigma_T = 0.25$ K would, for a reading near the same value, reduce it to the order of 10^{-5}; Chapter 18 ranks such measurements before commitment by the width each would remove. The posterior is inferred first, and the planner's question is then answered from it:

$$\underbrace{p(T \mid y_T)}_{\text{inference}} \quad \longrightarrow \quad \underbrace{\Pr(T > T_{\max} \mid y_T)}_{\text{threshold question}}.$$

3.5.6 The two-entry hall and its physics row

The reading of the one-entry hall informed only the entry it measured; Section 3.5.1 claimed that a reading also reaches the conductance through the physics row. The hall of Chapter 1 shows it. Its state is $\mathbf{z} = (T,\ g)^{\mathsf{T}}$, and its three rows, equation (1.10), are the physics row with tolerance $\sigma_p = 0.01$ K, the reading row with $\sigma_T = 0.5$ K and the handbook prior row on g with standard deviation 0.3; Chapter 1 solves their normal equations by Gauss–Newton from $(23°\mathrm{C},\ 0)$ (Table 1.4) and reaches $\hat{\mathbf{z}} = (23.070°\mathrm{C},\ -0.0139)^{\mathsf{T}}$. Equation (3.27) holds row by row: with $\mathbf{J}_k$ the k-th row of $\mathbf{J}$, $\mathbf{J}^{\mathsf{T}}\mathbf{J} = \sum_k \mathbf{J}_k^{\mathsf{T}}\mathbf{J}_k$, each row contributing the outer product of its own sensitivities. At $\hat{\mathbf{z}}$ the rows of $\mathbf{J}$ are $(100,\ 507.0)$, $(2,\ 0)$ and $(0,\ 3.333)$, and

$$\mathbf{J}^{\mathsf{T}}\mathbf{J} = \underbrace{\begin{bmatrix} 10\,000 & 50\,701 \\ 50\,701 & 257\,056 \end{bmatrix}}_{\text{physics}} + \underbrace{\begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}}_{\text{reading}} + \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & 11.11 \end{bmatrix}}_{\text{prior}} = \begin{bmatrix} 10\,004 & 50\,701 \\ 50\,701 & 257\,067 \end{bmatrix}.$$

The reading's and the prior's terms are diagonal, because each of those rows carries one entry. The only off-diagonal entry, 50 701, comes from the physics row, the one row with sensitivities in both columns: this is equation (3.20) in vector form, the reading informing T and the physics row carrying that information to g. Without the reading row, the physics and prior terms alone invert to $\sqrt{\Sigma_{22}} = 0.3$, the prior spread; with it, the covariance of Chapter 1, equation (1.21), has $\sqrt{\Sigma_{11}} = 0.481$ K, $\sqrt{\Sigma_{22}} = 0.0948$ and correlation -0.9998. The reading has cut the conductance spread from 0.3 to 0.0948 although its own row has a zero in the column of g, and the posterior lies along the physics curve, on the short arc of the state plane of Figure 1.5 inside the reading band. Because the physics row is nonlinear, these spreads include the curvature term that Chapter 4 derives; $(\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}$ alone gives 0.475 K and 0.094. The exact posterior of these rows has the standard deviation 0.4813 K against the Laplace 0.4806 K, a gap of 0.0007 K, and the mean 23.0300°C against the mode 23.0701°C.

3.5.7 Dependence inside and outside the physics graph

The common cause of Section 3.4.4 cost a factor of nine because it was left out of the product. A plant graph built on conservation relations does not leave out a shared cause that has a physical path: the cause is an entry of the state, and every row that depends on it carries it. The two-entry hall above shows this. The reading and the conductance share no row, yet the physics row couples them, and the posterior correlation of T and g, -0.9998, comes out of $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ without being assumed. Had the site-wide workload been a source of the plant graph,

feeding the load of each hall, with a row for each hall's exceedance given its state, the joint exceedance probability (3.17) would have followed from the rows, and the independent product of Section 3.4.5 could not have been formed. Conditional independence given the state is then the statement of Section 3.4.6 with the common cause known: it holds because the cause is in the state.

What the graph does not supply is the independence of the errors. Equation (3.20) assumes that a reading's error is independent of the inputs and, when there are several readings, of the other readings' errors, so that the posterior (3.24) is a product over rows. No conservation relation states that, and none checks it directly. Suppose the hall of Section 3.5.4 carries two sensors from one calibration batch, each with $\sigma_T = 0.5\,\mathrm{K}$, both reading 27°C, with errors correlated through the offset they share, $\mathrm{corr}(\epsilon_1, \epsilon_2) = 0.5$. Treated as independent rows, the pair adds the precision $2/\sigma_T^2 = 8\,\mathrm{K}^{-2}$ to the prior's 0.25. With the correlation, equation (3.27) holds with $\mathbf{R}$ a full matrix, its off-diagonal entry $0.5\,\sigma_T^2$ and its rows whitened by $\mathbf{R}^{-1/2}$, and the pair adds only

$$\mathbf{A}^\mathsf{T}\mathbf{R}^{-1}\mathbf{A} = \frac{2}{\sigma_T^2\,(1 + 0.5)} = 5.33\,\mathrm{K}^{-2}, \qquad \mathbf{A} = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

the precision of the pair's average, whose error has the variance $\sigma_T^2\,(1 + 0.5)/2$. Table 3.8 compares the two treatments with the single sensor of Table 3.7. Its columns are the precision the readings add, the posterior mean, the posterior standard deviation and the probability of exceeding $T_{\max} = 28$°C.

Table 3.8: Two hall sensors of $\sigma_T = 0.5\,\mathrm{K}$, both reading 27°C, against the prior $\mathcal{N}(26, 2^2)$ and the limit $T_{\max} = 28$°C: the pair treated as independent, the pair with $\mathrm{corr}(\epsilon_1, \epsilon_2) = 0.5$, and one sensor, which is also the limit of a correlation of one.

Readings	precision added [K^{-2}]	mean [°C]	σ_{post} [K]	$\Pr(T > T_{\max} \mid \mathbf{y})$ [–]
two, treated as independent	8.00	26.9697	0.3482	0.00154
two, correlation 0.5	5.33	26.9552	0.4232	0.00678
one sensor	4.00	26.9412	0.4851	0.0145

The independent treatment reports an exceedance probability of 0.00154 against the 0.00678 the correlated pair supports, 4.4 times too small, and a posterior standard deviation of 0.348 K against 0.423 K. The error has the direction of Section 3.4.5's: a posterior narrower than the evidence supports, and a limit that looks safer than it is. At a correlation of one the second sensor adds nothing, and the posterior is that of one sensor.

The shared offset is a common cause, and the remedy of Section 3.4.6 applies: put it in the state. A bias entry $b \sim \mathcal{N}(0,\ 0.5\,\sigma_T^2)$ carried by both reading rows, with independent remainders of variance $0.5\,\sigma_T^2$, makes the two rows conditionally independent given (T, b), and the state (T, b) gives the posterior of T of the correlated row of Table 3.8 to every printed digit. The posterior of b, however, has the standard deviation 0.348 K against its prior's 0.354 K, with $\mathrm{corr}(T, b \mid \mathbf{y}) = -0.81$. The readings barely inform it: a shared offset moves both readings together, as a change of T does, and only the prior on b bounds it.

The same assumption hides in four other places a physics graph does not reach: a prior shared by several entries, such as two conductances taken from one handbook; a physics tolerance σ_p treated as independent from row to row where the model's error is systematic; repeated readings averaged as if their errors were independent when they are correlated in time; and a path of the conserved quantity that the graph omits. Each makes the posterior narrower

than the evidence supports. What the graph adds is redundancy. When more rows bear on the state than it has entries, a misfit of the rows taken together, a reading far from what the other rows predict, or a comparison of drawings with and without the suspected cause can reveal the dependence, and Chapter 8 carries the halls' shared workload as entries of the state. No test reveals a dependence that moves the readings along a direction the physics also allows, as the shared offset moves them along T.

Principle 3.3. *A physics graph carries dependence through shared physical quantities as entries of the state, so the independent product cannot be formed for them. The independence it still assumes is that of the errors: it must be stated, tested where the rows are redundant, and, where it fails, removed by putting the shared cause in the state.*

3.6 Predictive variance and the covariance of a joint posterior

Two variances complete the chapter: the law of total variance, which separates the randomness of future exceedances from the uncertainty of their rate, and the covariance of a posterior over a joint state, which decides how uncertain a derived quantity is and why the shape of the posterior must be read rather than its marginals.

3.6.1 The law of total variance

Suppose the number K_τ of the hall's exceedances of its committed load during a future period of length τ is Poisson given the rate, $K_\tau \mid \lambda \sim \text{Poisson}(\lambda\tau)$, and the rate itself is uncertain, with posterior $p(\lambda \mid \mathbf{y})$. The predictive variance of K_τ is wanted. Two facts give it. The law of total expectation, which is the law of total probability applied inside an expectation, states that $\mathbb{E}[K_\tau \mid \mathbf{y}] = \mathbb{E}\big[\mathbb{E}[K_\tau \mid \lambda] \mid \mathbf{y}\big]$, the conditional mean averaged over the rate; and the variance of any quantity is the mean of its square minus the square of its mean. Then

$$\operatorname{Var}(K_\tau \mid \mathbf{y}) = \mathbb{E}\big[\mathbb{E}[K_\tau^2 \mid \lambda] \mid \mathbf{y}\big] - \big(\mathbb{E}\big[\mathbb{E}[K_\tau \mid \lambda] \mid \mathbf{y}\big]\big)^2,$$

and writing the conditional second moment as $\mathbb{E}[K_\tau^2 \mid \lambda] = \operatorname{Var}(K_\tau \mid \lambda) + \mathbb{E}[K_\tau \mid \lambda]^2$,

$$\operatorname{Var}(K_\tau \mid \mathbf{y}) = \mathbb{E}\big[\operatorname{Var}(K_\tau \mid \lambda) \mid \mathbf{y}\big] + \Big\{\mathbb{E}\big[\mathbb{E}[K_\tau \mid \lambda]^2 \mid \mathbf{y}\big] - \big(\mathbb{E}\big[\mathbb{E}[K_\tau \mid \lambda] \mid \mathbf{y}\big]\big)^2\Big\}.$$

The brace is the variance of the conditional mean, which gives the *law of total variance*,

$$\operatorname{Var}(K_\tau \mid \mathbf{y}) = \mathbb{E}\big[\operatorname{Var}(K_\tau \mid \lambda) \mid \mathbf{y}\big] + \operatorname{Var}\big(\mathbb{E}[K_\tau \mid \lambda] \mid \mathbf{y}\big). \tag{3.32}$$

The first term is the expected variability of the exceedance count when the rate is specified, the aleatory part. The second is the additional predictive variability because the rate is uncertain, the epistemic part. For a Poisson count the mean and the variance are equal, $\mathbb{E}[K_\tau \mid \lambda] = \operatorname{Var}(K_\tau \mid \lambda) = \lambda\tau$, so that $\operatorname{Var}(K_\tau \mid \mathbf{y}) = \mathbb{E}[\lambda\tau \mid \mathbf{y}] + \operatorname{Var}(\lambda\tau \mid \mathbf{y})$, and since τ is fixed,

$$\operatorname{Var}(K_\tau \mid \mathbf{y}) = \tau\,\mathbb{E}[\lambda \mid \mathbf{y}] + \tau^2\,\operatorname{Var}(\lambda \mid \mathbf{y}). \tag{3.33}$$

The predictive variance has been separated into a process term and a rate term.

3.6.2 Exceedance counts over the next design day

Take the posterior of the hall's exceedance rate from the state, the normal of the log-rate of equation (2.41), with $\hat{\omega} = \ln(8/3)$ and $\sigma_\omega = 0.5$, so that λ is lognormal about $\hat{\lambda} = 4/1500\,\text{h}$. The standard lognormal moments give

$$\mathbb{E}[\lambda \mid \mathbf{y}] = \hat{\lambda}\,\mathrm{e}^{\sigma_\omega^2/2} = 3.022 \times 10^{-3}\,\text{h}^{-1}, \qquad \operatorname{Var}(\lambda \mid \mathbf{y}) = \hat{\lambda}^2\,\mathrm{e}^{\sigma_\omega^2}\big(\mathrm{e}^{\sigma_\omega^2} - 1\big) = 2.593 \times 10^{-6}\,\text{h}^{-2}.$$

Over the peak hours of the next design day, $\tau = 12\,\text{h}$, the expected exceedance count is $\mathbb{E}[K_\tau \mid \mathbf{y}] = 12 \times 3.022 \times 10^{-3} = 0.0363$, and by equation (3.33) the predictive variance is

$$\text{Var}(K_\tau \mid \mathbf{y}) = 0.0363 + 12^2 \times 2.593 \times 10^{-6} = 0.0363 + 0.0004 = 0.0366.$$

The process term is 99.0 % of the predictive variance and the rate term 1.02 %. The count 0.0363 is a mean, not the probability of an exceedance: a day with two exceedances counts twice in the mean, so the probability of at least one, 0.0354 from the same normal, lies below it.

The exact posterior $\text{Gamma}(a, b) = \text{Gamma}(4, 1500\,\text{h})$ stands beside these, with the standard gamma moments $\mathbb{E}[\lambda \mid \mathbf{y}] = a/b = 2.667 \times 10^{-3}\,\text{h}^{-1}$, one exceedance in about 375 h, and $\text{Var}(\lambda \mid \mathbf{y}) = a/b^2 = 1.778 \times 10^{-6}\,\text{h}^{-2}$. A rate of one exceedance of the committed load in 375 logged hours is that of a hall running at its committed load, as this one does today at 30 kW, whose brief workload peaks cross the allocation now and then. The same two terms are

$$\text{Var}(K_\tau \mid \mathbf{y}) = 12 \left(\frac{4}{1500} \right) + 12^2 \left(\frac{4}{1500^2} \right) = 0.032000 + 0.000256 = 0.032256,$$

the mean 0.0320 against the reported 0.0363, a gap of 0.0043, and the variance 0.0323 against 0.0366; the probability of an exceedance, 0.0314, again lies below its mean. The share of the rate term under the gamma is

$$\frac{\tau^2 a/b^2}{\tau a/b + \tau^2 a/b^2} = \frac{\tau}{b + \tau},$$

which for $\tau = 12\,\text{h}$ and $b = 1500\,\text{h}$ is $12/1512 = 0.0079$, against the reported 0.0102. The gamma's result can be checked against its full predictive distribution. A Poisson count whose mean is averaged over a gamma rate has, by a standard result, the negative binomial distribution with shape a and success probability $\pi = b/(b + \tau)$, whose mean is $a(1 - \pi)/\pi = a\tau/b$ and whose variance is $a(1 - \pi)/\pi^2 = a\tau/b + a\tau^2/b^2$: exactly the two terms of equation (3.33). For $\pi = 1500/1512 = 0.99206$ they are 0.032000 and 0.032256.

Table 3.9 extends the calculation to longer periods. The first column is the period; the next four are the reported values from the state, the expected count, which is also the process term, the rate term, the total and the share of the rate term; and the last three the gamma's mean, variance and share. Over a design day almost all of the predictive variance comes from the random exceedance process, and a longer exceedance log, which shrinks only the rate term, would barely change the forecast of the next design day's exceedances. The two terms are equal at about 1165 h (at 1500 h under the gamma, as long as the logged hours behind the posterior), and over 3000 h the rate term is 0.720 of the total (0.667 under the gamma). The gaps grow with the period, because the lognormal's mean rate lies above the gamma's and its upper tail is heavier. Whether a longer log will meaningfully improve an exceedance forecast depends on this split: more evidence shrinks the second term, and the first, which the design must carry, remains as long as the future count is modeled as a Poisson process.

3.6.3 One posterior over a joint state

The examples so far had one unknown at a time. A plant graph carries many at once. Suppose the state vector holds the air temperatures of two halls, the log-conductance of an air path and the rates at which the two halls' loads exceed their committed loads,

$$\mathbf{z} = \begin{bmatrix} T_A & T_B & g & \lambda_A & \lambda_B \end{bmatrix}^\mathsf{T},$$

and the evidence $\mathbf{y}$ gathers temperature readings, the exceedance logs and test results. Everything known after the evidence is the joint posterior $p(\mathbf{z} \mid \mathbf{y}, M)$, where M names the model: the graph, its rows and its priors. Different questions are asked of this same distribution:

Table 3.9: Predictive mean and variance of the hall's count of committed-load exceedances over a future period τ, equation (3.33): reported from the normal of the log-rate, with the posterior Gamma$(4, 1500\,\mathrm{h})$ beside it; the first row is the twelve peak hours of one design day.

τ [h]	mean = process term [–]	rate term [–]	variance [–]	rate share [–]	gamma mean [–]	gamma variance [–]	gamma share [–]
12	0.0363	0.0004	0.0366	0.0102	0.0320	0.0323	0.0079
100	0.3022	0.0259	0.3281	0.0790	0.2667	0.2844	0.0625
500	1.5109	0.6483	2.1592	0.3003	1.3333	1.7778	0.2500
1500	4.5326	5.8351	10.3677	0.5628	4.0000	8.0000	0.5000
3000	9.0652	23.3405	32.4057	0.7203	8.0000	24.0000	0.6667

State estimation, $p(T_A \mid \mathbf{y}, M)$: the marginal of one entry.

Threshold probability, $\Pr(T_B > T_{\max} \mid \mathbf{y}, M)$: a tail of another marginal.

Parameter comparison, $\Pr(\lambda_A > \lambda_B \mid \mathbf{y}, M)$: a tail of the derived quantity $\lambda_A - \lambda_B$, which depends on the joint distribution of the two rates.

Expected derived quantity, $\mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}, M]$, such as the expected number of committed-load exceedances in the two halls over the next design day: the average of a function of the state.

Prediction, $\Pr(\text{no exceedance in } \tau \mid \mathbf{y}, M)$: a posterior predictive probability, equation (2.41) with the rate taken from the state.

The methods of computation differ, but every one of them depends on a consistent probability model of the physical plant and the evidence, and none needs a separate inference.

3.6.4 The posterior covariance and the variance budget

When the posterior can be approximated locally by a multivariate Gaussian, $\mathbf{z} \mid \mathbf{y}, M \approx \mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$, the covariance matrix $\boldsymbol{\Sigma}$ describes the uncertainty of each entry and how the entries vary together. The approximation takes the most probable state as the mean and the inverse curvature of the objective there as the covariance. For rows linear in the state it is exact, $\boldsymbol{\Sigma} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}$ by Section 3.5.3; for nonlinear rows it is the Laplace approximation, which Chapter 4 derives with its curvature term. For two entries,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}, \qquad \operatorname{corr}(z_1, z_2 \mid \mathbf{y}) = \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}}.$$

The diagonal entries are the marginal variances, $\Sigma_{11} = \operatorname{Var}(z_1 \mid \mathbf{y})$; the off-diagonal entry is the covariance, and $\operatorname{corr}(z_1, z_2 \mid \mathbf{y})$ the correlation. The off-diagonal entries matter because physical relations and shared observations introduce posterior dependence between unknowns; they are not zero merely because the priors of the entries were independent. The hall of Section 3.5.6 is one case. Another carries the supply air as an entry: the state $\mathbf{z} = (T,\ T_c,\ g)^{\mathsf{T}}$ for the illustrative hall with a weak air path of Figure 3.2(c), with $Q = 30\,\mathrm{kW}$, $G_{\mathrm{ref}} = 3\,\mathrm{kW/K}$, independent priors $T_c \sim \mathcal{N}(20, 1^2)$ and $g \sim \mathcal{N}(0, 0.05^2)$ (the $0.15\,\mathrm{kW/K}$ of the figure), a reading with $\sigma_T = 0.5\,\mathrm{K}$ equal to the prior prediction 30°C, and the physics row $(T - T_c - 10\,\mathrm{K}\ \mathrm{e}^{-g})/\sigma_p$ with $\sigma_p = 0.01\,\mathrm{K}$. At the prior means every residual is zero and the rows of $\mathbf{J}$ are $(1,\ -1,\ 10)/\sigma_p$ for the physics, $(1,\ 0,\ 0)/\sigma_T$ for the reading, and $(0,\ 1,\ 0)$ and $(0,\ 0,\ 1)/0.05$ for the priors; equation (3.26) gives

$$\boldsymbol{\Sigma} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1} = \begin{bmatrix} 0.2083 & 0.1667 & -0.0042 \\ 0.1667 & 0.3334 & 0.0167 \\ -0.0042 & 0.0167 & 0.0021 \end{bmatrix},$$

and the correlation of T_c and g is $0.0167/\sqrt{0.3334 \times 0.0021} = 0.632$ (unrounded entries; 0.6325 at $\sigma_p = 0.001\,\mathrm{K}$). Without the reading row the same computation gives zero. A single reading leaves T_c and G positively correlated, since a warmer supply air and a larger conductance explain the same reading, and the correlation enters through the physics row, the only row with entries in the columns of both.

Suppose a derived quantity $f(\mathbf{z})$ has the gradient $\mathbf{g} = \partial f/\partial \mathbf{z}$ at $\hat{\mathbf{z}}$, a column with one entry per entry of the state. With $\boldsymbol{\delta} = \mathbf{z} - \hat{\mathbf{z}}$ the deviation of the state from its mean, the first-order change is $f - f(\hat{\mathbf{z}}) \approx \mathbf{g}^\mathsf{T}\boldsymbol{\delta}$, a scalar equal to its own transpose, and the covariance is $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{\delta}\boldsymbol{\delta}^\mathsf{T}]$, so that

$$\mathrm{Var}(f) \approx \mathbb{E}[\mathbf{g}^\mathsf{T}\boldsymbol{\delta}\,\boldsymbol{\delta}^\mathsf{T}\mathbf{g}] = \mathbf{g}^\mathsf{T}\,\mathbb{E}[\boldsymbol{\delta}\boldsymbol{\delta}^\mathsf{T}]\,\mathbf{g} = \mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g}. \tag{3.34}$$

This is the *variance budget* of the output. Written out entry by entry, $\mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g}$ has one term per entry of the state, the squared sensitivity of the output to the entry times the variance of the entry, plus a cross term $2\,\nabla_k\nabla_l\,\Sigma_{kl}$, with $\nabla_k = \partial f/\partial z_k$, for every pair of entries that move together. For the sum $f = z_1 + z_2$ the gradient is $\mathbf{g} = (1,\ 1)^\mathsf{T}$, and the budget is exact because f is linear,

$$\mathrm{Var}(z_1 + z_2) = \begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12} & \Sigma_{22}\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix} = \Sigma_{11} + \Sigma_{22} + 2\Sigma_{12}. \tag{3.35}$$

With $\Sigma_{11} = \Sigma_{22} = 1\,\mathrm{K}^2$ (illustrative), the variance of the sum is $0.4\,\mathrm{K}^2$ for $\Sigma_{12} = -0.8\,\mathrm{K}^2$, $2\,\mathrm{K}^2$ for independent entries and $3.6\,\mathrm{K}^2$ for $\Sigma_{12} = 0.8\,\mathrm{K}^2$. A negative covariance makes the output more precisely known than the individual entries suggest; a positive one makes it less.

Table 3.10 works the budget for the hall air of equation (3.19), with the load known, $Q = 30\,\mathrm{kW}$, and the two unknowns of Figure 3.2(c) as the state $\mathbf{z} = (T_c,\ G)^\mathsf{T}$: T_c with standard deviation 1 K and G with mean 3 kW/K and standard deviation 0.15 kW/K. The gradient of T and the covariance are

$$\mathbf{g} = \begin{bmatrix}\partial T/\partial T_c\\ \partial T/\partial G\end{bmatrix} = \begin{bmatrix}1\\ -Q/G^2\end{bmatrix} = \begin{bmatrix}1\\ -3.333\end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix}1 & 0.15\,c\\ 0.15\,c & 0.0225\end{bmatrix}, \qquad c = \mathrm{corr}(T_c, G),$$

with $\mathbf{g}$ in K per unit of each entry and $\boldsymbol{\Sigma}$ in the products of the entries' units, so that $\mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g} = (1.00 + 0.25 - 1.00\,c)\,\mathrm{K}^2$: the supply-air term $1^2 \times 1$, the conductance term $3.333^2 \times 0.0225$ and the cross term $2 \times 1 \times (-3.333) \times 0.15\,c$. The columns of the table are the correlation, the three terms, the variance and the standard deviation of T. The two marginals are the same in every row, and the standard deviation of the hall air ranges from 0.67 to 1.43 K with the correlation alone. The budget is first order, since T is curved in G; a quadrature over the same joint normal gives 0.6701, 1.1203 and 1.4357 K against the budget's 0.6708, 1.1180 and 1.4318 K, gaps of at most 0.0039 K.

Table 3.10: Variance budget of the hall air $T = T_c + Q/G$, equation (3.34), with $Q = 30\,\mathrm{kW}$, G of mean 3 kW/K and standard deviation 0.15 kW/K, T_c of standard deviation 1 K, and correlation $c = \mathrm{corr}(T_c, G)$ (illustrative).

c [–]	supply-air term [K^2]	conductance term [K^2]	cross term [K^2]	Var(T) [K^2]	sd [K]
−0.8	1.00	0.25	+0.80	2.05	1.43
0	1.00	0.25	0	1.25	1.12
0.8	1.00	0.25	−0.80	0.45	0.67

Two consequences follow from equation (3.34). An entry can be very uncertain and still not matter: if the output is insensitive to it, its whole term is near zero, and narrowing its variance buys nothing. And an entry can be poorly known while the output is well known:

when the sensitivities and the covariance have opposite signs, the cross term subtracts from the diagonal terms. A measurement before committing is worth buying when it shrinks a variance Σ_{kk} whose sensitivity is large, and worthless when the sensitivity is near zero. Chapter 17 builds the variance budget of the site's design inputs on equation (3.34), and Chapter 18 ranks the measurements by the width each takes off it.

3.6.5 Reading the shape of the posterior

The second consequence deserves a worked case, because it is where marginal spreads mislead. Carry the hall air, the load and the conductance as entries, the last two through the logarithms $\ell_Q = \ln(Q/Q_{\mathrm{ref}})$ and $g = \ln(G/G_{\mathrm{ref}})$, each with an independent prior of mean 0 and standard deviation 0.10, that is about 10 %, and with the supply air pinned. The references are chosen so that $Q_{\mathrm{ref}}/G_{\mathrm{ref}} = 10\,\mathrm{K}$, a temperature rise of 10 K at the prior means, and a sensor with $\sigma_T = 0.2\,\mathrm{K}$ reads that temperature, $y_T = T_c + 10\,\mathrm{K}$. The state, its four rows and their 4×3 Jacobian are

$$\mathbf{z} = \begin{bmatrix} T \\ \ell_Q \\ g \end{bmatrix}, \qquad r(\mathbf{z}) = \begin{bmatrix} (T - y_T)/\sigma_T \\ \ell_Q/0.10 \\ g/0.10 \\ (T - T_c - 10\,\mathrm{K}\,\mathrm{e}^{\ell_Q - g})/\sigma_p \end{bmatrix},$$

$$\mathbf{J} = \begin{bmatrix} 1/\sigma_T & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \\ 1/\sigma_p & -10\,\mathrm{e}^{\ell_Q - g}/\sigma_p & 10\,\mathrm{e}^{\ell_Q - g}/\sigma_p \end{bmatrix},$$

with the physics row of equation (3.19) carrying the tolerance $\sigma_p = 0.001\,\mathrm{K}$. At the prior means, $\mathbf{z}_0 = (T_c + 10\,\mathrm{K},\ 0,\ 0)^{\mathsf{T}}$, every residual is zero, the normal equations (3.25) give a zero step, and $\hat{\mathbf{z}} = \mathbf{z}_0$; the curvature term that Chapter 4 adds for nonlinear rows is a sum of residuals times second derivatives and vanishes with the residuals. There the Jacobian is

$$\mathbf{J} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \\ 1000 & -10\,000 & 10\,000 \end{bmatrix},$$

whose last two columns carry the relation between load and conductance: both are zero in the reading row and equal and opposite in the physics row, so the direction $(0,\ 1,\ 1)^{\mathsf{T}}$, load and conductance raised together, gives $\mathbf{J}\,(0,\ 1,\ 1)^{\mathsf{T}} = (0,\ 10,\ 10,\ 0)^{\mathsf{T}}$ and is seen by the prior rows alone. Multiplying by $\mathbf{J}^{\mathsf{T}}$,

$$\mathbf{J}^{\mathsf{T}}\mathbf{J}\,(0,\ 1,\ 1)^{\mathsf{T}} = 100\,(0,\ 1,\ 1)^{\mathsf{T}},$$

so $(0,\ 1,\ 1)^{\mathsf{T}}$ is an eigenvector of $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ with the eigenvalue $100 = 1/0.10^2$, the prior precision alone, and of $\boldsymbol{\Sigma} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}$ with the eigenvalue 0.0100, the prior variance: the reading has left that direction untouched. Inverting $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ numerically, equation (3.26) gives

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.039216 & 0.001961 & -0.001961 \\ 0.001961 & 0.005098 & 0.004902 \\ -0.001961 & 0.004902 & 0.005098 \end{bmatrix}.$$

For a Gaussian, the integral of equation (3.16) over T leaves the block of $\boldsymbol{\Sigma}$ in the rows and columns of ℓ_Q and g, the lower right 2×2 block. Its precision is not the corresponding block of $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ but, by the block-inversion formula of Section 3.5.3 applied to $\boldsymbol{\Lambda} = \mathbf{J}^{\mathsf{T}}\mathbf{J}$, the Schur complement of the T block, $\boldsymbol{\Lambda}_{qq} - \boldsymbol{\Lambda}_{qT}\boldsymbol{\Lambda}_{TT}^{-1}\boldsymbol{\Lambda}_{Tq}$ with q the indices of ℓ_Q and g:

$$\boldsymbol{\Lambda}_{qq} - \boldsymbol{\Lambda}_{qT}\boldsymbol{\Lambda}_{TT}^{-1}\boldsymbol{\Lambda}_{Tq} = \begin{bmatrix} 2599.94 & -2499.94 \\ -2499.94 & 2599.94 \end{bmatrix} \longrightarrow \begin{bmatrix} 2600 & -2500 \\ -2500 & 2600 \end{bmatrix} \quad (\sigma_p \to 0),$$

against the block $\mathbf{\Lambda}_{qq}$ of $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ itself, whose entries are $10^8 + 100$ on the diagonal and -10^8 off it. Each marginal standard deviation is $\sqrt{0.005098} = 0.0714$, down only from 0.100, and the correlation is $0.004902/0.005098 = 0.962$. The eigenvectors of the block are $(1,\ 1)^{\mathsf{T}}/\sqrt{2}$ and $(1,\ -1)^{\mathsf{T}}/\sqrt{2}$, with the eigenvalues 0.005098 ± 0.004902, that is 0.010000 and 0.000196: the posterior is as wide as the prior along the first and 7.1 times narrower along the second. The budget (3.34) with the gradients $\mathbf{g} = (0,\ 1,\ -1)^{\mathsf{T}}$ and $(0,\ 1,\ 1)^{\mathsf{T}}$ gives the variance of the difference, $\mathbf{g}^{\mathsf{T}}\mathbf{\Sigma}\mathbf{g} = 0.000392$, a standard deviation of 0.0198, and that of the sum, 0.0200, a standard deviation of 0.1414, exactly its prior value $\sqrt{2} \times 0.10$. Repeated at $\sigma_p = 0.01\,\mathrm{K}$, the block entries print the same, 0.005098 and 0.004902, and the correlation prints 0.961 against 0.962 at 0.001 K; the limit is reached at 0.001 K.

Remark 3.2. Eliminating T through the physics row turns the reading into the linearized row $(\ell_Q - g)/0.02$, since $\sigma_T/(T - T_c) = 0.02$, and the same block follows from a 3×2 Jacobian; along the difference alone, the prior precision 50 plus the reading's 2500 gives $1/2550 = 0.000392$.

Figure 3.7 draws the two distributions in the plane of the two entries. The horizontal axis is ℓ_Q and the vertical axis g; the dashed gray circle encloses 95 % of the prior, the green ellipse 95 % of the posterior, and the thin line is $\ell_Q - g = 0$, the value of the difference that the reading fixed. The posterior is a narrow ellipse lying along that line, its axes the two eigenvectors of the block: it is long in the direction the reading cannot see, $(1,\ 1)^{\mathsf{T}}$, where it reaches the prior circle because the reading added nothing there, and short in the direction it measured, $(1,\ -1)^{\mathsf{T}}$.

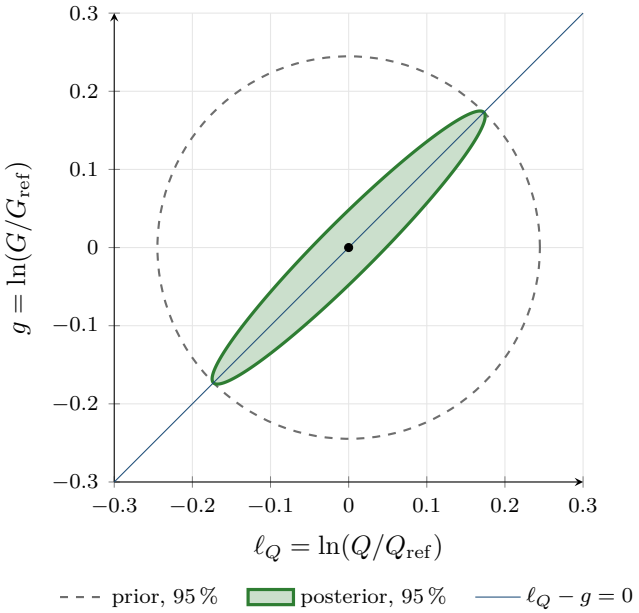


Figure 3.7: Prior and posterior of the log-load ℓ_Q and the log-conductance g after one hall-air reading that fixes $\ell_Q - g$ within 0.02 (illustrative): the (ℓ_Q, g) block of the posterior of $\mathbf{z} = (T, \ell_Q, g)^{\mathsf{T}}$. Prior: independent, standard deviations 0.10. Posterior: standard deviations 0.0714, correlation 0.962; standard deviation of $\ell_Q - g$, 0.0198; of $\ell_Q + g$, 0.1414. Contours enclose 95 % ($\chi^2_{2,0.95} = 5.991$).

Read by its marginals, this posterior says that neither the load nor the conductance is well known, about 7 % each. Read by its shape, it says that their ratio, and with it the hall air at the present load, is known to 2 %. A prediction at a different, specified load Q', such as the design load, $T' = T_c + Q'/G$, has a gradient with respect to (T, ℓ_Q, g) proportional to $(0,\ 0,\ 1)^{\mathsf{T}}$, so $\mathbf{g}^{\mathsf{T}}\mathbf{\Sigma}\mathbf{g}$ picks out Σ_{gg} and the prediction inherits its 7 % spread; a prediction that scales the present load, such as the hall at twice the unknown Q, has a gradient proportional to

$(0,\ 1,\ -1)^{\mathsf{T}}$, the short axis of the ellipse, and its temperature rise is known to the same 2 % relative as the present one; the absolute spread doubles. Which statement holds depends on the gradient of the question, and the covariance, not the list of marginal variances, carries the answer.

Principle 3.4. *Read the posterior's shape, not its marginals. Wide marginal spreads with a strong correlation mean that the evidence determines a combination of the entries; a quantity that depends on that combination is well known, and the entries are not unobservable merely because each alone is uncertain.*

3.6.6 Mixtures over structure

A Gaussian approximation is most useful for continuous uncertainty near a well-identified posterior mode. Conditional probability is more general. If the plant may operate under either of two structures, $M \in \{M_1, M_2\}$ (two drawings of the plant), the law of total probability, applied over the models with $\mathbf{z}$ conditioned on the evidence throughout, gives

$$p(\mathbf{z} \mid \mathbf{y}) = \sum_j p(\mathbf{z}, M_j \mid \mathbf{y}) = \sum_j p(\mathbf{z} \mid \mathbf{y}, M_j)\ \Pr(M_j \mid \mathbf{y}), \tag{3.36}$$

where the second step is the multiplication rule. The weights follow from Bayes' theorem over the models, $\Pr(M_j \mid \mathbf{y}) = p(\mathbf{y} \mid M_j)\ \Pr(M_j) / \sum_i p(\mathbf{y} \mid M_i)\ \Pr(M_i)$, so that the evidence of each model, the normalizing constant of its own posterior, sets its share. The expression assumes that the relevant state $\mathbf{z}$ has a common interpretation across the models. Such a posterior need not be Gaussian, even when each model's posterior is.

A two-structure illustration on the hall air shows what is lost. Let the drawing M_1, the air path as designed, give $T \mid \mathbf{y}, M_1 \sim \mathcal{N}(26.9, 0.5^2)$, and a drawing M_2 with a degraded path give $\mathcal{N}(28.5, 0.5^2)$, with posterior model probabilities 0.7 and 0.3 (illustrative). By the law of total expectation the mixture mean is $0.7 \times 26.9 + 0.3 \times 28.5 = 27.38\,^{\circ}\mathrm{C}$; by the law of total variance, equation (3.32) with the model in place of the rate, its variance is the mean of the within-model variances plus the variance of the model means,

$$\mathrm{Var}(T \mid \mathbf{y}) = 0.5^2 + 0.7 \times 0.3 \times (28.5 - 26.9)^2 = 0.25 + 0.5376 = 0.7876\,\mathrm{K}^2,$$

a standard deviation of 0.887 K. The probability of exceeding 28 °C is the weighted sum of the two tails,

$$\Pr(T > 28\,^{\circ}\mathrm{C} \mid \mathbf{y}) = 0.7\,(1 - \Phi(2.2)) + 0.3\,(1 - \Phi(-1)) = 0.7 \times 0.0139 + 0.3 \times 0.8413 = 0.262.$$

The most probable drawing alone would report 0.014, and a single Gaussian with the mixture's mean and variance would report $1 - \Phi(0.699) = 0.242$ and hide the two separate modes, 3.2 standard deviations apart. The law of total probability combines the uncertainty over the structure with the uncertainty over the numerical state, and this combination becomes central in deciding which drawing of the plant the operating record supports before a design is built on it. The next chapter assembles these pieces into the complete workflow of probabilistic inference, from the plant graph and its likelihood to the posterior, its marginals, its expectations and the probabilities of functions of the state.

3.7 What to retain

Definition 3.1 (Conditioning, marginalization and dependence)**.** A conditional probability restricts attention to the situations consistent with the evidence and renormalizes, $\Pr(A \mid B) =$

$\Pr(A \cap B)/\Pr(B)$; Bayes' theorem reverses the conditioning through the prior and the evidence; marginalization sums or integrates over every unknown the question does not name; and the posterior covariance of a joint state, not its marginal variances, decides how uncertain a derived quantity is.

Four ideas carry forward. Design questions contain two kinds of uncertainty: aleatory uncertainty is the day-to-day scatter of the next reading, load or exceedance count, which the design must carry, and epistemic uncertainty is incomplete knowledge of the plant, its conductance or a meter's calibration, which a test or a survey removes; more windows reduce the second and leave the first, and the law of total variance separates them in a prediction. Conditional probability is how information is incorporated: Bayes' theorem updates a probability when an observation arrives, the base rate decides how far a flag moves it, and the fraction of flags that mark a real shortfall is not the flag's probability of detection. Plant quantities are often dependent: a shared workload, a conservation relation or a common supply air creates relations among halls, units and unknown states, the joint probability of the halls' peaks lies beneath the coincidence factor, and conditional independence simplifies a calculation only where the model justifies it. A physics graph puts the shared physical causes in the state; the independence it still assumes is that of the errors, and a shared error, such as a calibration offset, is removed by carrying it as an entry too. Marginalization answers a focused question without pretending that the other unknowns are known, whether they are nuisance quantities, competing conditions of a unit or competing drawings of the plant.

Together these establish the mechanism of inference: the prior $p(\mathbf{z})$ and the observation model $p(\mathbf{y} \mid \mathbf{z})$ combine into the posterior $p(\mathbf{z} \mid \mathbf{y})$, the update multiplying the prior by the likelihood and normalizing the product. For Gaussian rows the product is $\exp(-\frac{1}{2}r^{\mathsf{T}}r)$ over the whole state, physics rows included; the normal equations give its most probable state, $\boldsymbol{\Sigma} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}$ its covariance, exact when the rows are linear, and the conditional of the joint Gaussian of state and readings, through the Schur complement, gives the same answer. Relations between quantities are read in that space: a reading informs an entry it does not measure through a row of $\mathbf{J}$ that carries both, and a combination of entries is well known when it lies along a short eigenvector of $\boldsymbol{\Sigma}$, which $\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}$ measures. Once the posterior is available, the planner's question decides the next calculation: a marginal probability, an expectation, a prediction, a model comparison or a decision. Table 3.11 lists the chapter's terms with their expressions, units and the subsection that defines each.

Table 3.11: The terms of uncertainty and conditional probability and where each is defined.

Term	Expression	Units	Meaning	Defined in
Aleatory uncertainty	$\Pr(y_T > c \mid T) = 1 - \Phi((c - T)/\sigma_T)$	–	readings vary with the hall known	3.1.1
Epistemic uncertainty	$p(g \mid \mathbf{y})$	–	knowledge of a fixed quantity	3.1.2
Predictive exceedance	$\int \Pr(y_{\text{new}} > c \mid T)\, p(T \mid \mathbf{y})\, \mathrm{d}T$	–	next reading, averaged over the hall	3.1.3
Conditional probability	$\Pr(A \cap B)/\Pr(B)$	–	probability of A among cases of B	3.2.2
Multiplication rule	$\Pr(A \cap B) = \Pr(B \mid A)\Pr(A)$	–	joint from conditional	3.2.3
Bayes' theorem	$\Pr(B \mid A)\Pr(A)/\Pr(B)$	–	reverses the conditioning	3.2.4
Posterior odds	likelihood ratio $\times$ prior odds	–	Bayes without the evidence	3.3.3
Law of total probability	$\sum_s \Pr(H \mid s)\Pr(s)$	–	marginal of an observation	3.4.1
Marginal likelihood	$\int p(\mathbf{y} \mid \lambda) p(\lambda)\, \mathrm{d}\lambda$	–	evidence; normalizer	3.4.2
Design day within the committed load	$\sum_j w_j\, \mathrm{e}^{-\lambda_{\text{ref}} \mathrm{e}^{\omega_j \tau}} = 0.96457$	–	reported, no exceedance over the log-rate, equation (2.41)	3.4.2
Its exact value	$(b/(b+\tau))^a = 0.96863$	–	same average over the gamma; gap 0.0041	3.4.2
Nuisance marginal	$\int p(g, b_Q \mid \mathbf{y})\, \mathrm{d}b_Q$	varies	integrate out, do not maximize	3.4.3
Conditional independence	$X_1 \perp X_2 \mid C$	–	joint given c factorizes	3.4.6
Common-cause joint	$\sum_c \Pr(F_A \mid c)\Pr(F_B \mid c)\Pr(c)$	–	both halls above committed with a shared cause	3.4.5
Correlated reading errors	$\mathbf{A}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{A}$, $\mathbf{R}$ full; or a shared bias entry	K^{-2}	the precision the readings truly add	3.5.7
Gaussian rows	$p(\mathbf{z} \mid \mathbf{y}) \propto \exp(-\frac{1}{2} r^{\mathsf{T}} r)$	–	posterior of the whole state	3.5.2
Normal equations	$(\mathbf{J}^{\mathsf{T}}\mathbf{J})\,\boldsymbol{\Delta} = -\mathbf{J}^{\mathsf{T}} r$, $\boldsymbol{\Sigma} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}$	varies	most probable state, covariance	3.5.3
Precisions add	$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_0^{-1} + \mathbf{J}_y^{\mathsf{T}}\mathbf{J}_y$	varies	prior plus reading information	3.5.3
Gaussian conditional	$\boldsymbol{\Sigma}_{zz} - \boldsymbol{\Sigma}_{zy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yz}$	varies	Schur complement	3.5.3
Exceedance probability	$1 - \Phi\big((T_{\max} - \mu_{\text{post}})/\sigma_{\text{post}}\big)$	–	tail of the posterior	3.5.5
Law of total variance	$\mathbb{E}[\mathrm{Var}(K_\tau \mid \lambda)] + \mathrm{Var}(\mathbb{E}[K_\tau \mid \lambda])$	–	process plus rate terms	3.6.1
Posterior covariance	$\boldsymbol{\Sigma}$, $\mathrm{corr}(z_1, z_2) = \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$	varies	joint spread of the state	3.6.4
Variance budget	$\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}$	varies	sensitivity^2 $\times$ variance, plus cross terms	3.6.4
Mixture posterior	$\sum_j p(\mathbf{z} \mid \mathbf{y}, M_j)\Pr(M_j \mid \mathbf{y})$	varies	uncertainty over structure	3.6.6

3.8 Exercises

Exercise 3.1 (Formulation)**.** (a) In 2000 comparable hall-days of a peak-day log the load rose above the committed load on 100, an inlet excursion occurred on 60, and both on 25. Compute $\Pr(H)$, $\Pr(E)$, $\Pr(H \cap E)$, $\Pr(E \mid H)$ and $\Pr(H \mid E)$, name what the last two mean for the design, and state whether load above committed and excursion are independent. (0.05, 0.03, 0.0125, $25/100 = 0.25$, the fraction of above-committed days that brought an excursion, and $25/60 = 0.417$, the fraction of excursions that came on above-committed days, the rest being airflow, not load; not independent, since $0.0125 \neq 0.05 \times 0.03 = 0.0015$.) (b) Two halls driven by one site-wide workload have $\Pr(F_A) = 0.05$, $\Pr(F_B) = 0.04$ for an exceedance of the committed load on a day and $\Pr(F_A \cap F_B) = 0.015$. Compute the joint probability under an incorrect independence assumption, the probability that at least one hall stays within its committed load, and $\Pr(F_B \mid F_A)$, and state what it means for the shared plant. ($0.05 \times 0.04 = 0.002$; $1 - 0.015 = 0.985$, against 0.998 under independence; $0.015/0.05 = 0.30$; both halls are above their committed loads together 7.5 times as often as independence predicts, so a day on which the shared plant must carry both halls' excess at once is far more likely than the product suggests.)

Exercise 3.2 (Formulation)**.** A hall satisfies $T = T_c + Q/G$ with T_c known, $G > 0$, and both Q and G uncertain. Starting from $\Pr(T > T_{\max} \mid \mathbf{y})$, write an equivalent event in Q and G and express the probability as an integral over their joint posterior, without assuming independence. (Since $G > 0$, $T_c + Q/G > T_{\max} \iff Q > G(T_{\max} - T_c)$, so $\Pr(T > T_{\max} \mid \mathbf{y}) = \iint \mathbf{1}[Q > G(T_{\max} - T_c)]\, p(Q, G \mid \mathbf{y})\, \mathrm{d}Q\, \mathrm{d}G$; the joint posterior keeps the correlation between Q and G that a reading of T creates.)

Exercise 3.3 (Derivation)**.** (a) A fan-airflow flag monitors a cooling unit whose fan delivers less than its rated airflow with $\Pr(F) = 0.02$ per day, $\Pr(H \mid F) = 0.95$ and $\Pr(H \mid \neg F) = 0.05$. Derive the probability that the fan is short of its rating when the flag comes up, and explain why it is not 95 %. ($\Pr(H) = 0.95 \times 0.02 + 0.05 \times 0.98 = 0.068$; $\Pr(F \mid H) = 0.019/0.068 = 0.279$; 0.95 is $\Pr(H \mid F)$, not $\Pr(F \mid H)$, and the low base rate makes most flags false.) (b) With prior $\Pr(F) = 0.10$ that the unit is short of its rating, the capacity-shortfall flag A has $\Pr(H_A \mid F) = 0.9$, $\Pr(H_A \mid \neg F) = 0.1$, and the hall-temperature flag B has $\Pr(H_B \mid F) = 0.8$, $\Pr(H_B \mid \neg F) = 0.2$; the flags are conditionally independent given the unit's state. Derive $\Pr(F \mid H_A, H_B)$ and explain why the assumption is needed. ($\Pr(H_A, H_B \mid F) = 0.9 \times 0.8 = 0.72$ and $\Pr(H_A, H_B \mid \neg F) = 0.1 \times 0.2 = 0.02$, so $\Pr(F \mid H_A, H_B) = 0.072/(0.072 + 0.018) = 0.80$; sequentially, flag A gives 0.5 and flag B then 0.8; the likelihoods multiply only under conditional independence given each state of the unit, and a load above the committed load that raises both flags whatever the unit's state would invalidate it.)

Exercise 3.4 (Derivation)**.** Future exceedances of a hall's committed load satisfy $K_\tau \mid \lambda \sim \text{Poisson}(\lambda\tau)$ with $\lambda \mid \mathbf{y} \sim \text{Gamma}(3, 1000\,\text{h})$ (shape and rate) and $\tau = 200\,\text{h}$, the peak hours of about 16.7 design days. Derive the expected count, the predictive variance and its two contributions, and state what happens to each when a much longer exceedance log supporting a similar rate arrives. ($\mathbb{E}[K_\tau \mid \mathbf{y}] = 200 \times 0.003 = 0.6$; by equation (3.33), $\text{Var}(K_\tau \mid \mathbf{y}) = 0.6 + 200^2 \times 3/1000^2 = 0.6 + 0.12 = 0.72$; process 0.6, rate 0.12, a share $200/1200 = 0.167$. With a and b growing in proportion, the rate term $\tau^2(a/b)/b$ tends to zero while the process term stays near 0.6.)

Exercise 3.5 (Application)**.** A hall operates in mild weather with probability 0.90 per day and an excursion probability of 0.01, and in hot weather near the design day, with the chiller at its limit, with probability 0.10 and an excursion probability of 0.15. Derive the unconditional excursion probability of a day, and the probability that the weather was hot given that an excursion occurred. ($\Pr(E) = 0.90 \times 0.01 + 0.10 \times 0.15 = 0.024$; $\Pr(\text{hot} \mid E) = 0.015/0.024 = 0.625$.)

Exercise 3.6 (Application)**.** (a) A hall air temperature has prior $\mathcal{N}(25, 1.5^2)$, and a sensor with $\sigma_T = 0.5\,\mathrm{K}$ reads $y_T = 26°\mathrm{C}$. For the state, the one number T, write the two residuals and their slopes j_k, the information Λ, take the Newton step from the prior mean, and compute σ^2_{post} and $\Pr(T > 27°\mathrm{C} \mid y_T)$. ($r_1 = (T-25)/1.5$ and $r_2 = (T-26)/0.5$, $j_1 = 0.6667$ and $j_2 = 2\,\mathrm{K}^{-1}$, $\Lambda = 0.4444 + 4 = 4.4444\,\mathrm{K}^{-2}$; at $T = 25$, $r_1 = 0$ and $r_2 = -2$, so $\mathcal{L}' = j_1 r_1 + j_2 r_2 = -4$, $\Delta T = 4/4.4444 = 0.9$ and $\hat{T} = 25.90°\mathrm{C}$; $\sigma^2_{\mathrm{post}} = 1/4.4444 = 0.225\,\mathrm{K}^2$, $\sigma_{\mathrm{post}} = 0.4743\,\mathrm{K}$; $1 - \Phi(1.10/0.4743) = 1 - \Phi(2.319) = 0.0102$.) (b) For the two-entry hall of equation (1.10), whose Jacobian at the forward state (23°C, 0) has the rows (100, 500), (2, 0) and (0, 3.333), compute the three outer products $\mathbf{J}_k^{\mathsf{T}}\mathbf{J}_k$ and their sum, and state which row creates the off-diagonal entry and what the rank of its term is. (Physics $\begin{bmatrix} 10\,000 & 50\,000 \\ 50\,000 & 250\,000 \end{bmatrix}$, reading diag(4, 0), prior diag(0, 11.11); sum $\begin{bmatrix} 10\,004 & 50\,000 \\ 50\,000 & 250\,011 \end{bmatrix}$; only the physics row has entries in both columns, and its term has rank one, determinant $10^4 \times 2.5 \times 10^5 - (5 \times 10^4)^2 = 0$.)

Chapter 4

The Anatomy of Probabilistic Inference

4.1 From a plant graph to a design probability

At the design review that is to release the equipment hall for its 60 kW design load, the planner has the hall's commissioning record on the table: one cooling unit holding the supply air at 18°C, a test load of 30 kW, and one hall sensor reading 23.2°C. The commissioning test writes its acceptance limit at the design load as 28°C. The planner has that one measurement, the design sheet's value for the thermal resistance of the hall's air path and nothing at all from full load. The quick answer of the design sheet uses its handbook value: a resistance of 0.160 K/kW puts the hall at $18 + 60 \times 0.160 = 27.6$°C at full load, under the limit, so the hall can be released. The test gives a second quick answer: its rise of 5.2 K above the supply air, doubled with the load, puts the hall at 28.4°C, just above the limit. These are the commissioning record's own entries, not the handbook conductance of Chapter 1, and the design-load limit of this chapter is the acceptance limit of 28°C, a different limit from the 27.5°C of Chapter 1: this is the design review's version of that chapter's question.

The two quick answers disagree, and neither says how sure the planner can be: the handbook value ignores the reading, and the scaled reading ignores the handbook and the error of the sensor. What neither settles is what one reading at the test condition, the handbook value and the physics of the hall together say about a condition that has never been run. The question of this chapter is

how a posterior distribution over the state of a plant graph is obtained from physical relations, readings and prior information, and how that posterior is turned into the probability that answers a planner's question.

Chapter 1 built the plant graph, Chapter 2 classified the questions and Chapter 3 derived the rules one at a time; here they are used together for the first time, on a state of three entries of which the reading touches only one.

Jargon: calibration (state estimation), maximum a posteriori (MAP) estimate, Hessian, Laplace approximation, delta method, change of variables, credible interval, scenario node, prior and posterior predictive distributions, marginal likelihood, Bayes factor, rank-one update, reduced state

The *calibration* of a plant model (its *state estimation*) is the fit of the state of the plant to its measurements and priors; its results are the MAP state and the posterior covariance defined next. The *maximum a posteriori (MAP) estimate* is the value of the unknowns at which the posterior density is highest; it is found by minimizing the negative logarithm of the posterior, the *objective*. The *Hessian* is the matrix of second derivatives of the objective; at the MAP estimate it measures the curvature of the objective, and for rows linear in the state and divided by their standard deviations it is $\mathbf{J}^\mathsf{T}\mathbf{J}$, the information matrix. The *Laplace approximation* replaces the posterior by the Gaussian whose mean is the MAP estimate and whose covariance is the inverse of the Hessian there. The *delta method* approximates the variance of a function of the unknowns by linearizing the function at the estimate. A *change of variables* carries a density from one coordinate to another, multiplying it by the derivative of the old coordinate with respect to the new so that each interval keeps its probability. A *credible interval* contains the unknown with a stated posterior probability. A *scenario node* (a what-if) is a quantity of the plant at an input value it has not yet seen, such as the hall temperature at a load never carried, obtained by solving the plant at that value and carrying the posterior to it. The *prior predictive* and *posterior predictive* distributions are the distributions of a reading not yet seen, averaged over the prior and over the posterior respectively. The *marginal likelihood*, or *evidence*, of a model is the probability density of the observed readings averaged over the prior; the *Bayes factor* is the ratio of the evidences of two models. A *rank-one update* adds one new reading to an existing Gaussian posterior by a closed-form correction of its mean and covariance, without solving the problem again. The *reduced state* is the vector of free entries that remain when the physical rows are taken as exact; every other entry of the state is then a function of it.

Key idea

A design probability is computed in two operations. Inference, the calibration of the plant, turns the plant graph, the measurements and the priors into a posterior over the state: its mean is the MAP state and its covariance is the inverse curvature of the objective. Query evaluation turns that posterior into the planner's answer: an entry of the state and its variance, or a derived quantity $f(\mathbf{z})$ with variance $\mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g}$, and then a normal tail at the limit.

4.1.1 The complete inference workflow

The method fixed in Chapter 2 (Figure 2.6) is a progression: the plant graph (step 0), then the questions written over its state (step 1), the graph extended for them with the evidence attached (step 2), the posterior of the state that the rows give, derived and computed once (step 3), the answers read from that posterior (step 4), their checks (step 5), the report with its bands (step 6), and the next measurement when a band is too wide (step 7). Figure 4.1 draws the progression. The wide box at the top is the plant graph, built once; the seven boxes beneath are one pass that carries every question, read left to right along the top row and right to left along the bottom row; the second line of each box names what the step produces. The dashed arrow from step 7 back to step 3 is the loop taken when a new measurement is made: its row is added and the posterior is rebuilt.

The distinction between steps 3 and 4 carries the method, and it is the distinction between the two operations of every inference. Step 3 is *inference*: the rows of step 2 are combined into one posterior of the state, its MAP state, information and covariance, derived and then

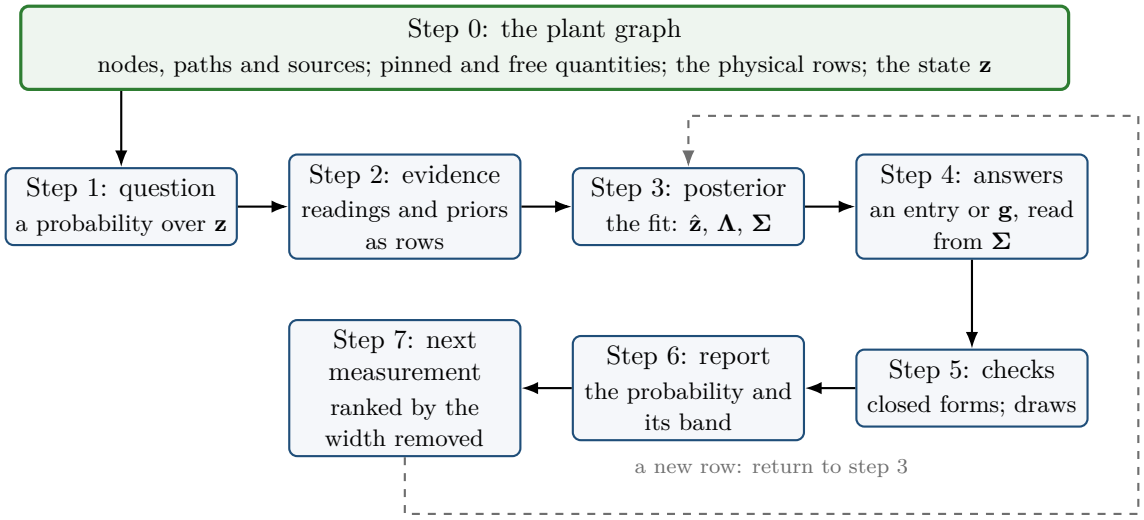


Figure 4.1: The inference workflow. Step 0 is built once; steps 1 to 7 are one pass, carrying every question of the plant. The second line of each box names what the step produces.

computed for the plant; it does not depend on which question is asked. Step 4 is *query evaluation*: each question is reduced, from the planner's statement through its general and specialized forms, to quantities read from that posterior, in this chapter the posterior mean and variance of one state entry, the full-load temperature, and the standard normal distribution function evaluated at the limit. Keeping them apart is what lets one posterior serve every question, and a new measurement change step 3 alone.

4.1.2 The posterior and the questions put to it

The calculation divides into two operations of different kinds:

$$\underbrace{(\text{graph},\ \mathbf{y}) \longrightarrow p(\mathbf{z} \mid \mathbf{y})}_{\text{inference}}, \qquad \underbrace{p(\mathbf{z} \mid \mathbf{y}) \longrightarrow \text{design probability or decision}}_{\text{query evaluation}}, \tag{4.1}$$

the second applied to the result of the first. Inference is done once for a given graph and a given body of evidence. Under a Gaussian posterior, exact or approximate, its result is the triple of the MAP state $\hat{\mathbf{z}}$, the posterior covariance $\boldsymbol{\Sigma}$, and the Jacobian of the model at $\hat{\mathbf{z}}$, from which the gradient $\mathbf{g}$ of any derived quantity is formed. Query evaluation is then a formula on the triple. An output that is itself an entry z_i of the state, such as a temperature at a node, has mean $\hat{z}_i$ and variance Σ_{ii}, read straight off the posterior. An output that is a function $f(\mathbf{z})$ of several entries needs its gradient and the quadratic form $\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}$. A probability that the output stays below a limit is a normal tail on either pair.

Two kinds of further work follow from this division and must be kept apart. A new question asked of the same evidence needs no new inference: a new function of the state is pushed through the existing $\boldsymbol{\Sigma}$. New evidence changes the posterior itself: either the inference is repeated with the new row, or the existing Gaussian is conditioned on the new row by a rank-one update, which Section 4.6.4 derives. The posterior represents what is known about the plant under the model; the queries represent what the planner wants to determine. A what-if at an input the plant has not seen is such a query: Section 4.3.8 reads it from the same posterior.

4.1.3 The inference in the terms of graphical models

The operations of this book are those of inference in a probabilistic graphical model, under the names a design office uses. The plant graph with its rows is a factor graph (Section 1.1.3): each

row is a factor over the entries it touches, and the posterior $e^{-\mathcal{L}(\mathbf{z})}$ is the normalized product of the factors. Table 4.1 sets each operation beside its name in the literature on graphical models. Its first column is the operation as this book writes it, the second the standard name, and the third the chapters that derive it.

Table 4.1: The operations of the book and their names in the literature on probabilistic graphical models.

Operation in this book	Name in graphical models	Derived in
a row: prior, physics or reading	a factor of a factor graph	Chapters 1, 3
the posterior $e^{-\mathcal{L}(\mathbf{z})}$	the normalized product of the factors	Chapter 3
the normal equations, $\mathbf{\Lambda} = \mathbf{J}^\mathsf{T}\mathbf{J}$, $\mathbf{\Sigma} = \mathbf{\Lambda}^{-1}$	exact inference in a linear-Gaussian factor graph, in information form	Chapter 3; this chapter
the gamma updates of an exceedance rate	exact inference with conjugate factors	Chapters 2, 3
Gauss–Newton and the Laplace covariance	the Gaussian approximation at the mode of a nonlinear factor graph (factor-graph optimization)	this chapter; Chapter 5
the Schur complement; the marginal of a joint Gaussian	variable elimination of the entries not asked about	Chapter 3
the enumeration of units up and down; total probability	exact inference over discrete states by summation	Chapter 9
the evidence $p(\mathbf{y} \mid M)$	the normalizing constant (partition function) of the product	Chapter 2

The general algorithms of that literature are not needed for the plants of this book, and the table shows why: every posterior here is solved jointly, by factoring the information matrix of the whole state, which is exact for linear-Gaussian rows and costs little at a few dozen entries. Message passing, belief propagation (the sum-product algorithm) and the junction tree compute the same marginals by local operations on the graph, and they earn their place on graphs too large to factor jointly or with discrete entries of many levels; on a Gaussian graph without loops, belief propagation returns exactly the Schur complements of the joint factorization. Factoring a large sparse information matrix is itself variable elimination, and its cost is set by the order in which the entries are eliminated: eliminated in time order, a state indexed by time, such as the monthly windows of an operating record, gives the Kalman filter and smoother. Markov chain Monte Carlo (Gibbs sampling and the Metropolis algorithm) draws from a posterior that the Laplace normal cannot approximate and reweighting cannot correct: one with several modes, or the funnel of a hierarchical model, in which the spread of a population of rates is itself unknown; a later volume of *Reliability Engineering by Derivation* meets it in a fleet of pumps. Variational inference is not used in this book.

These algorithms differ in what they give up. Variable elimination, the junction tree, belief propagation on a graph without loops and the Kalman smoother on linear-Gaussian rows are exact: they compute the posterior the joint factorization computes, in another order. The Laplace approximation, loopy belief propagation, the linearized Kalman filters and variational inference are approximate, and their error is a bias that more computation does not remove; loopy belief propagation and variational inference in particular tend to report variances too small, which for a tail probability against a limit is the error that matters most. Importance sampling and Markov chain Monte Carlo are exact in the limit of many draws, and their error is random and can be estimated, although a chain that never visits a second mode can look converged. The two kinds of error check each other, which is why the draws serve as the other route against the Laplace answer in step 5. The route also decides which objects come cheaply. The joint factorization gives $\hat{\mathbf{z}}$, $\mathbf{\Sigma}$, the gradient and the preposterior covariance $\mathbf{\Sigma}_a$ of a rank-one

update in one pass, and the evidence in closed form under the Laplace approximation; draws give the posterior as a sample, from which a variance is an estimate, a preposterior needs the candidate's outcomes simulated and the posterior refitted or reweighted, and the evidence needs a further computation, since the draws do not carry the normalizing constant. The method is unchanged by the choice: the plant graph, the questions, the rows and the objective $\mathcal{L}(\mathbf{z})$ are the same for every algorithm, the general forms of Chapter 2 are expectations over the posterior whatever computes it, and only the computations of steps 3 and 4, and the checks of step 5 that test them, depend on the route.

4.1.4 The plant graph under two operating conditions (step 0)

The plant is an equipment hall with a heat-generating load and a single effective heat-removal path to the supply air, which the cooling unit's discharge controller holds at $T_c = 18°\mathrm{C}$. Two operating conditions are considered: a test condition with a heat load of $Q_1 = 30\,\mathrm{kW}$, and the full design load $Q_2 = 60\,\mathrm{kW}$. At steady state the heat supplied to the hall leaves it through the path, and energy conservation at the hall node reads

$$Q = G\,(T - T_c), \tag{4.2}$$

where T is the hall air temperature in °C, Q the heat load in kW, and G the effective thermal conductance of the air path in kW/K. The supply air is pinned: the controller holds it at its set point, which is what makes T_c a boundary of the graph rather than an unknown, and it fails if the unit runs out of capacity and the discharge drifts up, which Section 4.6.6 takes up. The balance is taken at steady state through one effective path: this reduces the cooling of the hall to a single unknown, and it fails in the minutes after a load change, while heat stored in the hall and its equipment is still moving, which Section 4.6.6 and Exercise 4.6 take up. It is convenient to work with the thermal resistance of the path,

$$\rho = \frac{1}{G}, \tag{4.3}$$

in K/kW, with which equation (4.2) becomes

$$T = T_c + Q\rho. \tag{4.4}$$

Carrying the resistance in the state, rather than the conductance, is a choice: the temperature is then a linear function of the state, which makes every row of the example linear, the posterior exactly Gaussian and every result below exact rather than approximate, and it fails only where a Gaussian on the resistance puts real probability below zero, which Section 4.1.7 measures for the handbook prior and Section 4.6.6 discusses. Chapter 1 carries the conductance in the state as the log-conductance $g = \ln(G/G_{\mathrm{ref}})$, which keeps it positive but makes the temperature nonlinear in the state; the cooling units of hall B in Chapter 5 carry their coil conductances the same way, as log-derates.

At the two operating conditions equation (4.4) gives

$$T_1 = 18 + 30\rho, \tag{4.5}$$

$$T_2 = 18 + 60\rho, \tag{4.6}$$

with T_1 the hall temperature during the test and T_2 the temperature at full load, both in °C with ρ in K/kW. Figure 4.2 draws the graph. Each row is one operating condition: the box on the left is the heat source, the green box in the middle the hall node carrying a free temperature, and the gray box on the right the supply-air node, pinned at 18°C. The solid arrows are the physical paths, the heat entering the hall and the heat leaving it through the resistance. The green node between the rows is the resistance ρ, free and shared by both paths,

as the dashed links show. The dashed arrows leave the graph for the evidence: the reading of the test temperature and the handbook prior on ρ, attached in Section 4.1.7. The full-load node carries no reading; it is the scenario node of a what-if (contingency) analysis that the question of Section 4.1.6 requires, drawn here with the rest of the graph so that the figure shows the graph as extended for that question.

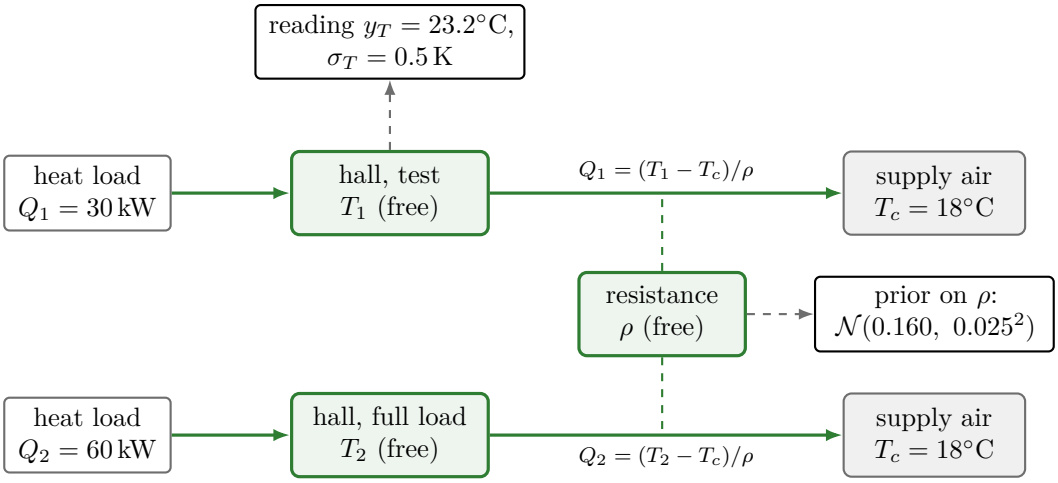


Figure 4.2: The plant graph of the equipment hall under its two operating conditions. Free: the resistance ρ and the temperatures T_1 and T_2 (green). Pinned: the supply-air temperature T_c, held by the cooling unit's discharge controller, and the loads Q_1, Q_2 (gray and white). Observed: T_1, by one reading (dashed arrow, top); ρ carries the handbook prior (dashed arrow, right). The full-load node T_2 is the scenario node of the question and has no reading; it is linked to the test only through the shared resistance.

The resistance is taken to be the same at 60 kW as at 30 kW, one free entry shared by both rows: this is what lets a reading taken at the test condition carry information about the full-load condition, although the hall has never been operated at full load, and it fails if the path changes with load, for example when the unit's fans or its coil change state, which one test reading cannot detect (Section 4.6.6) and a reading at full load would test (Section 4.6.3). The link between the two rows of Figure 4.2 is not a device of the calculation; it is the physical statement that both conditions are the same hall with the same heat-removal path, and the prediction at full load rests on it.

4.1.5 The state vector and the physical rows

The free quantities of the graph form the state vector,

$$\mathbf{z} = \begin{bmatrix} \rho \\ T_1 \\ T_2 \end{bmatrix}, \tag{4.7}$$

with entries in K/kW, °C and °C. The physical relations are written as rows that vanish when they hold:

$$r_1(\mathbf{z}) = T_1 - 18 - 30\rho = 0, \tag{4.8}$$
$$r_2(\mathbf{z}) = T_2 - 18 - 60\rho = 0. \tag{4.9}$$

These are the conservation rows of the graph, one for each hall node. The balance is taken to hold at both conditions up to a small model tolerance σ_p, a standard deviation that represents

model discrepancy, and the answers are reported in the limit $\sigma_p \to 0$ of exact balances: the finite tolerance keeps every matrix of the calculation invertible while the reported numbers are those of an exact balance, and the limit understates the spread if the real discrepancy is finite, which Section 4.4.5 measures over three tolerances and Section 4.6.6 discusses. Section 4.2.3 puts the rows, with their tolerance, into one residual vector with the reading and the prior.

The three entries of the state are therefore not three independent unknowns. The two physical rows leave one direction of the state free, a change of ρ with both temperatures following it; in the limit of exact rows the state has three entries but one degree of freedom, and this fact governs the shape of the posterior covariance derived in Section 4.2.5 and the reduced state of Section 4.2.8.

4.1.6 The planner's questions (step 1)

The release for the design load raises the heat load from 30 kW to 60 kW. The maximum acceptable hall temperature, the acceptance limit of the commissioning test, is $T_{\max} = 28°\mathrm{C}$. In the planner's words the question is: given the hall sensor's reading and what is known of the air path, what is the probability that the hall stays below 28°C at full load? As a probability over the state, with the unknown on the left of the bar and everything treated as known on the right,

$$\Pr(T_2 < T_{\max} \mid \mathbf{y}, M), \tag{4.10}$$

where T_2 is the full-load entry of the state, $\mathbf{y}$ the evidence (the test reading and the handbook prior), and M the assumed steady-state thermal model, equation (4.4) with the rows (4.8)–(4.9). The model is suppressed from the notation below when it is not in question. This is a question of the category *prediction under a threshold*: a tail probability of one marginal of the posterior.

Three estimation questions are asked of the same evidence, one for each entry of the state: the thermal resistance of the air path, the hall temperature at the test condition, and the hall temperature at full load. Table 4.2 lists the five questions: the first column is the question in the planner's words; the second the general form that every question of its category takes, written for a quantity $f(\mathbf{z})$ derived from the state and a threshold f^* on it; the third its mathematical form for the hall; the fourth its category; and the last the subsection where it is answered. Which further measurement would narrow the answers most asks nothing new of the present evidence: it is the next measurement for these questions, answered in Section 4.6.

Table 4.2: The questions asked of the hall's test reading, the general form of their category, their mathematical form for the hall, their category and where each is answered. $\mathbf{y}$ is the test reading and the handbook prior; $T_{\max} = 28°\mathrm{C}$.

Planner's question	General form	Mathematical form	Category	Answered in
What is the thermal resistance of the hall's air path?	$\Pr(f(\mathbf{z}) \le f^* \mid \mathbf{y})$, moments	$p(\rho \mid \mathbf{y})$: $\hat{z}_1$, $\sqrt{\Sigma_{11}}$	estimation	4.3.6
How warm is the hall at the 30 kW test?	$\Pr(f(\mathbf{z}) \le f^* \mid \mathbf{y})$, moments	$p(T_1 \mid \mathbf{y})$: $\hat{z}_2$, $\sqrt{\Sigma_{22}}$	estimation	4.3.6
How warm will the hall be at 60 kW?	$\Pr(f(\mathbf{z}) \le f^* \mid \mathbf{y})$, moments, quantiles	$p(T_2 \mid \mathbf{y})$: $\hat{z}_3$, $\sqrt{\Sigma_{33}}$, credible interval	estimation	4.3.6
Can the hall take 60 kW and stay below 28°C?	$\Pr(f(\mathbf{z}) \ge f^* \mid \mathbf{y})$	$\Pr(T_2 < T_{\max} \mid \mathbf{y})$, with $f = -T_2$	prediction under a threshold	4.3.7
Would the hall stay below 28°C at a load it has not yet carried?	$\Pr(f(\mathbf{z}) \ge f^* \mid \mathbf{y})$ at a scenario node	$\Pr(T^*(Q^*) \le T_{\max} \mid \mathbf{y})$, $T^* = T_c + Q^* z_1$, with $f = -T^*$	prediction under a threshold	4.3.8

Because the full-load temperature is carried as an entry of the state, the threshold question needs exactly one marginal of the posterior: the mean $\hat{z}_3$ and the variance Σ_{33} of the third entry: the question is written over the state, and the position of T_2 in the state shows which part of the posterior it needs. Each estimation question decomposes in the same way, into one entry of $\hat{\mathbf{z}}$ and one diagonal entry of $\boldsymbol{\Sigma}$.

4.1.7 The evidence as rows (step 2)

The temperature reading. During the 30 kW test the hall sensor reports $y_T = 23.2°\text{C}$, with a measurement error of standard deviation $\sigma_T = 0.5\,\text{K}$. The observation model writes the random reading Y as the true test temperature, the second entry of the state, plus an error,

$$Y = T_1 + \epsilon = \mathbf{h}^\top \mathbf{z} + \epsilon, \qquad \mathbf{h} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \epsilon \sim \mathcal{N}(0, \sigma_T^2), \tag{4.11}$$

where $\mathbf{h}$ selects the entry the sensor measures. The error is taken as Gaussian, unbiased, of known spread and independent of the state: this makes the reading one linear row of fixed weight $1/\sigma_T^2$, and it fails if the sensor is biased, which would shift every inferred quantity; the prior predictive check of Section 4.4.2 would flag a large bias, and the comparison of models of Section 4.2.2 is the test built for it. The reading lands on T_1 only; it is the physical row of the test hall, not a substitution, that will carry it to ρ.

Prior information. The design sheet gives a preliminary value of the resistance, the handbook figure for an air path of this kind, represented by the Gaussian prior

$$\rho \sim \mathcal{N}(\mu_0, \sigma_0^2), \qquad \mu_0 = 0.160\,\text{K/kW}, \quad \sigma_0 = 0.025\,\text{K/kW}. \tag{4.12}$$

The prior mean corresponds to a conductance $G_0 = 1/0.160 = 6.25\,\text{kW/K}$. Taking the handbook value as the prior for this hall is a choice: it is what places the full-load temperature at 27.6°C before any reading and what keeps the estimate from following the one reading alone, and it misleads if the handbook describes air paths unlike this one, a conflict that the prior predictive check of Section 4.4.2 would show as a reading far from 22.8°C. A Gaussian prior on a resistance assigns some probability to negative, nonphysical values; here it is $\Phi(-0.160/0.025) = \Phi(-6.4) \approx 8 \times 10^{-11}$, negligible, and the Gaussian is kept for its analytical simplicity. A production model enforces positivity through a logarithmic parameterization, such as the log-conductance g of Chapter 1, or a prior with positive support.

The rows and the entries they act on. The physics and the evidence put four rows on the state: the physical rows of the test hall and of the full-load hall, equations (4.8)–(4.9), the reading row on T_1, and the prior row on ρ. Which row acts on which entry is a property of the graph, and it fixes where the Jacobian of the rows can be nonzero, marked $*$:

$$\mathbf{J} = \begin{bmatrix} * & * & 0 \\ * & 0 & * \\ 0 & * & 0 \\ * & 0 & 0 \end{bmatrix} \begin{matrix} \text{physics, test hall} \\ \text{physics, full-load hall} \\ \text{reading} \\ \text{prior} \end{matrix}, \qquad \text{columns } \rho, T_1, T_2. \tag{4.13}$$

A line of the pattern lists the entries one row sees, and a column lists the rows that see one entry. The third column has a single mark: the full-load temperature is seen by its own physical row and by nothing else. The second column shows that the reading and the test hall's balance are the two rows that see T_1, and the first that the two balances and the prior see ρ. No row joins the reading to T_2; the two physical rows, through the column ρ they share, are the only path from the one to the other.

Knowns and unknowns. Table 4.3 lists every quantity of the problem. The first column is the symbol, the second its value, and the third its role in the graph: pinned boundary condition, evidence, or free entry of the state. The graph and the probabilistic model are now specified completely enough for the inference to be carried out.

Table 4.3: Knowns and unknowns of the equipment-hall example.

Quantity	Value	Role
T_c [°C]	18	pinned: supply air, held by the discharge controller
Q_1 [kW]	30	pinned: test load
Q_2 [kW]	60	pinned: design load
y_T [°C]	23.2	evidence: the test reading
σ_T [K]	0.5	evidence: standard deviation of the reading error
μ_0 [K/kW]	0.160	evidence: prior mean of the resistance
σ_0 [K/kW]	0.025	evidence: prior standard deviation of the resistance
σ_p [K]	0.01, then $\to 0$	model: tolerance of the two physical rows
ρ [K/kW]	unknown	free: thermal resistance, the first entry of the state
T_1, T_2 [°C]	unknown	free: test and full-load temperatures, tied to ρ by the physical rows

Remark 4.1. Four spreads appear in what follows and must not be confused. The reading error has standard deviation σ_T (no index), a property of the sensor. The prior on the resistance has standard deviation σ_0. The physical rows have the model tolerance σ_p, a property of the steady-state model. The posterior standard deviation of the i-th entry of the state is $\sqrt{\Sigma_{ii}}$, the square root of a diagonal entry of the posterior covariance, written $\mathrm{sd}(T_2 \mid \mathbf{y})$ when the entry is named.

4.2 The posterior of the state (step 3)

Every question of Table 4.2 is a property of one density, the posterior of the state given the reading and the handbook prior. Each row of Section 4.1.7 contributes one factor to it, and the MAP state, the information matrix and the covariance are properties of that product. This section derives the posterior and computes it for the hall; Section 4.3 reads every answer from it.

4.2.1 The likelihood of the reading and the physical rows

The likelihood states how probable the observed reading would be under a proposed state. By equation (4.11), for a fixed state the reading is the test temperature T_1 plus a Gaussian error, so $Y \mid \mathbf{z} \sim \mathcal{N}(T_1,\ \sigma_T^2)$, and its density at the observed value is

$$p(y_T \mid \mathbf{z}) = \frac{1}{\sqrt{2\pi\sigma_T^2}} \exp\left[-\frac{(T_1 - y_T)^2}{2\sigma_T^2}\right]. \tag{4.14}$$

For the reading $y_T = 23.2\,^{\circ}\mathrm{C}$ and $\sigma_T = 0.5\,\mathrm{K}$, so that $2\sigma_T^2 = 0.5\,\mathrm{K}^2$,

$$p(y_T \mid \mathbf{z}) \propto \exp\left[-\frac{(T_1 - 23.2)^2}{0.5}\right]. \tag{4.15}$$

The likelihood is a function of the state with the reading held fixed; it is not a density in $\mathbf{z}$ and need not integrate to one over it. It depends on the first temperature only and is flat in ρ and T_2: the sensor measures a temperature and says nothing directly about the resistance or about the full-load hall.

What connects the three entries is the pair of physical rows. A steady-state balance that holds up to a model tolerance σ_p (a standard deviation in kelvin, the amount by which the balance of equation (4.4) is allowed to fail) is the density

$$p(T_1 \mid \rho) = \mathcal{N}(T_1;\ 18 + 30\rho,\ \sigma_p^2), \qquad p(T_2 \mid \rho) = \mathcal{N}(T_2;\ 18 + 60\rho,\ \sigma_p^2), \tag{4.16}$$

where $\mathcal{N}(x; m, v)$ denotes the Gaussian density with mean m and variance v evaluated at x. The exact rows (4.8)–(4.9) are the limit $\sigma_p \to 0$, in which each density concentrates on its row. The derivation below keeps σ_p finite, carries all three entries, and takes the limit at the end, where it is checked numerically; no row is used to eliminate an entry.

4.2.2 The prior density and Bayes' theorem

By equation (4.12) the prior density of the resistance is

$$p(\rho) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{(\rho - \mu_0)^2}{2\sigma_0^2}\right] \propto \exp\left[-\frac{(\rho - 0.160)^2}{2(0.025)^2}\right]. \tag{4.17}$$

The prior of the state is the product of this density with the two physical densities (4.16), $p(\mathbf{z}) = p(T_1 \mid \rho)\, p(T_2 \mid \rho)\, p(\rho)$, the factorization of Chapter 3 read along the graph: each hall temperature depends on the resistance, and on nothing else. Bayes' theorem, derived in that chapter, gives the posterior of the state as

$$p(\mathbf{z} \mid y_T) = \frac{p(y_T \mid \mathbf{z})\, p(\mathbf{z})}{p(y_T)}, \qquad p(y_T) = \int p(y_T \mid \mathbf{u})\, p(\mathbf{u})\, \mathrm{d}\mathbf{u},$$

with $\mathbf{u}$ a variable of integration over the state. The denominator does not depend on $\mathbf{z}$, so

$$p(\mathbf{z} \mid y_T) \propto p(y_T \mid \mathbf{z})\, p(\mathbf{z}).$$

The evidence of the model. The denominator has a name of its own. Writing the model M explicitly, it is the evidence, or marginal likelihood, of the model,

$$p(y_T \mid M) = \int p(y_T \mid \mathbf{z}, M)\, p(\mathbf{z} \mid M)\, \mathrm{d}\mathbf{z}\,: \tag{4.18}$$

the density of the observed reading averaged over the prior of the state. For the linear-Gaussian model it is the prior predictive density of the reading evaluated at the observed value (Section 4.4.2). It also follows from the objective and its curvature, which is how it is computed when no closed form exists. The integrand is $\exp[-\mathcal{L}(\mathbf{z})]$, with the objective $\mathcal{L}$ of equation (4.24), times the normalizing constants $(2\pi\sigma_k^2)^{-1/2}$ of the four rows; when the objective is quadratic about its minimum $\hat{\mathbf{z}}$ with matrix $\mathbf{J}^\mathsf{T}\mathbf{J}$, whose inverse is the covariance $\boldsymbol{\Sigma}$ of Section 4.2.5, the Gaussian integral over the three entries gives

$$p(y_T \mid M) = \frac{(2\pi)^{3/2}\, \det(\boldsymbol{\Sigma})^{1/2}}{(2\pi)^2\, \sigma_p^2\, \sigma_T\, \sigma_0}\, \mathrm{e}^{-\mathcal{L}(\hat{\mathbf{z}})}. \tag{4.19}$$

The two directions of the state that the physical rows fix have posterior widths proportional to σ_p, and they cancel the two physical rows' normalizing constants: $\det(\boldsymbol{\Sigma})^{1/2}/\sigma_p^2$ tends to the width of the posterior along the free direction. The result is the volume of the posterior,

$\det(\boldsymbol{\Sigma})^{1/2}$, times the height of the integrand at the MAP state. The formula is exact when the objective is quadratic, as it is here; in general it is the Laplace approximation to the evidence.

The evidence normalizes the posterior, and its second use is to compare models. Alternatives for the hall include M_1, a single thermal resistance; M_2, two distinct heat-removal paths; and M_3, a model with a biased hall sensor. Each gives its own evidence for the same reading, and their ratio is a Bayes factor,

$$B_{21} = \frac{p(y_T \mid M_2)}{p(y_T \mid M_1)}, \tag{4.20}$$

which multiplies the prior odds of the two models to give their posterior odds, $\Pr(M_2 \mid y_T)/\Pr(M_1 \mid y_T) = B_{21} \cdot \Pr(M_2)/\Pr(M_1)$. That comparison is the category *hypothesis comparison*, whose general form Chapter 2 states.

The posterior up to its constant. The data $\mathbf{y}$, the evidence in the sense of what is known about the hall, consists of the reading and the handbook prior; the likelihood involves only the reading, and the prior enters through $p(\mathbf{z})$, so $p(\mathbf{z} \mid \mathbf{y})$ and $p(\mathbf{z} \mid y_T)$ name the same density. Substituting equations (4.14), (4.16) and (4.17) and collecting the four exponents,

$$p(\mathbf{z} \mid \mathbf{y}) \propto \exp\left[-\frac{1}{2}\left(\frac{(T_1 - 18 - 30\rho)^2}{\sigma_p^2} + \frac{(T_2 - 18 - 60\rho)^2}{\sigma_p^2} + \frac{(T_1 - y_T)^2}{\sigma_T^2} + \frac{(\rho - \mu_0)^2}{\sigma_0^2}\right)\right]. \tag{4.21}$$

This is the posterior of the state up to its normalizing constant. It combines the physical relations, the reading, the reading's uncertainty and the prior knowledge in one probability model, one term per row. Every later result of the chapter is a property of this expression.

4.2.3 The residual vector, the Jacobian and the information matrix

The maximum a posteriori estimate is the state at which the posterior density is greatest,

$$\hat{\mathbf{z}} = \arg\max_{\mathbf{z}} \; p(\mathbf{z} \mid \mathbf{y}). \tag{4.22}$$

The logarithm is increasing, so maximizing the posterior is equivalent to minimizing its negative logarithm. Dropping the terms that do not depend on $\mathbf{z}$ (the normalizing constants of the four densities and $\ln p(y_T)$) leaves half the sum of the squared exponents of (4.21). Each exponent is a row divided by its standard deviation, and the four rows stack into one residual vector,

$$\mathbf{r}(\mathbf{z}) = \begin{bmatrix} (T_1 - 18 - 30\rho)/\sigma_p \\ (T_2 - 18 - 60\rho)/\sigma_p \\ (T_1 - y_T)/\sigma_T \\ (\rho - \mu_0)/\sigma_0 \end{bmatrix} \begin{matrix} \text{physics, test hall} \\ \text{physics, full-load hall} \\ \text{reading} \\ \text{prior} \end{matrix} \tag{4.23}$$

whose entries are dimensionless, so that the objective is

$$\mathcal{L}(\mathbf{z}) = \tfrac{1}{2}\,\mathbf{r}(\mathbf{z})^{\mathsf{T}}\mathbf{r}(\mathbf{z}), \qquad p(\mathbf{z} \mid \mathbf{y}) \propto \exp[-\mathcal{L}(\mathbf{z})]. \tag{4.24}$$

In the notation of the book it is $\mathcal{L} = \frac{1}{2}\sum_k w_k r_k^2$ over the unnormalized rows r_1, r_2 of equations (4.8)–(4.9), $T_1 - y_T$ and $\rho - \mu_0$, with weights $w_k = 1/\sigma_k^2$; the normalized rows carry their weights inside them. The first two terms penalize a state that unbalances a hall, in units of the model tolerance; the third, a state that disagrees with the reading, in units of the sensor spread; the fourth, a resistance far from its handbook value, in units of the prior spread. The weight of a physical row has a statistical meaning: $1/\sigma_p^2$ is the inverse variance of a model discrepancy, the amount by which the steady-state balance may fail.

The Jacobian. The derivatives of the four rows with respect to the three entries (ρ, T_1, T_2) form the 4×3 Jacobian $\mathbf{J} = \partial\mathbf{r}/\partial\mathbf{z}$, one line per row and one column per entry:

$$\mathbf{J} = \begin{bmatrix} -30/\sigma_p & 1/\sigma_p & 0 \\ -60/\sigma_p & 0 & 1/\sigma_p \\ 0 & 1/\sigma_T & 0 \\ 1/\sigma_0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} -3000 & 100 & 0 \\ -6000 & 0 & 100 \\ 0 & 2 & 0 \\ 40 & 0 & 0 \end{bmatrix} \quad (\sigma_p = 0.01\,\mathrm{K}), \tag{4.25}$$

with the first column in kW/K and the other two in K^{-1}. The entry J_{ki} is the change in row k, counted in units of that row's standard deviation, per unit change of entry i. Its nonzero entries sit where the pattern (4.13) of the graph places them. The entries $2 = 1/0.5$ and $40 = 1/0.025$ are the normalizing factors of the reading and the prior. Every row is linear in the state, so $\mathbf{J}$ is the same at every linearization point. The tolerance $\sigma_p = 0.01\,\mathrm{K}$, one-fiftieth of the sensor's spread, is tight enough that the answers no longer move (the limit is checked in Section 4.4.5) and loose enough, within two decades of σ_T, to keep the normal equations conditioned well enough for double precision.

The information matrix. For rows normalized by their standard deviations the information matrix is $\mathbf{\Lambda} = \mathbf{J}^\mathsf{T}\mathbf{J}$, equation (3.26). Written with the unnormalized rows, whose Jacobian is $\tilde{\mathbf{J}}$, and the diagonal weight matrix $\mathbf{W} = \mathrm{diag}(1/\sigma_p^2,\ 1/\sigma_p^2,\ 1/\sigma_T^2,\ 1/\sigma_0^2)$, it is the same matrix in the form used throughout the book,

$$\mathbf{\Lambda} = \mathbf{J}^\mathsf{T}\mathbf{J} = \tilde{\mathbf{J}}^\mathsf{T}\mathbf{W}\tilde{\mathbf{J}} = \mathbf{J}_\mathrm{d}^\mathsf{T}\mathbf{W}_\mathrm{d}\mathbf{J}_\mathrm{d} + \mathbf{\Lambda}_\mathrm{prior}, \tag{4.26}$$

where the last form keeps the prior rows apart: $\mathbf{J}_\mathrm{d}$ and $\mathbf{W}_\mathrm{d}$ hold the physical and reading rows, and $\mathbf{\Lambda}_\mathrm{prior}$, here $\mathrm{diag}(1/\sigma_0^2, 0, 0)$, is the prior precision. For rows that are nonlinear in the state, equation (4.26) is the Gauss–Newton approximation to the Hessian of $\mathcal{L}$: it drops the curvature term $\sum_k r_k\, \partial^2 r_k/\partial\mathbf{z}\,\partial\mathbf{z}^\mathsf{T}$, the second derivatives of the rows weighted by their residuals, which vanishes for linear rows. Writing $\mathbf{j}_k$ for the k-th line of $\mathbf{J}$, the product is a sum of one outer product per row, so the information of independent rows adds, as equation (3.27) states for priors and readings; with the numbers of equation (4.25),

$$\mathbf{J}^\mathsf{T}\mathbf{J} = \sum_{k=1}^{4} \mathbf{j}_k^\mathsf{T}\mathbf{j}_k = \begin{bmatrix} 45\,001\,600 & -300\,000 & -600\,000 \\ -300\,000 & 10\,004 & 0 \\ -600\,000 & 0 & 10\,000 \end{bmatrix}. \tag{4.27}$$

The entry in the corner is $3000^2 + 6000^2 + 40^2$: the two physical rows and the prior. The reading contributes only the $4 = 1/\sigma_T^2$ of the (T_1, T_1) entry. The reading and prior rows alone would give $\mathrm{diag}(1600, 4, 0)$: no information at all would reach T_2, and none would pass from the reading to ρ. It is the off-diagonal entries of the physical rows that carry the reading at T_1 to ρ and from ρ to T_2. The matrix does not contain the reading's value y_T, only its standard deviation: the information is fixed by the design of the measurement and by the prior, not by the value observed. Its eigenvalues are 1.155, 1.0003×10^4 and 4.5012×10^7, all positive, with a condition number of 3.9×10^7 that grows as $1/\sigma_p^2$.

4.2.4 The normal equations and the MAP state

The stationary condition. By the chain rule the gradient of the objective (4.24) is

$$\nabla\mathcal{L}(\mathbf{z}) = \mathbf{J}^\mathsf{T}\mathbf{r}(\mathbf{z}), \tag{4.28}$$

and the MAP state is the point where it vanishes, $\mathbf{J}^\mathsf{T}\mathbf{r}(\hat{\mathbf{z}}) = \mathbf{0}$: three equations, one per entry.

The Gauss–Newton step. The step that sets the gradient of the linearized rows to zero solves the normal equations of equation (3.25),

$$(\mathbf{J}^{\mathsf{T}}\mathbf{J})\,\Delta\mathbf{z} = -\mathbf{J}^{\mathsf{T}}\mathbf{r}(\mathbf{z}), \qquad \mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + \Delta\mathbf{z}, \tag{4.29}$$

the
and one step lands on $\hat{\mathbf{z}}$ from any starting state.

The free direction. The two physical lines of $\mathbf{J}$ annihilate one direction of the state,

$$\mathbf{J}_{\text{phys}}\,\mathbf{v} = \mathbf{0}, \qquad \mathbf{J}_{\text{phys}} = \begin{bmatrix} -3000 & 100 & 0 \\ -6000 & 0 & 100 \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} 1 \\ 30 \\ 60 \end{bmatrix}, \tag{4.30}$$

the free direction: a change of ρ with T_1 and T_2 following by $Q_1 = 30\,\text{kW}$ and $Q_2 = 60\,\text{kW}$ times that change, which unbalances neither hall. A pull of the reading on T_1 is answered by a move along $\mathbf{v}$, because any move off it would cost the physical rows far more than it gains on the reading; from a state that balances both halls, the step of the normal equations therefore lies, to within the tolerance, along $\mathbf{v}$.

4.2.5 The posterior covariance from the information matrix

The point estimate does not answer the design question, which depends on how uncertain the state remains.

The covariance. The rows are linear and Gaussian, so by equation (3.26) the posterior is exactly $\mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1},$$

the inverse of the matrix (4.27), with rows and columns ordered (ρ, T_1, T_2), so that its entries carry $(\text{K/kW})^2$, K^2, and K·K/kW for the products. A larger information means a narrower posterior: the covariance is the inverse of the steepness of the objective.

The marginals. The estimation questions ask for entries. The marginals of a Gaussian are read off its mean and covariance,

$$\rho \mid \mathbf{y} \sim \mathcal{N}(\hat{z}_1,\ \Sigma_{11}), \qquad T_1 \mid \mathbf{y} \sim \mathcal{N}(\hat{z}_2,\ \Sigma_{22}), \qquad T_2 \mid \mathbf{y} \sim \mathcal{N}(\hat{z}_3,\ \Sigma_{33}), \tag{4.31}$$

so the resistance is $\hat{\rho} = \hat{z}_1$ with standard deviation $\sqrt{\Sigma_{11}}$, the test temperature $\hat{z}_2$ with $\sqrt{\Sigma_{22}}$, and the full-load temperature, the entry the threshold question asks about, $\hat{T}_2 = \hat{z}_3$ with $\sqrt{\Sigma_{33}}$. How the entries move together is read off the correlations $\Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$, which the physical rows drive toward one.

Coordinates along and across the physical rows. The shape of $\boldsymbol{\Sigma}$, and its limit as the tolerance shrinks, are read by measuring the state along the free direction and across the two physical rows. Write

$$\mathbf{z} = \mathbf{c} + \mathbf{B}\mathbf{u}, \qquad \mathbf{u} = \begin{bmatrix} \rho \\ d_1 \\ d_2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 30 & 1 & 0 \\ 60 & 0 & 1 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 0 \\ 18 \\ 18 \end{bmatrix}, \tag{4.32}$$

where $d_1 = T_1 - 18 - 30\rho$ and $d_2 = T_2 - 18 - 60\rho$ are the two hall imbalances in kelvin and the first column of $\mathbf{B}$ is $\mathbf{v}$. The rows are linear in $\mathbf{u}$ with Jacobian $\mathbf{JB}$, whose columns are $\mathbf{Jv}$, $\mathbf{Je}_2$ and $\mathbf{Je}_3$,

$$\mathbf{JB} = \begin{bmatrix} 0 & 1/\sigma_p & 0 \\ 0 & 0 & 1/\sigma_p \\ 30/\sigma_T & 1/\sigma_T & 0 \\ 1/\sigma_0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 100 & 0 \\ 0 & 0 & 100 \\ 60 & 2 & 0 \\ 40 & 0 & 0 \end{bmatrix}.$$

The first column, $\mathbf{Jv} = (0,\ 0,\ 60,\ 40)^{\mathsf{T}}$, is seen only by the reading (with slope $2 \times 30 = 60$) and the prior (slope 40); a move along $\mathbf{v}$ costs the information $\|\mathbf{Jv}\|^2 = 60^2 + 40^2 = 5200\,(\mathrm{kW/K})^2$. The other two columns carry the entries $1/\sigma_p$, so every direction that unbalances a hall costs about $1/\sigma_p^2 = 10^4\,\mathrm{K}^{-2}$: a tiny tolerance leaves the state free to move cheaply only along $\mathbf{v}$, a narrow ridge in the objective. The third column is orthogonal to the other two, so the information matrix in these coordinates is block diagonal,

$$(\mathbf{JB})^{\mathsf{T}}(\mathbf{JB}) = \begin{bmatrix} 1/\sigma_0^2 + 30^2/\sigma_T^2 & 30/\sigma_T^2 & 0 \\ 30/\sigma_T^2 & 1/\sigma_p^2 + 1/\sigma_T^2 & 0 \\ 0 & 0 & 1/\sigma_p^2 \end{bmatrix} = \begin{bmatrix} 5200 & 120 & 0 \\ 120 & 10\,004 & 0 \\ 0 & 0 & 10\,000 \end{bmatrix},$$

and its inverse, the covariance $\mathbf{\Sigma}_u$ of $\mathbf{u}$, is

$$\mathbf{\Sigma}_u = \begin{bmatrix} 1.9236 \times 10^{-4} & -2.3074 \times 10^{-6} & 0 \\ -2.3074 \times 10^{-6} & 9.9988 \times 10^{-5} & 0 \\ 0 & 0 & 1.0000 \times 10^{-4} \end{bmatrix} \quad (\sigma_p = 0.01\,\mathrm{K}).$$

Three results follow. First, the full-load imbalance d_2 is seen by its own row only: $(\mathbf{\Sigma}_u)_{33} = \sigma_p^2$, the variance that row allows, and it is uncorrelated with ρ and d_1. Second, by the block-inversion formula of Section 3.5.3, the first diagonal entry of $\mathbf{\Sigma}_u$ is the inverse of the Schur complement of the (d_1, d_1) entry in the upper 2×2 block,

$$\Sigma_{11} = (\mathbf{\Sigma}_u)_{11} = \left[\frac{1}{\sigma_0^2} + \frac{30^2}{\sigma_T^2} - \frac{(30/\sigma_T^2)^2}{1/\sigma_p^2 + 1/\sigma_T^2} \right]^{-1} = \left[\frac{1}{\sigma_0^2} + \frac{30^2}{\sigma_T^2 + \sigma_p^2} \right]^{-1}, \tag{4.33}$$

the first row of $\mathbf{B}$ being $\mathbf{e}_1^{\mathsf{T}}$; a tolerance acts on the reading as extra sensor variance and widens the spread of ρ. Third, the covariance of the state is $\mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_u\mathbf{B}^{\mathsf{T}}$; its third diagonal entry, with $T_2 = 18 + 60\rho + d_2$ and d_2 uncorrelated with ρ, is

$$\mathrm{Var}(T_2 \mid \mathbf{y}) = \Sigma_{33} = \mathbf{g}^{\mathsf{T}}\mathbf{\Sigma}\mathbf{g} + \sigma_p^2 = 60^2\,\Sigma_{11} + \sigma_p^2, \qquad \mathbf{g} = (60,\ 0,\ 0)^{\mathsf{T}}, \tag{4.34}$$

with $\mathbf{g}$ the gradient of the prediction $18 + 60\rho$. A model discrepancy therefore enters the answer twice: it widens Σ_{11}, and it adds its own variance at the node that has no reading.

The limit of exact rows. As $\sigma_p \to 0$, $(\mathbf{\Sigma}_u)_{22}$ and $(\mathbf{\Sigma}_u)_{33}$, both at most σ_p^2, vanish and $(\mathbf{\Sigma}_u)_{11}$ tends to $1/5200$ by equation (4.33), so $\mathbf{\Sigma}_u \to \mathrm{diag}(1/5200,\ 0,\ 0)$ and

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_u\mathbf{B}^{\mathsf{T}} \ \longrightarrow\ \frac{\mathbf{B}\mathbf{e}_1\mathbf{e}_1^{\mathsf{T}}\mathbf{B}^{\mathsf{T}}}{5200} = \frac{\mathbf{v}\mathbf{v}^{\mathsf{T}}}{5200} = \frac{1}{5200}\begin{bmatrix} 1 & 30 & 60 \\ 30 & 900 & 1800 \\ 60 & 1800 & 3600 \end{bmatrix}, \tag{4.35}$$

of rank one. The matrix has two zero eigenvalues and no ordinary inverse, and the posterior correlation of T_1 and T_2 is exactly one. This is not a failure of the inference; it is the consequence of exact physical rows, which leave the state one degree of freedom. The state-vector form needs no special treatment of it: the normal equations are solved at a finite tolerance, where $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ is invertible, and the limit is derived in the coordinates of equation (4.32) and read from the sequence of tolerances. The distinction between exact constraints and uncertain physical relations matters whenever a Hessian and a covariance are constructed over the full state, and the tolerance is the device that keeps both defined.

The objective along the free direction. On the line $\mathbf{z} = \hat{\mathbf{z}} + t\,\mathbf{v}$ through the MAP state, in the limit of exact rows, the physical rows keep their values at $\hat{\mathbf{z}}$ (zero in the limit) and only the reading and prior rows change, as $r_3(\hat{\mathbf{z}}) + 60t$ and $r_4(\hat{\mathbf{z}}) + 40t$, the slopes being the entries of $\mathbf{J}\mathbf{v} = (0,\ 0,\ 60,\ 40)^{\mathsf{T}}$. The objective on the line is the reading term $\frac{1}{2}(r_3(\hat{\mathbf{z}}) + 60t)^2$ plus the prior term $\frac{1}{2}(r_4(\hat{\mathbf{z}}) + 40t)^2$. At the minimum the slopes of the two terms cancel, $\mathbf{v}^{\mathsf{T}}\mathbf{J}^{\mathsf{T}}\mathbf{r}(\hat{\mathbf{z}}) = 60\,r_3 + 40\,r_4 = 0$: the reading pulls with weight 60 and the prior with weight 40. The gradient vanishes at $\hat{\mathbf{z}}$, so the objective on the line is

$$\mathcal{L}(\hat{\mathbf{z}}+t\mathbf{v}) = \mathcal{L}(\hat{\mathbf{z}})+\tfrac{1}{2}\,t^2\,\|\mathbf{J}\mathbf{v}\|^2, \qquad \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{L}(\hat{\mathbf{z}}+t\mathbf{v}) = \mathbf{v}^{\mathsf{T}}\mathbf{J}^{\mathsf{T}}\mathbf{J}\,\mathbf{v} = 60^2+40^2 = 5200\,(\mathrm{kW/K})^2, \tag{4.36}$$

a parabola whose curvature is the information along $\mathbf{v}$, equal to $1/\Sigma_{11}$ in the limit; the rise of $\frac{1}{2}$ is reached at $t = \pm 1/\sqrt{5200} = \pm\sqrt{\Sigma_{11}}$.

Remark 4.2. Substituting the exact rows $T_1 = 18 + 30\rho$ and $T_2 = 18 + 60\rho$ into the objective gives the same curve as a function of the resistance alone,

$$\mathcal{L}(\rho) = \frac{1}{2}\left[\frac{(\rho-\mu_0)^2}{\sigma_0^2} + \frac{(y_T - 18 - 30\rho)^2}{\sigma_T^2}\right], \tag{4.37}$$

whose second derivative is the 5200 of equation (4.36) and whose minimum is $\hat{\rho} = 880/5200$.

4.2.6 The Gaussian form of the posterior

In this example the prior is Gaussian, the reading error and the model tolerance are Gaussian, and every row is linear in the state. The negative log posterior is therefore exactly quadratic in $\mathbf{z}$, the curvature term $\sum_k r_k\,\partial^2 r_k/\partial\mathbf{z}\,\partial\mathbf{z}^{\mathsf{T}}$ is zero, the Gauss–Newton matrix $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ is the exact Hessian, and its inverse is the exact posterior covariance under the stated model. Nothing in equations (4.29)–(4.31) is an approximation; the only limit taken, $\sigma_p \to 0$, is derived in equation (4.35).

For a nonlinear physical model, for instance one written in the log-conductance g, the negative log posterior is not quadratic. The same construction is then made locally: the objective is expanded to second order about the MAP state,

$$\mathcal{L}(\mathbf{z}) \approx \mathcal{L}(\hat{\mathbf{z}}) + \tfrac{1}{2}(\mathbf{z}-\hat{\mathbf{z}})^{\mathsf{T}}\mathbf{H}(\mathbf{z}-\hat{\mathbf{z}}), \qquad \mathbf{H} = \left.\frac{\partial^2\mathcal{L}}{\partial\mathbf{z}\,\partial\mathbf{z}^{\mathsf{T}}}\right|_{\hat{\mathbf{z}}} = \mathbf{J}^{\mathsf{T}}\mathbf{J} + \sum_k r_k \left.\frac{\partial^2 r_k}{\partial\mathbf{z}\,\partial\mathbf{z}^{\mathsf{T}}}\right|_{\hat{\mathbf{z}}},$$

with no first-order term because the gradient vanishes at the minimum, and the posterior is replaced by the Gaussian that this expansion defines, the Laplace approximation:

$$\mathbf{z} \mid \mathbf{y} \approx \mathcal{N}(\hat{\mathbf{z}},\ \mathbf{H}^{-1}). \tag{4.38}$$

It holds when the posterior is close to Gaussian over its own width and when $\mathbf{H}$ is positive definite in the chosen free coordinates. The distinction matters throughout the book: the calculation of this chapter is exact, while a Laplace approximation in general is not, and its adequacy is one of the checks of step 5.

4.2.7 Rows that are not Gaussian

Every row of the hall is Gaussian, and its term in the objective is half its squared residual. Chapter 2's exceedance log already has a row of another kind, a count of exceedances of the hall's committed load in a number of logged hours. The construction above carries over to it unchanged once the objective is defined by the posterior rather than by a sum of squares.

The objective. The objective is the negative logarithm of the posterior up to a constant. Each row, a likelihood factor or a prior factor, contributes one term, its negative log-probability:

$$\mathcal{L}(\mathbf{z}) = \sum_k \mathcal{L}_k(\mathbf{z}), \qquad \mathcal{L}_k(\mathbf{z}) = -\ln p_k(\cdot \mid \mathbf{z}), \qquad p(\mathbf{z} \mid \mathbf{y}) \propto \exp[-\mathcal{L}(\mathbf{z})]. \tag{4.39}$$

A Gaussian row is the case $\mathcal{L}_k = \frac{1}{2} r_k^2$, and the objective (4.24) of the hall is this sum for its four rows.

What the fit uses. The MAP state, the Newton step and the Laplace covariance use the objective only through its gradient and its Hessian, and both are sums over the rows:

$$\nabla\mathcal{L} = \sum_k \nabla\mathcal{L}_k, \qquad \mathbf{H} = \sum_k \nabla^2\mathcal{L}_k, \qquad \mathbf{H}\,\Delta\mathbf{z} = -\nabla\mathcal{L}, \qquad \mathbf{\Sigma} = \mathbf{H}(\hat{\mathbf{z}})^{-1}. \tag{4.40}$$

The value of $\mathcal{L}$ enters only the evidence, equation (4.19). When the state has one entry ξ, the gradient is the slope $\mathcal{L}'(\xi)$, the Hessian the curvature $\mathcal{L}''(\xi)$, the step $\Delta\xi = -\mathcal{L}'/\mathcal{L}''$ and the variance $1/\mathcal{L}''(\hat{\xi})$; nothing else changes.

The Gaussian row. For $\mathcal{L}_k = \frac{1}{2} r_k^2$ with Jacobian line $\mathbf{j}_k = \partial r_k / \partial \mathbf{z}^\mathsf{T}$, the chain rule gives

$$\nabla\mathcal{L}_k = \mathbf{j}_k^\mathsf{T} r_k, \qquad \nabla^2\mathcal{L}_k = \mathbf{j}_k^\mathsf{T}\mathbf{j}_k + r_k \frac{\partial^2 r_k}{\partial \mathbf{z}\,\partial \mathbf{z}^\mathsf{T}}. \tag{4.41}$$

Summed over the rows these are $\mathbf{J}^\mathsf{T}\mathbf{r}$ and $\mathbf{J}^\mathsf{T}\mathbf{J}$ plus the curvature term of Section 4.2.6.

The count row. A record of n_k events (the count k of Chapter 2) whose expected number is $\mu_k(\mathbf{z})$, Poisson in windows, has the term $\mathcal{L}_k = \mu_k - n_k \ln \mu_k$. Its gradient and Hessian are

$$\nabla\mathcal{L}_k = (\mu_k - n_k)\,\nabla \ln \mu_k, \qquad \nabla^2\mathcal{L}_k = \mu_k\,\nabla \ln \mu_k\,\nabla \ln \mu_k^\mathsf{T} + (\mu_k - n_k)\,\nabla^2 \ln \mu_k. \tag{4.42}$$

The first term of the Hessian is the product of one line with itself, so it is $\mathbf{j}_k^\mathsf{T}\mathbf{j}_k$ for the line and the residual

$$\mathbf{j}_k = \sqrt{\mu_k}\,\nabla \ln \mu_k^\mathsf{T}, \qquad r_k = \frac{\mu_k - n_k}{\sqrt{\mu_k}}, \tag{4.43}$$

which also give $\mathbf{j}_k^\mathsf{T} r_k = \nabla\mathcal{L}_k$. This pair, the *working residual* and the *working Jacobian line*, is how a count row joins the Jacobian beside Gaussian rows: one more line of $\mathbf{J}$, one more entry of $\mathbf{r}$, and the same products (4.40). The remaining term $(\mu_k - n_k)\,\nabla^2 \ln \mu_k$ plays the part of $r_k\,\partial^2 r_k$ in a Gaussian row. It vanishes when $\ln \mu_k$ is linear in the state, which is the case when the state carries the log-rate, the canonical parameter of the Poisson distribution.

The working pair reproduces the row's gradient and the Gauss–Newton part of its Hessian, and nothing more. Its half square is not the row's term, $\frac{1}{2} r_k^2 \neq \mathcal{L}_k$, and $\mathbf{j}_k$ is not the derivative of r_k except where $r_k = 0$; the fit never uses either, because it uses only the products (4.40). For a Gaussian row that is linear in the state, the working pair is the row's own residual and Jacobian line, which is why the two forms are the same in the hall.

The information and its curvature term. With every row in this form, the Hessian is

$$\mathbf{H} = \mathbf{J}^\mathsf{T}\mathbf{J} + \mathbf{C}, \qquad \mathbf{\Lambda} = \mathbf{J}^\mathsf{T}\mathbf{J}, \tag{4.44}$$

with $\mathbf{C}$ the sum of the rows' remaining terms. The matrix $\mathbf{\Lambda}$ is the information; it is the average of $\mathbf{H}$ over the records the model could have produced, the *expected information*. The two are equal when $\mathbf{C} = \mathbf{0}$: for linear Gaussian rows, and for count rows carried on their canonical parameter. Where $\mathbf{C}$ is not zero, the chapters name it, keep it in the Newton step and report the covariance from $\mathbf{H}$. Once $\hat{\mathbf{z}}$ and $\mathbf{\Sigma}$ are in hand, the kind of row no longer matters: every answer of step 4 is read from them in the same way.

4.2.8 The reduced state

The full state carries every free quantity of the graph, and in the limit of exact rows the physical rows make some of its entries functions of the others; in the hall, two of the three entries follow from the third. The same posterior can then be written over the entries that remain free, the reduced state, as a projection of the rows onto the null space of the physical rows; hall B of Chapter 5 and every plant after it are fitted in this form, so the derivation below is written for any plant, with the hall as its instance.

The null space of the physical rows. Let the full state $\mathbf{z}$ have n entries and the graph m physical rows, with Jacobian $\mathbf{A} = \mathbf{J}_{\text{phys}}$, an $m \times n$ matrix. Split the full state into $n_{\text{f}} = n - m$ free entries $\mathbf{z}_{\text{f}}$ and m dependent entries $\mathbf{z}_{\text{d}}$, and $\mathbf{A}$ into the matching columns, $\mathbf{A} = [\,\mathbf{A}_{\text{free}}\ \ \mathbf{A}_{\text{dep}}\,]$, an $m \times n_{\text{f}}$ and an $m \times m$ block (the entries ordered free first). The split is a choice: any split serves for which $\mathbf{A}_{\text{dep}}$ is square and invertible, that is, for which the physical rows can be solved for the dependent entries at given free ones. A move $\Delta\mathbf{z}$ of the full state keeps the linearized balances when $\mathbf{A}\,\Delta\mathbf{z} = \mathbf{0}$, that is $\mathbf{A}_{\text{free}}\,\Delta\mathbf{z}_{\text{f}} + \mathbf{A}_{\text{dep}}\,\Delta\mathbf{z}_{\text{d}} = \mathbf{0}$, so

$$\Delta\mathbf{z}_{\text{d}} = \mathbf{N}_{\text{b}}\,\Delta\mathbf{z}_{\text{f}}, \qquad \mathbf{N}_{\text{b}} = -\mathbf{A}_{\text{dep}}^{-1}\mathbf{A}_{\text{free}}, \qquad \mathbf{N} = \begin{bmatrix} \mathbf{I}_{n_{\text{f}}} \\ \mathbf{N}_{\text{b}} \end{bmatrix}, \qquad \mathbf{A}\mathbf{N} = \mathbf{0}. \tag{4.45}$$

Each physical row's tolerance divides both of its blocks and cancels in $\mathbf{N}_{\text{b}}$. The n_{f} columns of the $n \times n_{\text{f}}$ matrix $\mathbf{N}$ are independent, through its identity block, and $\mathbf{A}$ has rank m, so they span the null space of the physical rows, the tangent of the set on which the balances hold; each column moves one free entry by one unit and the dependent entries as the linearized balances require. In the hall $n = 3$ and $m = 2$; with $\mathbf{z}_{\text{f}} = \rho$ and $\mathbf{z}_{\text{d}} = (T_1, T_2)^{\mathsf{T}}$,

$$\mathbf{A}_{\text{free}} = \begin{bmatrix} -Q_1/\sigma_p \\ -Q_2/\sigma_p \end{bmatrix}, \qquad \mathbf{A}_{\text{dep}} = \begin{bmatrix} 1/\sigma_p & 0 \\ 0 & 1/\sigma_p \end{bmatrix}, \qquad \mathbf{N}_{\text{b}} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}, \qquad \mathbf{N} = \begin{bmatrix} 1 \\ 30 \\ 60 \end{bmatrix},$$

the entries of $\mathbf{N}_{\text{b}}$ the loads in kW: a unit change of resistance raises each hall temperature by its load. The single column of $\mathbf{N}$ is the free direction $\mathbf{v}$ of equation (4.30). Taking T_1 as the free entry would serve as well, since the block of (ρ, T_2) is invertible too; the resistance is chosen because the prior acts on it and each hall temperature follows from it by its row of $\mathbf{N}_{\text{b}}$.

The rows of the reduced state. With exact balances the full state lies on the set where the physical rows vanish, and the reduced state $\mathbf{z}_{\text{f}}$ names a point of that set; the point itself is $\mathbf{z}_{\text{full}}(\mathbf{z}_{\text{f}}) = (\mathbf{z}_{\text{f}},\ \mathbf{z}_{\text{d}}(\mathbf{z}_{\text{f}}))$, with $\mathbf{z}_{\text{d}}(\mathbf{z}_{\text{f}})$ the solution of the physical rows at the given free entries, and its derivative with respect to $\mathbf{z}_{\text{f}}$ is $\mathbf{N}$. Write $\mathbf{r}_{\text{rows}}$ for the evidence rows, the readings and the priors, as functions of the full state. The rows of the reduced state are the evidence rows at the point of the set,

$$\mathbf{r}(\mathbf{z}_{\text{f}}) = \mathbf{r}_{\text{rows}}\big(\mathbf{z}_{\text{full}}(\mathbf{z}_{\text{f}})\big), \tag{4.46}$$

and the objective on the reduced state is the full objective (4.24) on the set, where its physical rows contribute nothing. In the hall, with $\mathbf{z}_{\text{full}}(\rho) = (\rho,\ T_c + Q_1\rho,\ T_c + Q_2\rho)^{\mathsf{T}}$, the rows are

$$\mathbf{r}(\rho) = \begin{bmatrix} (T_c + Q_1\rho - y_T)/\sigma_T \\ (\rho - \mu_0)/\sigma_0 \end{bmatrix} \begin{matrix} \text{reading} \\ \text{prior} \end{matrix},$$

and the objective is the curve (4.37), reached here as a projection and a limit rather than by a substitution made by hand.

The reduced Jacobian. A move $\Delta\mathbf{z}_\mathrm{f}$ of the reduced state moves the point of the set by $\mathbf{N}\,\Delta\mathbf{z}_\mathrm{f}$, so the chain rule gives the Jacobian of the reduced rows as the Jacobian of the evidence rows projected onto the null space,

$$\mathbf{J}_\mathrm{red} = \mathbf{J}_\mathrm{rows}\mathbf{N}, \tag{4.47}$$

where $\mathbf{J}_\mathrm{rows} = \partial\mathbf{r}_\mathrm{rows}/\partial\mathbf{z}$ is the Jacobian of the evidence rows (readings and priors) with respect to the full state, the lines of $\mathbf{J}$ that are not physical rows. A reading whose row touches only dependent entries sees the free entries through $\mathbf{N}_\mathrm{b}$. In the hall $\mathbf{J}_\mathrm{rows}$ is the reading and prior lines of equation (4.25),

$$\mathbf{J}_\mathrm{rows} = \begin{bmatrix} 0 & 1/\sigma_T & 0 \\ 1/\sigma_0 & 0 & 0 \end{bmatrix}, \qquad \mathbf{J}_\mathrm{red} = \begin{bmatrix} Q_1/\sigma_T \\ 1/\sigma_0 \end{bmatrix}.$$

The reading, whose row in the full state touches only T_1, sees the resistance on the reduced state with the slope Q_1/σ_T: its entry is nonzero, so the reading alone determines the one free entry (the hall is observable from its one sensor). The information of the reduced state is $\mathbf{J}_\mathrm{red}^\mathsf{T}\mathbf{J}_\mathrm{red} = Q_1^2/\sigma_T^2 + 1/\sigma_0^2$, the reading's term and the prior's, equal to $\|\mathbf{J}\mathbf{v}\|^2$ of equation (4.36).

Nonlinear physical rows. When the physical rows are nonlinear in the state, as they are when a conductance is carried as a log-derate or a recirculation fraction multiplies a temperature difference, the same objects are evaluated point by point. $\mathbf{z}_\mathrm{full}(\mathbf{z}_\mathrm{f})$ is the solution of the physical rows at the given free entries, found by Newton's method on those rows alone; $\mathbf{N}(\mathbf{z}_\mathrm{f})$, from the blocks of $\mathbf{A}$ at that solution, and $\mathbf{J}_\mathrm{red}(\mathbf{z}_\mathrm{f}) = \mathbf{J}_\mathrm{rows}\,\mathbf{N}(\mathbf{z}_\mathrm{f})$ are evaluated there and change with the free entries. The Gauss–Newton iteration (4.29) runs on the reduced rows, $(\mathbf{J}_\mathrm{red}^\mathsf{T}\mathbf{J}_\mathrm{red})\,\Delta\mathbf{z}_\mathrm{f} = -\mathbf{J}_\mathrm{red}^\mathsf{T}\mathbf{r}(\mathbf{z}_\mathrm{f})$, each iterate solving the physical rows again. At the MAP state, $\mathbf{\Sigma}_\mathrm{red} = (\mathbf{J}_\mathrm{red}^\mathsf{T}\mathbf{J}_\mathrm{red})^{-1}$ is the Gauss–Newton form of the Laplace covariance (4.38); the term it drops holds the second derivatives of the reduced rows, which now include the curvature of the set, and each chapter checks it numerically. The hall's rows are linear, $\mathbf{N}$ is constant, one step is exact and $\mathbf{\Sigma}_\mathrm{red}$ is the exact covariance.

The reduced covariance as the limit. The covariance of the reduced state is $\mathbf{\Sigma}_\mathrm{red} = (\mathbf{J}_\mathrm{red}^\mathsf{T}\mathbf{J}_\mathrm{red})^{-1}$, and it is the free-entry block $\mathbf{\Sigma}_\mathrm{ff}$ of the full $\mathbf{\Sigma}$ in the limit of exact rows. Write a move of the full state in the coordinates $\Delta\mathbf{z} = \mathbf{B}\,\Delta\mathbf{u}$ with $\mathbf{B} = [\,\mathbf{N}\;\;\mathbf{B}_\mathrm{d}\,]$, where $\mathbf{B}_\mathrm{d} = [\,\mathbf{0}\;\;\mathbf{I}_m\,]^\mathsf{T}$ is the $n \times m$ matrix that moves only the dependent entries. The square matrix $\mathbf{B}$ is invertible and its first block row is $[\,\mathbf{I}_{n_\mathrm{f}}\;\;\mathbf{0}\,]$, so the first n_f coordinates are the move of the free entries; in the hall $\mathbf{B}$ is the matrix of equation (4.32). Since $\mathbf{A}\mathbf{N} = \mathbf{0}$ and $\mathbf{A}\mathbf{B}_\mathrm{d} = \mathbf{A}_\mathrm{dep}$, the Jacobian in these coordinates, physical rows first, and its information matrix are

$$\mathbf{J}\,[\,\mathbf{N}\;\;\mathbf{B}_\mathrm{d}\,] = \begin{bmatrix} \mathbf{0} & \mathbf{A}_\mathrm{dep} \\ \mathbf{J}_\mathrm{red} & \mathbf{J}_\mathrm{rows}\mathbf{B}_\mathrm{d} \end{bmatrix}, \qquad \begin{bmatrix} \mathbf{J}_\mathrm{red}^\mathsf{T}\mathbf{J}_\mathrm{red} & \mathbf{E} \\ \mathbf{E}^\mathsf{T} & \mathbf{F} + \mathbf{A}_\mathrm{dep}^\mathsf{T}\mathbf{A}_\mathrm{dep} \end{bmatrix},$$

with $\mathbf{E} = \mathbf{J}_\mathrm{red}^\mathsf{T}\mathbf{J}_\mathrm{rows}\mathbf{B}_\mathrm{d}$ and $\mathbf{F} = \mathbf{B}_\mathrm{d}^\mathsf{T}\mathbf{J}_\mathrm{rows}^\mathsf{T}\mathbf{J}_\mathrm{rows}\mathbf{B}_\mathrm{d}$. By the block-inversion formula of Section 3.5.3 the free-entry block of the covariance in these coordinates, which is $\mathbf{\Sigma}_\mathrm{ff}$ because the first block row of $\mathbf{B}$ is $[\,\mathbf{I}_{n_\mathrm{f}}\;\;\mathbf{0}\,]$, is the inverse of the Schur complement of the lower block, and that block grows without bound as the tolerances shrink: each line of $\mathbf{A}_\mathrm{dep}$ carries its row's $1/\sigma_p$, so with a common tolerance $\mathbf{A}_\mathrm{dep}^\mathsf{T}\mathbf{A}_\mathrm{dep} = \sigma_p^{-2}\,\tilde{\mathbf{A}}_\mathrm{dep}^\mathsf{T}\tilde{\mathbf{A}}_\mathrm{dep}$, with $\tilde{\mathbf{A}}_\mathrm{dep}$ the invertible block of the unnormalized rows:

$$\mathbf{\Sigma}_\mathrm{ff} = \Big[\mathbf{J}_\mathrm{red}^\mathsf{T}\mathbf{J}_\mathrm{red} - \mathbf{E}\big(\mathbf{F} + \mathbf{A}_\mathrm{dep}^\mathsf{T}\mathbf{A}_\mathrm{dep}\big)^{-1}\mathbf{E}^\mathsf{T}\Big]^{-1} \;\longrightarrow\; \big(\mathbf{J}_\mathrm{red}^\mathsf{T}\mathbf{J}_\mathrm{red}\big)^{-1} = \mathbf{\Sigma}_\mathrm{red} \qquad (\sigma_p \to 0). \tag{4.48}$$

The correction the bracket subtracts is of order σ_p^2; for nonlinear rows the equality holds with every block evaluated at the point of linearization, the MAP state. In the hall $\mathbf{\Sigma}_\mathrm{ff}$ is the scalar

Σ_{11}, $\mathbf{E} = (Q_1/\sigma_T^2,\ 0)$, $\mathbf{F} = \mathrm{diag}(1/\sigma_T^2,\ 0)$ and $\mathbf{A}_{\mathrm{dep}}^{\mathsf{T}}\mathbf{A}_{\mathrm{dep}} = \sigma_p^{-2}\,\mathbf{I}_2$, and the bracket is the Schur complement of equation (4.33). The same coordinates give the covariance of the dependent entries in the limit, $\mathbf{N}_{\mathrm{b}}\boldsymbol{\Sigma}_{\mathrm{red}}\mathbf{N}_{\mathrm{b}}^{\mathsf{T}}$, in the hall the lower block of the limit (4.35): the dependent entries are known through the free ones, each by its row of $\mathbf{N}_{\mathrm{b}}$. The exact balance is this limit, never a substitution made by hand.

4.2.9 The fit of the hall state

The fit is run at the tolerance $\sigma_p = 0.01\,\mathrm{K}$ of Table 4.3, and the limit of exact rows is read off equation (4.35).

The step from the handbook state. The iteration starts at the forward calculation with the handbook resistance, $\mathbf{z}^{(0)} = (0.160,\ 22.8,\ 27.6)^{\mathsf{T}}$, where both physical rows and the prior row vanish and the reading row is $(22.8 - 23.2)/0.5 = -0.8$, so that $\mathcal{L}(\mathbf{z}^{(0)}) = 0.32$ and $\mathbf{J}^{\mathsf{T}}\mathbf{r}(\mathbf{z}^{(0)}) = (0,\ -1.6,\ 0)^{\mathsf{T}}$: only the reading is out of balance, and it pulls on T_1 alone. The normal equations (4.29) are

$$\begin{bmatrix} 45\,001\,600 & -300\,000 & -600\,000 \\ -300\,000 & 10\,004 & 0 \\ -600\,000 & 0 & 10\,000 \end{bmatrix} \Delta\mathbf{z} = \begin{bmatrix} 0 \\ 1.6 \\ 0 \end{bmatrix}, \qquad \Delta\mathbf{z} = \begin{bmatrix} 0.0092296 \\ 0.27694 \\ 0.55378 \end{bmatrix}.$$

Divided by its first entry, the step is $\Delta\mathbf{z} = 0.0092296 \times (1,\ 30.005,\ 60.000)^{\mathsf{T}}$: it lies, to within the tolerance, along the free direction $\mathbf{v}$ of equation (4.30). Table 4.4 lists the iteration. Each line is one iterate: its three entries, the objective, and the largest entry of the gradient (4.28). After the first step the gradient is at round-off and the second step moves no entry by as much as 10^{-9}.

Table 4.4: Gauss–Newton iterations for the hall state from the handbook state (0.160, 22.8, 27.6), with $\sigma_p = 0.01\,\mathrm{K}$.

Iterate k	ρ [K/kW]	T_1 [°C]	T_2 [°C]	$\mathcal{L}$ [–]	$\max_i \lvert(\mathbf{J}^{\mathsf{T}}\mathbf{r})_i\rvert$ [kW/K or K^{-1}]
0	0.160000	22.8000	27.6000	0.32000	1.6
1	0.169230	23.0769	28.1538	0.09845	$< 10^{-8}$
2	0.169230	23.0769	28.1538	0.09845	$< 10^{-8}$

The MAP state. The fit returns

$$\hat{\mathbf{z}} = \begin{bmatrix} \hat{\rho} \\ \hat{T}_1 \\ \hat{T}_2 \end{bmatrix} = \begin{bmatrix} 0.16923 \\ 23.0769 \\ 28.1538 \end{bmatrix} : \tag{4.49}$$

a resistance of $0.16923\,\mathrm{K/kW}$, a test temperature of 23.077°C and a full-load temperature of 28.154°C, each read straight off its entry, with the effective conductance $\hat{G} = 1/\hat{\rho} = 5.909\,\mathrm{kW/K}$ evaluated at the estimated resistance.

4.2.10 The covariance and the resistance

The covariance. Inverting the matrix (4.27),

$$\boldsymbol{\Sigma} = \left(\mathbf{J}^{\mathsf{T}}\mathbf{J}\right)^{-1} = \begin{bmatrix} 1.9236 \times 10^{-4} & 5.7685 \times 10^{-3} & 1.1542 \times 10^{-2} \\ 5.7685 \times 10^{-3} & 0.17309 & 0.34611 \\ 1.1542 \times 10^{-2} & 0.34611 & 0.69260 \end{bmatrix} \quad (\sigma_p = 0.01\,\mathrm{K}). \tag{4.50}$$

Its eigenvalues are 0.866 along (nearly) $\mathbf{v}$, and 1.0×10^{-4} and 2.2×10^{-8} in the two directions the physics fixes, both shrinking as σ_p^2.

The entries read off. The resistance has

$$\hat{\rho} = \hat{z}_1 = 0.16923\,\mathrm{K/kW}, \qquad \sqrt{\Sigma_{11}} = \sqrt{1.9236 \times 10^{-4}} = 0.01387\,\mathrm{K/kW},$$

the test temperature $\sqrt{\Sigma_{22}} = 0.416\,\mathrm{K}$, and the full-load temperature $\hat{T}_2 = \hat{z}_3 = 28.154°\mathrm{C}$ with $\sqrt{\Sigma_{33}} = \sqrt{0.69260} = 0.8322\,\mathrm{K}$. The correlations $\Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$ are 0.99971 between ρ and T_1, 0.99993 between ρ and T_2, and 0.99964 between T_1 and T_2.

The limit of exact rows. The limit covariance (4.35) gives $\Sigma_{11} = 1/5200$, so that

$$\sqrt{\Sigma_{11}} = 0.013868\,\mathrm{K/kW}, \qquad \hat{\rho} = 880/5200 = 0.169231\,\mathrm{K/kW}, \qquad \hat{T}_2 = 18{+}60\hat{\rho} = 28.1538°\mathrm{C},$$

the resistance being the minimum of equation (4.37); the rest of the chapter carries these limit values.

The objective along the free direction. Figure 4.3 draws the objective (4.36) on the line $\mathbf{z} = \hat{\mathbf{z}} + t\,\mathbf{v}$ in the limit of exact rows, with t read on the horizontal axis as the resistance $\rho = \hat{z}_1 + t$ in K/kW and the dimensionless objective on the vertical axis: the solid curve is the objective, the dashed curve the prior term and the dotted curve the reading term; the dot marks the minimum and the double arrow joins the two points where the objective has risen by $\frac{1}{2}$. At the minimum $r_3(\hat{\mathbf{z}}) = -0.2462$ and $r_4(\hat{\mathbf{z}}) = 0.3692$, and the slopes cancel: $|60\,r_3(\hat{\mathbf{z}}) + 40\,r_4(\hat{\mathbf{z}})| < 10^{-9}$, round-off.

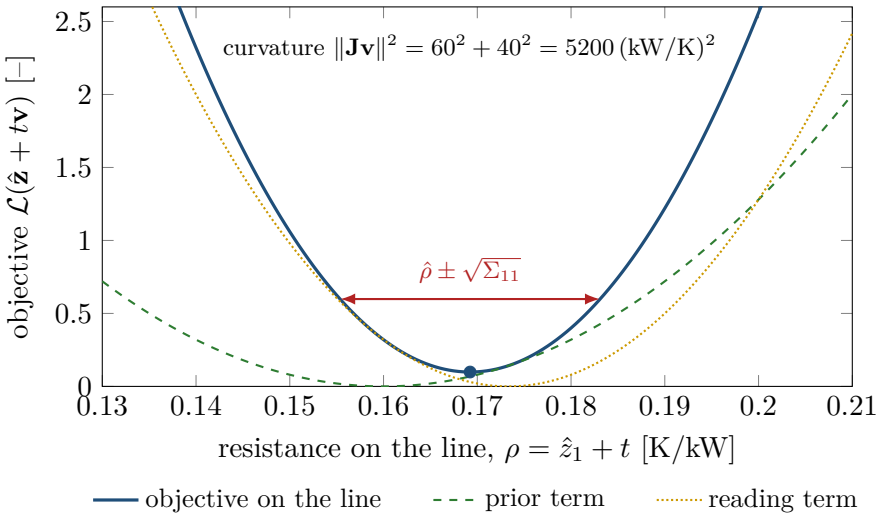


Figure 4.3: The objective on the line $\mathbf{z} = \hat{\mathbf{z}} + t\,\mathbf{v}$ along the free direction $\mathbf{v} = (1, 30, 60)^\mathsf{T}$, in the limit of exact physical rows, with its prior term $\frac{1}{2}(r_4(\hat{\mathbf{z}}) + 40t)^2$ and reading term $\frac{1}{2}(r_3(\hat{\mathbf{z}}) + 60t)^2$; the horizontal axis is the resistance $\hat{z}_1 + t$. Dot: the minimum at $\hat{\rho} = 0.16923\,\mathrm{K/kW}$, $\mathcal{L} = 0.0985$. Double arrow: the points $\hat{\rho} \pm \sqrt{\Sigma_{11}}$, $\sqrt{\Sigma_{11}} = 0.01387\,\mathrm{K/kW}$, where the objective has risen by $\frac{1}{2}$.

The posterior of the resistance. By the first marginal of equation (4.31),

$$\rho \mid \mathbf{y} \sim \mathcal{N}(\hat{z}_1,\ \Sigma_{11}) = \mathcal{N}(0.169231,\ 0.013868^2), \tag{4.51}$$

a standard deviation of $0.013868/0.169231 = 0.082$ of the estimate. Figure 4.4 compares it with the prior. The horizontal axis is the resistance and the vertical axis the probability density in kW/K; the dashed curve is the prior of equation (4.12) and the solid curve the posterior of equation (4.51). The reading has shifted the resistance upward, from 0.160 to 0.169, and narrowed it, from a standard deviation of 0.025 to 0.0139.

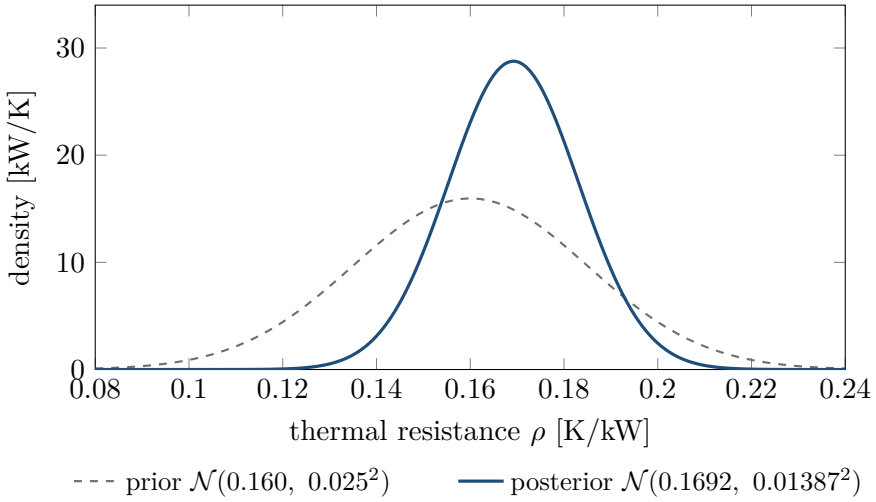


Figure 4.4: Prior and posterior densities of the thermal resistance, the first entry of the state: $\mathcal{N}(0.160, 0.025^2)$ and $\mathcal{N}(0.169231, 0.013868^2)$. The reading moves the resistance up and reduces its spread.

4.2.11 The reduced state of the hall

At the tolerance $\sigma_p = 0.01\,\mathrm{K}$ of equation (4.25), $\mathbf{A}_{\mathrm{free}} = (-3000,\ -6000)^{\mathsf{T}}$ and $\mathbf{A}_{\mathrm{dep}} = \mathrm{diag}(100,\ 100)$. From the handbook resistance 0.160 the reduced rows (4.46) are $(-0.8,\ 0)$, $\mathbf{J}_{\mathrm{red}}^{\mathsf{T}}\mathbf{r} = -48$, and one Gauss–Newton step $48/5200 = 0.0092308$ lands on $\hat{\rho} = 0.169231\,\mathrm{K/kW}$, which a second step does not move. The carried entries have $\mathbf{N}_{\mathrm{b}}\boldsymbol{\Sigma}_{\mathrm{red}}\mathbf{N}_{\mathrm{b}}^{\mathsf{T}}$ with diagonal 0.17308 and $0.69231\,\mathrm{K}^2$ and off-diagonal $0.34615\,\mathrm{K}^2$.

Table 4.5 sets the reduced state against the full one. Its first two rows give $\mathbf{N}$ and $\mathbf{J}_{\mathrm{red}}$, which do not depend on the tolerance; the remaining rows give, for the full state at three tolerances and for the reduced state in the last column, the (ρ, ρ) entry of the covariance, its square root, the MAP resistance, and the relative difference of the full entry from $\boldsymbol{\Sigma}_{\mathrm{red}}$. The difference shrinks a hundredfold per decade of σ_p, as the order σ_p^2 of the correction in equation (4.48) requires, and at $\sigma_p = 0.001\,\mathrm{K}$ the two agree to every printed digit.

Table 4.5: The reduced state of the hall against the free-entry marginal of the full state. $\mathbf{N}$ is in [–, kW, kW]; the last column is the reduced state, exact in the limit of exact rows.

Quantity	$\sigma_p = 0.1\,\mathrm{K}$	$0.01\,\mathrm{K}$	$0.001\,\mathrm{K}$	reduced state
null-space basis $\mathbf{N}$	$(1,\ 30,\ 60)^{\mathsf{T}}$ at every tolerance			
$\mathbf{J}_{\mathrm{red}} = \mathbf{J}_{\mathrm{rows}}\mathbf{N}$ [kW/K]	$(60,\ 40)^{\mathsf{T}}$, information $5200\,(\mathrm{kW/K})^2$			
Σ_{11}, $\boldsymbol{\Sigma}_{\mathrm{red}}$ $[(\mathrm{K/kW})^2]$	1.97568×10^{-4}	1.92361×10^{-4}	1.92308×10^{-4}	1.92308×10^{-4}
standard deviation [K/kW]	0.014056	0.013869	0.013868	0.013868
$\hat{\rho}$ [K/kW]	0.169119	0.169230	0.169231	0.169231
relative difference from $\boldsymbol{\Sigma}_{\mathrm{red}}$ [–]	2.7×10^{-2}	2.8×10^{-4}	2.8×10^{-6}	–

4.3 The answers (step 4)

Each question of Table 4.2 is read from the posterior of Section 4.2, and the last subsection carries the same posterior to loads the hall has not carried.

4.3.1 Uncertainty propagation and the delta method

The rows of the hall are linear, so the entries of $\boldsymbol{\Sigma}$ are exact variances. In most plant models an output is a nonlinear function of the state, $f(\mathbf{z})$, and the posterior is Gaussian at best approximately, $\mathbf{z} \mid \mathbf{y} \approx \mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$.

Linearization of the output. The first-order Taylor expansion of f about $\hat{\mathbf{z}}$ is

$$f(\mathbf{z}) \approx f(\hat{\mathbf{z}}) + \nabla f(\hat{\mathbf{z}})^{\mathsf{T}}(\mathbf{z} - \hat{\mathbf{z}}).$$

Writing $\mathbf{g} = \nabla f(\hat{\mathbf{z}})$ for the column gradient,

$$f(\mathbf{z}) \approx f(\hat{\mathbf{z}}) + \mathbf{g}^{\mathsf{T}}(\mathbf{z} - \hat{\mathbf{z}}). \tag{4.52}$$

Taking the posterior expectation, with $\mathbb{E}[\mathbf{z} - \hat{\mathbf{z}} \mid \mathbf{y}] = \mathbf{0}$ because $\hat{\mathbf{z}}$ is the mean of the Gaussian,

$$\mu_f = \mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}] \approx f(\hat{\mathbf{z}}).$$

Taking the variance, the constant $f(\hat{\mathbf{z}})$ drops out, and the variance of the linear combination $\mathbf{g}^{\mathsf{T}}(\mathbf{z} - \hat{\mathbf{z}}) = \sum_i g_i(z_i - \hat{z}_i)$ is the double sum of its covariances,

$$\operatorname{Var}(f(\mathbf{z}) \mid \mathbf{y}) \approx \operatorname{Var}\Big(\sum_i g_i z_i \;\Big|\; \mathbf{y}\Big) = \sum_i \sum_j g_i\, g_j\, \operatorname{Cov}(z_i, z_j \mid \mathbf{y}),$$

which in matrix form is

$$\operatorname{Var}(f(\mathbf{z}) \mid \mathbf{y}) \approx \mathbf{g}^{\mathsf{T}} \boldsymbol{\Sigma}\, \mathbf{g}. \tag{4.53}$$

This is the first-order delta method, one of the most reused formulas in probabilistic plant modeling. Its only approximation is the linearization (4.52); when f is linear and the posterior Gaussian, it is exact.

Application to the thermal system. The full-load temperature is the third entry of the state, so its gradient is the third unit vector and the quadratic form picks one diagonal entry,

$$\mathbf{g} = \frac{\partial T_2}{\partial \mathbf{z}} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \qquad \mathbf{g}^{\mathsf{T}} \boldsymbol{\Sigma}\, \mathbf{g} = \Sigma_{33}.$$

The same temperature can also be predicted as a derived quantity of the resistance alone, $f(\mathbf{z}) = T_c + Q_2 \rho = 18 + 60\rho$, with

$$\mathbf{g} = \frac{\partial f}{\partial \mathbf{z}} = \begin{bmatrix} 60 \\ 0 \\ 0 \end{bmatrix}, \qquad \mathbf{g}^{\mathsf{T}} \boldsymbol{\Sigma}\, \mathbf{g} = 60^2\, \Sigma_{11}.$$

In the limit the two are equal, since the limit covariance (4.35) has $\Sigma_{33} = 3600/5200 = 60^2\, \Sigma_{11}$; at a finite tolerance they differ by σ_p^2, equation (4.34), because the entry carries the discrepancy of the full-load balance and the prediction from ρ does not. The two calculations, reading the variance off the covariance and propagating the resistance's uncertainty through a gradient, agree in the limit because they are the same transformation of the same uncertainty: an output that is an entry of the state needs no propagation, and an output that is not needs its gradient.

4.3.2 The gradient and the variance budget

Physical meaning of the gradient. The first entry of the gradient $\mathbf{g} = (60,\ 0,\ 0)^{\mathsf{T}}$ of the prediction $18 + 60\rho$, 60 kW, means that a change of 0.01 K/kW in the resistance changes the full-load temperature by $\Delta T_2 = 60 \times 0.01 = 0.6\,\mathrm{K}$. The gradient is the link between physical uncertainty and design consequence. A state entry can have a large posterior variance and still contribute little to an output that is insensitive to it, with $g_i \approx 0$; conversely, a modest uncertainty in a highly influential entry can dominate the uncertainty of a prediction.

Several uncertain entries. Suppose the supply-air temperature is uncertain as well, the controller holding it only near its set point, so that the full-load temperature $T = T_c + Q\rho$ depends on the sub-vector $\boldsymbol{\theta} = (T_c, \rho)^{\mathsf{T}}$ of the state, with posterior covariance $\boldsymbol{\Sigma}_\theta$. The gradient is $\mathbf{g} = (\partial T/\partial T_c,\ \partial T/\partial \rho)^{\mathsf{T}} = (1,\ Q)^{\mathsf{T}}$, and equation (4.53) gives

$$\mathrm{Var}(T) = \mathbf{g}^{\mathsf{T}} \boldsymbol{\Sigma}_\theta\, \mathbf{g} = \begin{bmatrix} 1 & Q \end{bmatrix} \begin{bmatrix} \mathrm{Var}(T_c) & \mathrm{Cov}(T_c, \rho) \\ \mathrm{Cov}(T_c, \rho) & \mathrm{Var}(\rho) \end{bmatrix} \begin{bmatrix} 1 \\ Q \end{bmatrix}.$$

Multiplying out,

$$\mathrm{Var}(T) = \mathrm{Var}(T_c) + Q^2\ \mathrm{Var}(\rho) + 2Q\ \mathrm{Cov}(T_c, \rho). \tag{4.54}$$

This is the variance budget of the output: one term per entry, each its sensitivity squared times its variance, plus a cross term from their covariance. The first term is the spread the supply air alone puts on the temperature; the second, the spread the resistance alone puts on it, $Q^2\ \mathrm{Var}(\rho)$; the third corrects for the two moving together in the posterior. If the readings cannot separate the two entries, the posterior leaves them correlated, and the cross term either adds to the diagonal terms or cancels part of them. With a negative correlation and same-signed sensitivities it is negative, and the output can be well known although each entry is poorly known: the readings then fix the combination $T_c + Q\rho$ that the output depends on, not the entries separately. Adding the two individual variances without the cross term is wrong in either direction, which is why the joint posterior, not its marginals, must be kept; Exercise 4.2 works a case in which ignoring the cross term overstates the spread by more than half.

4.3.3 The density of a transformed quantity

The delta method gives a derived quantity a mean and a variance. Some questions need more: the whole density of the quantity, or a tail probability far from the center, where the shape matters and a Gaussian with the right variance can still be wrong. The tool for these is the *change of variables*, and it is used throughout the book whenever a quantity is computed in one coordinate and reported in another.

Probability is conserved; density is not. Let x have the density $p_x(x)$, and let $u = h(x)$ be a monotone function with inverse $x = h^{-1}(u)$. A short interval $[x, x + \mathrm{d}x]$ is mapped onto a short interval $[u, u + \mathrm{d}u]$, and the two carry the same probability, because they hold the same outcomes:

$$p_u(u)\,|\mathrm{d}u| = p_x(x)\,|\mathrm{d}x|.$$

Dividing by $|\mathrm{d}u|$ gives the density of the transformed quantity,

$$p_u(u) = p_x\big(h^{-1}(u)\big)\ \left|\frac{\mathrm{d}x}{\mathrm{d}u}\right|. \tag{4.55}$$

The first factor evaluates the old density at the point that maps to u. The second, the *change-of-variable factor*, records how much the map stretches or squeezes the axis there: where a small change of u corresponds to a large change of x, the probability of the long x-interval is packed

into a short u-interval, and the density of u must be larger to hold it. Without the factor, $p_x(h^{-1}(u))$ is not a density of u at all: it does not in general integrate to one. A monotone map also carries quantiles over unchanged: if h is increasing, $\Pr(u \le h(x^*)) = \Pr(x \le x^*)$, so the median of u is h of the median of x, while the mode, which depends on the factor, moves.

The conductance. For the conductance $G = 1/\rho$ of the hall, the inverse map is $\rho = 1/G$ and $|\mathrm{d}\rho/\mathrm{d}G| = 1/G^2$, so equation (4.55) gives

$$p_G(G) = \frac{1}{G^2}\, p(\rho = 1/G \mid \mathbf{y}).$$

It is skewed to the right of $\hat{G}$ because the factor $1/G^2$ is larger at small G, which squeezes the left of the curve and stretches its right.

Two routes to a derived quantity. The delta method and the change of variables answer the same question with different approximations. For a scalar entry with posterior $x \mid \mathbf{y} \approx \mathcal{N}(\hat{x}, \sigma^2)$ and a monotone output $u = h(x)$:

- the *delta method* linearizes h at $\hat{x}$, with gradient $g = \mathrm{d}u/\mathrm{d}x$ evaluated there, and reports $u \mid \mathbf{y} \approx \mathcal{N}(h(\hat{x}),\ g^2\sigma^2)$, a Gaussian in u;
- the *change of variables* keeps h exact and reports $p_u(u) = p_x(h^{-1}(u))\,|\mathrm{d}x/\mathrm{d}u|$, with the Gaussian of x as its only approximation.

The two factors are reciprocals: the delta method's gradient $\mathrm{d}u/\mathrm{d}x$ is the inverse of the change-of-variable factor $\mathrm{d}x/\mathrm{d}u$, but the delta method evaluates it once, at $\hat{x}$, while the change of variables evaluates it at every u. The delta method is the change of variables with the map replaced by its tangent, which is why the two agree on the spread near the center and differ in the tails. When h is linear the factor is constant and the two coincide. When h bends across the posterior, the delta method puts a symmetric Gaussian on a quantity whose density is skewed, and can put probability where the quantity cannot go: below zero for a conductance or a rate, outside $(0, 1)$ for a probability such as a probability of detection.

> **Principle 4.1.** *The gradient carries a spread; the change-of-variable factor carries a density. Report $\sqrt{\mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g}}$ when a standard deviation is the answer, and transform the density when the answer is a shape, a tail probability or a bound near the edge of what the quantity can be.*

4.3.4 Credible intervals

A design review often needs an interval for an uncertain quantity as well as its mean. For the Gaussian marginal $T_2 \mid \mathbf{y} \sim \mathcal{N}(\hat{z}_3, \Sigma_{33})$ of equation (4.31) the central 95 % credible interval is the interval that leaves 2.5 % of the posterior in each tail. Since $\Phi(1.960) = 0.975$,

$$\left[\hat{z}_3 - 1.960\,\sqrt{\Sigma_{33}},\ \hat{z}_3 + 1.960\,\sqrt{\Sigma_{33}}\right]. \tag{4.56}$$

The same form serves any entry, with $\hat{z}_i$ and $\sqrt{\Sigma_{ii}}$, and a derived quantity, with $f(\hat{\mathbf{z}})$ and $\sqrt{\mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g}}$ from the delta method (4.53).

The interval and a threshold probability carry different information. The interval describes the range that holds most of the posterior probability; a threshold probability answers directly whether the plant meets a stated limit. For a release decision the threshold probability is usually the more actionable of the two. Neither is meaningful apart from the model assumptions, the quality of the measurement and the sources of uncertainty that the model includes; Section 4.6.6 lists those it leaves out.

4.3.5 The temperature-threshold probability

The question of equation (4.10) needs one marginal of the posterior. The marginal of the full-load temperature is read off the third entry of the state, equation (4.31): $T_2 \mid \mathbf{y} \sim \mathcal{N}(\hat{z}_3,\ \Sigma_{33})$.

The standardized temperature. For a Gaussian variable the standardized variable $(T_2 - \hat{z}_3)/\sqrt{\Sigma_{33}}$ has the standard normal distribution, and subtracting a constant and dividing by a positive constant preserve the inequality, so

$$\Pr(T_2 < T_{\max} \mid \mathbf{y}) = \Pr\!\left(\frac{T_2 - \hat{z}_3}{\sqrt{\Sigma_{33}}} < \frac{T_{\max} - \hat{z}_3}{\sqrt{\Sigma_{33}}} \;\Big|\; \mathbf{y}\right) = \Phi\!\left(\frac{T_{\max} - \hat{z}_3}{\sqrt{\Sigma_{33}}}\right). \tag{4.57}$$

The probability of exceeding the limit is its complement, $1 - \Pr(T_2 < T_{\max} \mid \mathbf{y})$.

The prior marginal. Before the test reading the state had the prior formed by the physical and prior rows alone. Its Jacobian $\mathbf{J}_0$ is $\mathbf{J}$ of equation (4.25) without the reading's line, its mean is the handbook state $\mathbf{z}^{(0)} = (0.160,\ 22.8,\ 27.6)^{\mathsf{T}}$, and its covariance $\mathbf{\Sigma}_0 = (\mathbf{J}_0^{\mathsf{T}}\mathbf{J}_0)^{-1}$ tends, by the argument of equation (4.35) with $\mathbf{J}_0\mathbf{v} = (0,\ 0,\ 40)^{\mathsf{T}}$ in place of $\mathbf{J}\mathbf{v}$, to $\mathbf{v}\mathbf{v}^{\mathsf{T}}/40^2 = \sigma_0^2\,\mathbf{v}\mathbf{v}^{\mathsf{T}}$. Its third diagonal entry is $(\mathbf{\Sigma}_0)_{33} = 60^2 \times 0.025^2 = 2.25\,\mathrm{K}^2$, so the full-load temperature had the prior marginal $\mathcal{N}(27.6,\ 1.5^2)$, the third entry of $\mathbf{z}^{(0)}$ with $\sqrt{(\mathbf{\Sigma}_0)_{33}} = 1.5\,\mathrm{K}$.

The question through the resistance. In the limit of exact rows the posterior lies on the line $\mathbf{z} = \hat{\mathbf{z}} + t\,\mathbf{v}$ through the MAP state along the free direction, $\mathbf{\Sigma} = \mathbf{v}\mathbf{v}^{\mathsf{T}}/5200$ by equation (4.35). On that line the third entry is $\hat{z}_3 + 60t$ and the first $\hat{z}_1 + t$, so the event $\{T_2 < T_{\max}\}$ is the event $\{t < t^*\}$ with $t^* = (T_{\max} - \hat{z}_3)/60$, which is the event that the resistance stays below

$$\rho^* = \hat{z}_1 + t^* = \hat{z}_1 + \frac{T_{\max} - \hat{z}_3}{60}. \tag{4.58}$$

With the marginal (4.31) of the first entry,

$$\Pr(\rho < \rho^* \mid \mathbf{y}) = \Phi\!\left(\frac{\rho^* - \hat{z}_1}{\sqrt{\Sigma_{11}}}\right) = \Phi\!\left(\frac{t^*}{\sqrt{\Sigma_{11}}}\right),$$

which is the probability (4.57): both standardized distances are t^* divided by the spread of t, since the third entry moves 60 times as far as the first along $\mathbf{v}$ and $\sqrt{\Sigma_{33}} = 60\sqrt{\Sigma_{11}}$ in the limit. Reading the third entry and propagating the first one along the free direction are the same calculation; the full-load question is a question about the resistance.

The dependence on the value read. The information matrix (4.27), and with it $\mathbf{\Sigma}$, does not contain the value observed, so a different reading leaves every standard deviation where it is and moves only the mean, along one vector. The reading enters the normal equations only through the part $-y_T/\sigma_T$ of the third entry of the residual vector that depends on y_T; that row's line of the Jacobian is $\mathbf{h}^{\mathsf{T}}/\sigma_T$, so differentiating $\hat{\mathbf{z}} = \mathbf{z} - \mathbf{\Sigma}\mathbf{J}^{\mathsf{T}}\mathbf{r}(\mathbf{z})$ with respect to y_T gives the gain vector

$$\frac{\partial \hat{\mathbf{z}}}{\partial y_T} = \frac{\mathbf{\Sigma}\mathbf{h}}{\sigma_T^2}, \tag{4.59}$$

the second column of $\mathbf{\Sigma}$ divided by σ_T^2. The threshold probability (4.57) is computed with a spread fixed before the reading and a mean set by its value, so the reading can move it in either direction.

4.3.6 The posterior state and the full-load temperature

General form. The three estimation questions, the resistance of the air path and the hall temperature at the test and at full load, share one form: the posterior of a quantity $f(\mathbf{z})$, equation (2.9), summarized by its moments (2.10) and its central interval (2.11).

Specialized form. Equations (2.10) and (2.11) with $f(\mathbf{z}) = z_i$, the entry asked about, and $p(\mathbf{z} \mid \mathbf{y})$ the Gaussian posterior of step 3: the mean is $\hat{z}_i$, the variance Σ_{ii}, by the marginals (4.31), and the interval is equation (4.56). The rows are linear and Gaussian, so these are the exact moments, with no approximation to set beside them.

Building blocks. The MAP state and the limit covariance (4.35), with its limit values of Section 4.2.10; the prior marginal $\mathcal{N}(27.6,\ 1.5^2)$ of Section 4.3.5, for the prior interval.

The posterior state. Table 4.6 collects it: each row is one entry of $\hat{\mathbf{z}}$, with its posterior mean and its standard deviation $\sqrt{\Sigma_{ii}}$, in the limit $\sigma_p \to 0$; the full-load spread is twice the test spread, the ratio 60/30 of the entries of $\mathbf{v}$.

Table 4.6: The posterior state of the equipment hall after the test reading, in the limit of exact physical rows.

State entry	Posterior mean	Standard deviation
ρ [K/kW]	0.16923	0.01387
T_1 [°C; sd in K]	23.077	0.416
T_2 [°C; sd in K]	28.154	0.832

The diagonal of the limit covariance. The limit covariance (4.35) gives $\Sigma_{11} = 1/5200 = 1.9231 \times 10^{-4}\,(\mathrm{K/kW})^2$, $\Sigma_{22} = 900/5200 = 0.17308\,\mathrm{K}^2$ ($\sqrt{\Sigma_{22}} = 0.4160\,\mathrm{K}$) and

$$\Sigma_{33} = \mathrm{Var}(T_2 \mid \mathbf{y}) = \frac{3600}{5200} = 0.69231\,\mathrm{K}^2, \qquad \sqrt{\Sigma_{33}} = 0.8321\,\mathrm{K}. \tag{4.60}$$

The credible interval. By equation (4.56), $28.1538 \pm 1.960 \times 0.8321 = 28.1538 \pm 1.6308$, so that

$$T_2 \in [26.52,\ 29.78]^\circ\mathrm{C}$$

with 95 % posterior probability under the stated model; the prior interval was $[24.66,\ 30.54]^\circ$C, 1.8 times as wide.

The conductance beside the resistance. The conductance is a derived quantity, $G(\mathbf{z}) = 1/z_1$, so the delta method (4.53) reads it from the same posterior, with

$$\mathbf{g} = \frac{\partial G}{\partial \mathbf{z}} = \begin{bmatrix} -1/\hat{z}_1^2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -34.92 \\ 0 \\ 0 \end{bmatrix} (\mathrm{kW/K})^2, \qquad \mathrm{Var}(G \mid \mathbf{y}) \approx \mathbf{g}^{\mathsf{T}} \boldsymbol{\Sigma} \mathbf{g} = \frac{\Sigma_{11}}{\hat{z}_1^4},$$

which gives $\hat{G} = 1/\hat{\rho} = 5.909\,\mathrm{kW/K}$ with $\mathrm{sd}(G \mid \mathbf{y}) \approx 0.484\,\mathrm{kW/K}$ and the 5–95 % band $[5.113,\ 6.706]\,\mathrm{kW/K}$. The exact band, by the change of variables (4.55), is $[5.207,\ 6.830]\,\mathrm{kW/K}$, higher by 0.095 and $0.124\,\mathrm{kW/K}$; the median is $5.909\,\mathrm{kW/K}$ on both routes.

4.3.7 The threshold probability at full load

General form. A requirement that a quantity stay below a limit is the posterior probability of the event, equation (2.12), compared with an assurance level κ.

Specialized form. Equation (2.12) with $f(\mathbf{z}) = -T_2 = -z_3$, $f^* = -T_{\max}$ and $p(\mathbf{z} \mid \mathbf{y})$ the Gaussian posterior of step 3 becomes the normal tail (4.57) of the third entry. The commissioning test states no assurance level κ; the probability is reported without an accept verdict. The posterior is exactly Gaussian, so the tail is the exact probability under the model.

Building blocks. The third entry $\hat{z}_3$ and the variance Σ_{33} of equation (4.60), and the standard normal distribution function Φ.

The probability. With $\hat{z}_3 = 28.1538\,^\circ\mathrm{C}$ and $\sqrt{\Sigma_{33}} = 0.8321\,\mathrm{K}$, the standardized distance in equation (4.57) is $(28 - 28.1538)/0.8321 = -0.1849$, and

$$\Pr(T_2 < 28\,^\circ\mathrm{C} \mid \mathbf{y}) = \Phi(-0.1849) = 0.4267. \tag{4.61}$$

The probability of exceeding the limit is $1 - 0.4267 = 0.5733$. The fit at $\sigma_p = 0.01\,\mathrm{K}$ gives 0.4267 as well, with $\sqrt{\Sigma_{33}} = 0.83223\,\mathrm{K}$ against the limit's $0.83205\,\mathrm{K}$.

Against the prior. Under the prior marginal $\mathcal{N}(27.6,\ 1.5^2)$ of Section 4.3.5, before the test reading, the probability that the full-load temperature stays below the limit was

$$\Pr(T_2 < 28) = \Phi\left(\frac{28 - 27.6}{1.5}\right) = \Phi(0.2667) = 0.6051.$$

Figure 4.5 draws the answer. The horizontal axis is the full-load temperature and the vertical axis its density in K^{-1}. The solid curve is the posterior marginal, the dashed curve the prior marginal, and the vertical line the limit $T_{\max}$. The shaded area under the posterior to the left of the line is the probability (4.61); the unshaded area under it to the right is the probability of exceeding the limit. The posterior is narrower than the prior and centered to the right of it, across the limit.

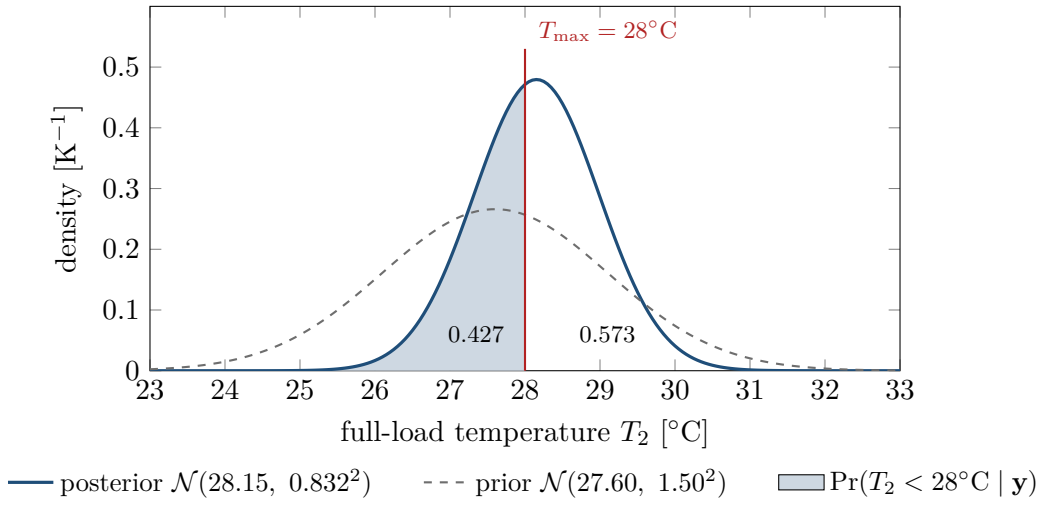


Figure 4.5: The threshold probability of equation (4.57). Posterior marginal of the full-load temperature, $\mathcal{N}(28.1538, 0.8321^2)$, with the area below $T_{\max} = 28\,^\circ\mathrm{C}$ shaded ($\Phi(-0.1849) = 0.4267$), and the prior marginal $\mathcal{N}(27.6, 1.5^2)$, under which the same probability is $\Phi(0.2667) = 0.6051$.

4.3.8 The scenario node

The design review asks one more kind of question of the same calibration, and every later chapter asks it: what the hall would do at a load it has not carried. The design load was carried in the state from the start, as the entry T_2 of Figure 4.2; a load of 70 or 80 kW was not, and adding an entry and a row for each load and fitting again would repeat the inference for a question that asks nothing new of the evidence. The device that answers it from the posterior already in hand is the *scenario node*, the state route of a what-if. The input the planner changes, here the load, is pinned at a value Q^* the hall has not seen; the physical rows are solved again at that value for the quantity asked about, the scenario temperature T^*; and the posterior is carried to the new node through its gradient, as the delta method carries it to any derived quantity. A star on a load, or on the quantity solved at it, marks a scenario quantity throughout the book; a starred threshold such as f^*, t^* or ρ^* remains a limit value, and the context tells the two apart.

General form. On the reduced state $\mathbf{z}_\mathrm{f}$ of Section 4.2.8, with its calibrated mean $\hat{\mathbf{z}}_\mathrm{f}$ and covariance $\mathbf{\Sigma}_\mathrm{red}$, the scenario node is

$$T^* = h(\mathbf{z}_\mathrm{f};\ Q^*), \qquad T^* \mid \mathbf{y} \approx \mathcal{N}\big(h(\hat{\mathbf{z}}_\mathrm{f};\ Q^*),\ (\mathbf{g}^*)^\mathsf{T}\mathbf{\Sigma}_\mathrm{red}\,\mathbf{g}^*\big), \qquad \mathbf{g}^* = \left.\frac{\partial h}{\partial \mathbf{z}_\mathrm{f}}\right|_{\hat{\mathbf{z}}_\mathrm{f}}, \tag{4.62}$$

where $h(\mathbf{z}_\mathrm{f};\ Q^*)$ is the solution of the physical rows for the scenario quantity at the free entries $\mathbf{z}_\mathrm{f}$ and the pinned input Q^*, and $\mathbf{g}^*$ its gradient with respect to the free entries at the calibrated state, a function of Q^*. The first relation is the plant solved at an input it has not seen; the second is the delta method (4.53) applied to that solution. The posterior itself is not touched: a scenario is one more function of the state pushed through the existing $\mathbf{\Sigma}_\mathrm{red}$, the division of Section 4.1.2. The form is exact when h is linear in the free entries and the posterior Gaussian, and a linearization otherwise, checked against draws in step 5.

Specialized form. In the hall the reduced state is the resistance ρ, and the physical row of a hall node at the load Q^*, $T^* - T_c - Q^*\rho = 0$, solved for the scenario temperature, gives equation (4.62) as

$$T^*(Q^*) = T_c + Q^*\rho, \qquad \frac{\partial T^*}{\partial \rho} = Q^*, \qquad \mathrm{sd}(T^* \mid \mathbf{y}) = Q^*\sqrt{\mathbf{\Sigma}_\mathrm{red}} = Q^* \times 0.013868\,\mathrm{K/kW}, \tag{4.63}$$

with Q^* in kW, the gradient in kW and $\sqrt{\mathbf{\Sigma}_\mathrm{red}}$ in K/kW. The row is linear in ρ, so the scenario node is exactly Gaussian, and the probability that it stays within the limit is the normal tail (4.57) at the node, $\Pr(T^* \le T_\mathrm{max} \mid \mathbf{y}) = \Phi\big((T_\mathrm{max} - T_c - Q^*\hat{\rho})/(Q^*\sqrt{\mathbf{\Sigma}_\mathrm{red}})\big)$. Under the handbook prior alone the same node has mean $T_c + Q^*\mu_0$ and standard deviation $Q^*\sigma_0$.

Building blocks. The calibrated resistance $\hat{\rho} = 0.169231\,\mathrm{K/kW}$ and its spread $\sqrt{\mathbf{\Sigma}_\mathrm{red}} = 0.013868\,\mathrm{K/kW}$ of Section 4.2.11; the handbook prior (4.12); the standard normal distribution function Φ.

The scenario loads. Table 4.7 applies equation (4.63) at six loads. Each row is one pinned load Q^*, from the 30 kW of the test to 80 kW in steps of 10 kW; the columns give the mean and the standard deviation of the scenario temperature under the calibration, the probability that it stays at or below 28°C, and the same probability under the handbook prior alone, before the test reading.

The spread of the scenario temperature grows in proportion to the load, from 0.4160 K at 30 kW to 1.1094 K at 80 kW, because the one uncertain entry, the resistance, is multiplied by

Table 4.7: The scenario node of the hall at six pinned loads, under the calibration (the posterior after the test reading, in the limit of exact physical rows) and under the handbook prior $\mathcal{N}(0.160,\ 0.025^2)$ alone. Probabilities of 1.0000 are rounded to four decimals.

Q^* [kW]	T^* [°C]	sd(T^*) [K]	Pr($T^* \le 28$°C), calibration [–]	Pr($T^* \le 28$°C), handbook prior [–]
30	23.0769	0.4160	1.0000	1.0000
40	24.7692	0.5547	1.0000	0.9998
50	26.4615	0.6934	0.9867	0.9452
60	28.1538	0.8321	0.4267	0.6051
70	29.8462	0.9707	0.0286	0.2464
80	31.5385	1.1094	0.0007	0.0808

the load. The calibration and the handbook prior part ways in both directions. At loads whose scenario temperature lies below the limit the narrower calibration is the more favorable, 0.9867 against 0.9452 at 50 kW; at loads above the design load the higher calibrated resistance is the less favorable, and at 70 kW the calibration gives 0.0286 against the handbook prior's 0.2464. The test reading moved the resistance up, and the scenario node carries that move to every load, the more strongly the farther the load lies from the test.

The design load as a scenario. The row at 60 kW is the entry T_2 of the state: 28.1538°C with standard deviation 0.8321 K, the values of equation (4.60), rounded to three decimals in Table 4.6, and the probability 0.4267 of equation (4.61); the row at 30 kW is likewise the entry T_1. A scenario node is therefore a state entry that the model did not need until the question asked for it: carried in the state, as T_2 was, it is fitted with the rest; added afterwards, it is read from the posterior already in hand, and the two give the same answer. Between 50 and 60 kW the probability falls through one half: the median of the scenario temperature meets the limit at 59.09 kW, just below the design load, the load that Chapter 5 names the binding load and derives with its band.

4.4 The checks (step 5)

Obtaining a posterior is not the same as establishing that it is credible.

4.4.1 Algebraic and physical verification

Algebraic verification. At the fitted state of equation (4.49), with $\sigma_p = 0.01$ K, the residual vector (4.23) is $\mathbf{r}(\hat{\mathbf{z}}) = (0.0049225,\ 0,\ -0.24612,\ 0.36919)^{\mathsf{T}}$, and the gradient (4.28), column by column of $\mathbf{J}$,

$$\mathbf{J}^{\mathsf{T}}\mathbf{r}(\hat{\mathbf{z}}) = \begin{bmatrix} -3000 \times 0.0049225 + 40 \times 0.36919 \\ 100 \times 0.0049225 + 2 \times (-0.24612) \\ 100 \times 0 \end{bmatrix} = \begin{bmatrix} -14.767 + 14.767 \\ 0.4922 - 0.4922 \\ 0 \end{bmatrix} = \mathbf{0},$$

to numerical precision: the state satisfies the first-order optimality condition. Each line reads as a balance of rows on one entry: the prior's pull on ρ is held by the first physical row, which in turn is held at T_1 by the reading row; nothing pulls on T_2 but its own physical row. The information matrix (4.27) has positive eigenvalues, so the objective is strictly convex and the stationary point is its unique minimum.

Physical consistency. The physical rows at the fitted state, in kelvin, are $\sigma_p r_1 = 4.9 \times 10^{-5}\,\mathrm{K}$ and $\sigma_p r_2 = 0$: both halls balance to well within the tolerance, and the first imbalance shrinks with σ_p. The difference between the observed and the fitted test temperature is $23.2 - 23.0769 = 0.1231\,\mathrm{K}$, a quarter of the assumed sensor standard deviation of 0.5 K. The model is consistent with this reading. Agreement with one reading does not validate the physical model: with one reading and one free direction, the fit has nearly enough freedom to reproduce any value.

4.4.2 Prior predictive check

Before the reading was used, the model and the prior predicted a distribution for it. By equation (4.11), $Y = \mathbf{h}^\mathsf{T}\mathbf{z} + \epsilon$, with the state under its prior $\mathcal{N}(\mathbf{z}^{(0)}, \mathbf{\Sigma}_0)$ of Section 4.3.5 and ϵ independent of it, so Y is Gaussian with

$$\mathbb{E}[Y] = \mathbf{h}^\mathsf{T}\mathbf{z}^{(0)} = 22.8^\circ\mathrm{C}, \qquad \mathrm{Var}(Y) = \mathbf{h}^\mathsf{T}\mathbf{\Sigma}_0\mathbf{h} + \sigma_T^2 = (\mathbf{\Sigma}_0)_{22} + \sigma_T^2 = 0.5625 + 0.25 = 0.8125\,\mathrm{K}^2,$$

the variances adding because the two terms are independent, and $(\mathbf{\Sigma}_0)_{22} = 30^2 \times 0.025^2$ in the limit of exact rows. Thus

$$Y \sim \mathcal{N}(22.8,\ 0.9014^2). \tag{4.64}$$

The observed reading lies $(23.2 - 22.8)/0.9014 = 0.444$ standard deviations from the predicted mean, an unremarkable distance: the reading is compatible with the prior information and the model. This check is distinct from computing the posterior. It asks whether the evidence was plausible under the assumptions before those assumptions were used to interpret it; a reading several predictive standard deviations away would have signaled a wrong prior, a wrong model or a faulty sensor before any posterior was formed.

4.4.3 The evidence by two routes

The evidence (4.18) of the model is the prior predictive density (4.64) evaluated at the reading,

$$p(y_T \mid M) = \mathcal{N}(y_T;\ \mathbf{h}^\mathsf{T}\mathbf{z}^{(0)},\ \mathbf{h}^\mathsf{T}\mathbf{\Sigma}_0\mathbf{h} + \sigma_T^2),$$

with $\mathcal{N}(y; m, v)$ the Gaussian density of Section 4.2.1. With $m = 22.8^\circ\mathrm{C}$ and $v = 0.8125\,\mathrm{K}^2$,

$$p(y_T = 23.2 \mid M) = \frac{1}{\sqrt{2\pi \times 0.8125}} \exp\left[-\frac{(23.2 - 22.8)^2}{2 \times 0.8125}\right] = 0.44259 \times 0.90623 = 0.4011\,\mathrm{K}^{-1}. \tag{4.65}$$

The Laplace form (4.19) gives the same number from the fit. At $\sigma_p = 0.01\,\mathrm{K}$, $\det(\mathbf{\Sigma})^{1/2} = 1.3867 \times 10^{-6}$ and $\mathcal{L}(\hat{\mathbf{z}}) = 0.09845$ (Table 4.4), so

$$p(y_T \mid M) = 0.39894 \times \frac{\det(\mathbf{\Sigma})^{1/2}}{\sigma_p^2} \times \frac{\mathrm{e}^{-\mathcal{L}(\hat{\mathbf{z}})}}{\sigma_T\,\sigma_0} = 0.39894 \times 0.013867 \times 0.90624 \times 80 = 0.4011\,\mathrm{K}^{-1},$$

with $\det(\mathbf{\Sigma})^{1/2}/\sigma_p^2 = 0.013867$ against its limit $1/\sqrt{5200} = 0.013868$, the width of the posterior along the free direction.

Common mistake

Reading an evidence value as a probability that the model is correct. The density $0.4011\,\mathrm{K}^{-1}$ is not a 40 % confidence in the thermal model: it is a density, in inverse units of the reading, whose value changes by a factor of 1000 if the reading is expressed in millikelvin. A posterior probability of a model requires the evidence of the competing models and their prior probabilities. Likewise, the posterior probability 0.4267 is a confidence in the claim about T_2 under the model, while the evidence ratio is a confidence in the model the claim was computed under; the two answer different questions.

4.4.4 Posterior predictive validation with a new observation

Suppose a further, independent reading is to be taken at the same 30 kW condition. Before its value is seen, its distribution given the evidence already in hand is the posterior predictive. The new reading is $Y_{\text{new}} = \mathbf{h}^{\mathsf{T}}\mathbf{z} + \epsilon_{\text{new}}$, with $\mathbf{z} \mid \mathbf{y} \sim \mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$ and $\epsilon_{\text{new}} \sim \mathcal{N}(0, \sigma_T^2)$ independent of it, so

$$Y_{\text{new}} \mid \mathbf{y} \sim \mathcal{N}(\mathbf{h}^{\mathsf{T}}\hat{\mathbf{z}},\ \mathbf{h}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{h} + \sigma_T^2) = \mathcal{N}(23.0769,\ 0.17308 + 0.25) = \mathcal{N}(23.0769,\ 0.6504^2), \quad (4.66)$$

with $\mathbf{h}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{h} = \Sigma_{22}$ from equation (4.35). The predictive variance has two parts: the remaining uncertainty about the test temperature, $0.173\,\text{K}^2$, and the error of the new reading itself, $0.25\,\text{K}^2$. A further reading of the same sensor can therefore never be predicted more precisely than σ_T, however well the resistance becomes known.

An independent new reading compared with this distribution assesses the predictive performance of the model: a reading a fraction of a predictive standard deviation from 23.0769°C passes, and one several standard deviations away does not. Repeated large discrepancies, or a systematic sign in them, would indicate that the model, the prior, the measurement-error assumptions or the representation of the operating condition need revision.

4.4.5 The limit of exact rows

Table 4.8 repeats the fit at three tolerances and in the limit $\sigma_p \to 0$. Each column is one tolerance; the rows give the MAP resistance and full-load temperature, their standard deviations, the correlation of the two hall temperatures, the threshold probability of Section 4.3.5, and the condition number of the information matrix. By $\sigma_p = 0.01\,\text{K}$ the full-load temperature and the threshold probability have stopped moving at the printed precision; the resistance and its spread still move in their sixth decimal, the full-load spread in its fourth; at $\sigma_p = 0.001\,\text{K}$ all have stopped. The fit at $\sigma_p = 0.001\,\text{K}$ reproduces the limit values of Section 4.2.10 to every printed digit.

Table 4.8: The fit of the hall state as the model tolerance of the physical rows shrinks. The last column is the limit of exact rows.

Quantity	$\sigma_p = 0.1\,\text{K}$	$0.01\,\text{K}$	$0.001\,\text{K}$	$\sigma_p \to 0$
$\hat{\rho}$ [K/kW]	0.169119	0.169230	0.169231	0.169231
$\sqrt{\Sigma_{11}}$ [K/kW]	0.014056	0.013869	0.013868	0.013868
$\hat{T}_2$ [°C]	28.1471	28.1538	28.1538	28.1538
$\sqrt{\Sigma_{33}}$ [K]	0.84926	0.83223	0.83205	0.83205
correlation of T_1, T_2 [–]	0.96522	0.99964	0.999996	1
$\Pr(T_2 < 28°\text{C})$ [–]	0.4312	0.4267	0.4267	0.4267
condition number of $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ [–]	4.0×10^5	3.9×10^7	3.9×10^9	–

The $\sqrt{\Sigma_{11}}$ row widens with the tolerance as equation (4.33) requires, the tolerance acting on the reading as extra sensor variance, and the limit column is the matrix of equation (4.35); its entry for $\sqrt{\Sigma_{11}}$ is the square root of $\boldsymbol{\Sigma}_{\text{red}}$, the covariance of the reduced state (Section 4.2.11). The fitted covariances satisfy equation (4.34) at every tolerance of the table ($0.69260 = 0.69250 + 0.0001$ at $\sigma_p = 0.01\,\text{K}$, $0.72125 = 0.71125 + 0.01$ at $0.1\,\text{K}$). A tighter tolerance than $0.001\,\text{K}$ buys nothing and costs conditioning: the condition number grows a hundredfold per decade, and at $\sigma_p = 10^{-6}\,\text{K}$ it reaches 3.9×10^{15}: round-off then moves the fit in its fifth digit from one arithmetic library to another, so that the fit there is good to three significant figures only, $\hat{\rho} = 0.169\,\text{K/kW}$, $\sqrt{\Sigma_{33}} = 0.832\,\text{K}$ and $\Pr(T_2 < 28°\text{C}) = 0.427$, and adds nothing to the limit column.

4.4.6 Second routes to the full-load answers

The propagated full-load temperature. As the third entry, with $\mathbf{g} = (0,\ 0,\ 1)^{\mathsf{T}}$, the full-load temperature has $\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g} = \Sigma_{33} = 0.69231\,\mathrm{K}^2$; as the prediction $18{+}60\rho$, with $\mathbf{g} = (60,\ 0,\ 0)^{\mathsf{T}}$, it has $\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g} = 60^2\,\Sigma_{11} = 3600/5200 = 0.69231\,\mathrm{K}^2$, the same number, as Section 4.3.1 requires in the limit.

The threshold through the resistance. By equation (4.58), $t^* = (28 - 28.1538)/60 = -0.0025641\,\mathrm{K/kW}$ and

$$\rho^* = \hat{z}_1 + t^* = 0.169231 - 0.002564 = 0.166667\,\mathrm{K/kW},$$
$$\Pr(\rho < \rho^* \mid \mathbf{y}) = \Phi\left(\frac{-0.002564}{0.013868}\right) = \Phi(-0.1849) = 0.4267,$$

the number of equation (4.61).

4.4.7 The propagation check against draws

For a linear output the delta method is exact, and the check is the algebraic agreement of the entry and the gradient already shown in Section 4.4.6. For a nonlinear output the linearized result must be compared with the full nonlinear evaluation at draws from the posterior. The procedure is the following.

1. Factor the covariance as $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$ with $\mathbf{L}$ lower triangular (the Cholesky factor, which exists at any finite tolerance), and draw N independent states $\mathbf{z}^{(j)} = \hat{\mathbf{z}} + \mathbf{L}\boldsymbol{\varepsilon}^{(j)}$, $j = 1, \ldots, N$, each $\boldsymbol{\varepsilon}^{(j)}$ a vector of three independent standard normal numbers, so that $\mathbf{z}^{(j)} \sim \mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$. At $\sigma_p = 0.01\,\mathrm{K}$,
$$\mathbf{L} = \begin{bmatrix} 0.013869 & 0 & 0 \\ 0.41592 & 0.0099980 & 0 \\ 0.83217 & 0 & 0.010000 \end{bmatrix},$$
whose first column is $\sqrt{\Sigma_{11}}$ times (nearly) $\mathbf{v}$ and whose other two columns have the width σ_p of the physical rows. (In the limit of exact rows the factor has the single column $\mathbf{v}/\sqrt{5200}$, and a draw is $\hat{\mathbf{z}} + \mathbf{v}\,\varepsilon^{(j)}/\sqrt{5200}$, along the free direction.)
2. Evaluate the output at each drawn state through the full model, for example $T_2^{(j)} = z_3^{(j)}$ or $G^{(j)} = 1/z_1^{(j)}$.
3. Estimate a threshold probability as the fraction of draws satisfying the event, $\hat{P} = N^{-1}\sum_j \mathbf{1}[f(\mathbf{z}^{(j)}) < f^*]$, whose standard error is $\sqrt{\hat{P}(1-\hat{P})/N}$; estimate quantiles as the empirical quantiles of the $f(\mathbf{z}^{(j)})$.
4. Compare with the closed forms of the delta method. Agreement within the standard error means the closed form can be trusted.

A few hundred draws resolve a probability near one half to about ± 0.03; $N = 10^5$ resolves it to about ± 0.002.

Table 4.9 applies the procedure to the full-load temperature, which is linear in ρ, and to the conductance $G = 1/\rho$, which is not. The first column names the quantity; the second gives the delta-method value, from a Gaussian with mean $f(\hat{\mathbf{z}})$ and standard deviation $\sqrt{\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}}$; the third gives the exact value of the push-forward of the posterior, the limit to which the draw estimates converge, computed here in closed form because both outputs are monotone in ρ (the q-quantile of G is $1/(\hat{\rho} - z_q\sqrt{\Sigma_{11}})$ with z_q the standard normal quantile); the fourth gives one run of $N = 10^5$ drawn states at $\sigma_p = 0.01\,\mathrm{K}$, generated with a seeded pseudo-random generator so that the run can be repeated exactly. The conductance carries the gradient of Section 4.3.6.

For the temperature the three columns agree to within the sampling error, and the closed form needs no correction. For the conductance the median is right in all three columns, since

Table 4.9: The propagation check: delta method against the exact push-forward of the posterior, whose first entry is $\mathcal{N}(0.169231, 0.013868^2)$, and one run of 10^5 states drawn as $\hat{\mathbf{z}} + \mathbf{L}\boldsymbol{\varepsilon}$ (standard error of a probability about 0.0015). The threshold $G = 6.25\,\mathrm{kW/K}$ is the handbook conductance.

Quantity	Delta method	Exact	10^5 draws
T_2, median [°C]	28.154	28.154	28.152
T_2, 5 % and 95 % [°C]	26.785, 29.522	26.785, 29.522	26.786, 29.516
$\Pr(T_2 < 28°\mathrm{C})$ [–]	0.4267	0.4267	0.4269
G, median [kW/K]	5.909	5.909	5.910
G, 5 % and 95 % [kW/K]	5.113, 6.706	5.207, 6.830	5.210, 6.828
$\Pr(G > 6.25\,\mathrm{kW/K})$ [–]	0.2407	0.2528	0.2527

$1/\rho$ is monotone and maps the median of ρ to the median of G, but the delta-method band is shifted: the exact band $[5.207, 6.830]$ is skewed to the right of $\hat{G}$, while the linearized band $[5.113, 6.706]$ is symmetric about it, and the delta method understates the probability that the conductance exceeds its handbook value by about 0.012. The draws reproduce the exact column to within their standard error. Figure 4.6 shows the cause. The horizontal axis is the conductance and the vertical axis its density; the solid curve is the exact density of $G = 1/\rho$, obtained by the change of variables $p_G(G) = p(\rho = 1/G \mid \mathbf{y})/G^2$ of Section 4.3.3, and the dashed curve the delta-method Gaussian. The exact density has its mode to the left of 5.909 and a longer right tail, and the Gaussian cannot represent the asymmetry.

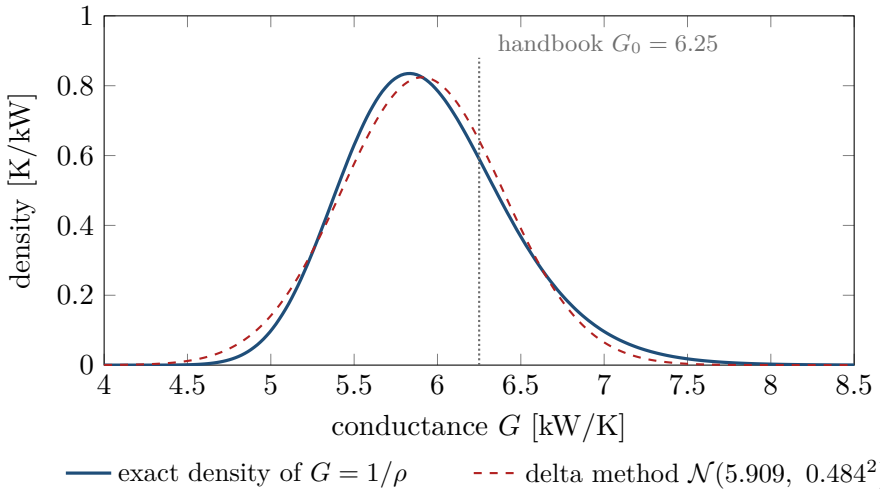


Figure 4.6: Posterior density of the conductance $G = 1/\rho$ under $\rho \mid \mathbf{y} \sim \mathcal{N}(0.169231, 0.013868^2)$: exact, by change of variables, $p_G(G) = \varphi((1/G - 0.169231)/0.013868)/(0.013868\,G^2)$ with φ the standard normal density; and the delta-method Gaussian $\mathcal{N}(1/\hat{\rho}, \Sigma_{11}/\hat{\rho}^4)$. Dotted line: the handbook conductance.

The conductance also shows why "the posterior mean of G" must be handled with care. Under a Gaussian posterior for ρ, the mean of $1/\rho$ does not exist in the strict sense, because the Gaussian has positive density at $\rho = 0$, where $1/\rho$ is not integrable; the mass involved here is about 10^{-34} and has no practical effect, but it is a reminder that moments of a transformed quantity are properties of the parameterization. Quantiles and threshold probabilities, which the push-forward preserves exactly for a monotone output, are the robust summaries, and they are what the check compares.

The scenario node of Section 4.3.8 passes the same check. Carried to the node at $Q^* = 70\,\mathrm{kW}$

by $T^{*(j)} = T_c + 70\, z_1^{(j)}$, the same 10^5 drawn states give a median of 29.845°C and $\Pr(T^* \le 28°\text{C}) = 0.0285$, against the 29.846°C and 0.0286 of equation (4.63) in Table 4.7: the scenario read from the posterior is the scenario that the drawn plants, each solved at 70 kW, produce.

4.4.8 Sensitivity to the value read

Table 4.10 repeats the assessment for other values of the single test reading. The first numeric column is the prior; each further column is the posterior after a different value of the reading, the actual one, 23.2°C, among them. The rows give the MAP resistance from equation (4.29), the mean and standard deviation of the full-load temperature, and the probability (4.57). The standard deviation is the same for every reading, and the mean moves along the gain vector (4.59), which with the limit covariance (4.35) is

$$\frac{\partial \hat{\mathbf{z}}}{\partial y_T} = \frac{\boldsymbol{\Sigma}\mathbf{h}}{\sigma_T^2} = 4 \begin{bmatrix} 0.0057692 \\ 0.17308 \\ 0.34615 \end{bmatrix} = \begin{bmatrix} 0.023077 \\ 0.69231 \\ 1.38462 \end{bmatrix},$$

four times the second column of the limit covariance, in K/kW per K and K per K. One column of the table, for $y_T = 22.4$°C, is the actual state moved by -0.8 K along it: $\hat{\rho} = 0.169231 - 0.8 \times 0.023077 = 0.150769$, $\mathbb{E}[T_2 \mid \mathbf{y}] = 28.1538 - 0.8 \times 1.38462 = 27.046$°C, and $\Phi((28 - 27.046)/0.8321) = \Phi(1.1464) = 0.8742$.

Table 4.10: The full-load assessment before the test and after a single test reading of various values ($\sigma_T = 0.5$ K, prior $\mathcal{N}(0.160, 0.025^2)$). The column 23.2°C is the observed reading.

Quantity	Prior	$y_T = 22.4$	22.8	23.2	23.6	24.0
$\hat{\rho}$ [K/kW]	0.16000	0.15077	0.16000	0.16923	0.17846	0.18769
$\mathbb{E}[T_2]$ [°C]	27.60	27.05	27.60	28.15	28.71	29.26
$\mathrm{sd}(T_2)$ [K]	1.500	0.832	0.832	0.832	0.832	0.832
$\Pr(T_2 < 28°\text{C})$ [–]	0.6051	0.8742	0.6846	0.4267	0.1975	0.0647

A reading of 22.8°C, equal to what the prior predicted, leaves the mean where it was and only narrows the distribution; because the mean then lies below the limit, narrowing alone raises the probability from 0.605 to 0.685. The observed 23.2°C moves the mean across the limit and lowers the probability to 0.427. The posterior probability equals the prior one for a reading of 22.93°C and falls below one half for any reading above 23.09°C.

4.5 The report (step 6)

One reading at the test condition overturns the handbook's answer. The handbook resistance put the hall at 27.6°C at full load, under the limit, with probability 0.6051 of staying below it; the reading of 23.2°C moves the resistance up, from 0.160 to 0.169 K/kW, and leaves the probability at 0.4267, near even odds, although it narrows the full-load spread from 1.5 to 0.832 K. The answers to the questions of Table 4.2 are given in the units in which the questions were asked, conditional on the steady-state model, the test reading and the handbook prior; the release decision follows from them.

4.5.1 The answers and their bands

Table 4.11 gives the answers. Each row is a question; the second column gives the answer in the limit of exact physical rows, and the third its band: the posterior standard deviation of an estimate, with the central 95 % credible interval of the full-load temperature, for the full-load

probability the range over the model tolerances of Table 4.8, and for the scenario loads the standard deviation of the scenario temperature with the draws check.

Table 4.11: The report for the equipment hall at 60 kW and at the scenario loads, after the test reading of 23.2°C. Answers in the limit of exact physical rows; bands are posterior standard deviations, a credible interval, the range over model tolerances σ_p from 0 to 0.1 K, or the draws check.

Planner's question	Answer	Band
What is the thermal resistance of the hall's air path?	0.16923 [K/kW]; conductance 5.909 [kW/K]	standard deviation 0.01387 [K/kW]; conductance 5–95 % band 5.113 to 6.706 [kW/K], exact 5.207 to 6.830 [kW/K]
How warm is the hall at the 30 kW test?	23.077 [°C]	standard deviation 0.416 [K]
How warm will the hall be at 60 kW?	28.15 [°C]	standard deviation 0.83 [K]; 26.52 to 29.78 [°C], central 95 % interval
Can the hall take 60 kW and stay below 28°C?	probability 0.427 [–]; of exceeding, 0.573 [–]	0.4267 to 0.4312 [–] over the model tolerances
Would the hall stay below 28°C at a load it has not yet carried?	probability 0.9867 [–] at 50 kW, 0.0286 [–] at 70 kW (Table 4.7)	standard deviation of T^* 0.6934 and 0.9707 [K]; 0.0285 [–] at 70 kW by draws

The observed temperature was slightly warmer than the prior model predicted (22.8°C, the second entry of $\mathbf{z}^{(0)}$), so the inferred resistance rose above its handbook value and the implied conductance fell below 6.25 kW/K. The inferred test temperature is not the observed 23.2°C: the reading is noisy, and the prior, carried to T_1 by the physical row, also contributes information. The conductance in the first row is $1/\hat{\rho}$, the conductance evaluated at the estimated resistance; it is not the posterior mean of the conductance (Section 4.4.7). With correlations above 0.9996 the three entries are known almost only as one combination, the free direction $\mathbf{v}$, and the full-load interval contains temperatures both below and above the limit.

4.5.2 The release decision

Conditional on the steady-state model, the reading and the prior, the estimated full-load temperature lies slightly above the limit, with substantial uncertainty on both sides of it. The reading reduced the uncertainty and, at the same time, lowered the probability of staying within the acceptance limit, from 0.605 before it to 0.427. The evidence does not support a high-confidence claim that the design load keeps the hall within the limit: the posterior probability of doing so is about 0.43. A decision to release the hall for its full design load would require an assurance level κ, stated explicitly, and a consideration of the consequences of exceeding the limit. Further measurements or more cooling capacity may be appropriate, but their value should be computed, not assumed. That is what an inference report gives the design review: the physical interpretation of a mathematical result, its uncertainty, and its implications for a decision.

Table 4.10 shows that the direction of the change was set by the value read, not by the act of reading. New evidence does not necessarily make the plant look safer: it can make it look less safe, or move the probability in either direction, depending on the value observed. The purpose of inference is to improve the knowledge of the plant, not to produce reassuring probabilities.

Principle 4.2. *A reading narrows the posterior by an amount fixed before it is taken, but the direction in which it moves a threshold probability is set by the value it returns. Evidence improves knowledge; it does not promise reassurance.*

4.6 The next measurement (step 7)

The report leaves the release question near even odds, with a band that a single reading at the test condition has set. A further measurement before the hall is committed to its design load is one more row of the residual vector, and the information it adds is known before its value is.

4.6.1 The preposterior covariance

A further reading at the test condition is the row $(T_1 - y)/\sigma_T$, whose line of the Jacobian is $\mathbf{j} = (0,\ 2,\ 0)$, the line of the first reading; a reading at full load, by a sensor of the same spread, is the row $(T_2 - y)/\sigma_T$, with line $\mathbf{j}_\mathrm{f} = (0,\ 0,\ 2)$. The information that a reading adds, $\mathbf{j}^\mathsf{T}\mathbf{j}$, is known before the reading is taken, because its Jacobian depends on the sensor and on where it is placed, not on the value. For a candidate reading a with information $\mathbf{\Lambda}_a$ the preposterior covariance is

$$\mathbf{\Sigma}_a = (\mathbf{\Sigma}^{-1} + \mathbf{\Lambda}_a)^{-1}, \tag{4.67}$$

formed at a finite tolerance, where $\mathbf{\Sigma}^{-1} = \mathbf{J}^\mathsf{T}\mathbf{J}$ exists, the object on which candidate readings are ranked. In the limit of exact rows the posterior stays on the free direction, and a candidate adds to the information $\|\mathbf{J}\mathbf{v}\|^2 = 5200$ along it the square of its own line times $\mathbf{v}$: $(\mathbf{j}\mathbf{v})^2 = 60^2 = 3600$ for a reading at 30 kW and $(\mathbf{j}_\mathrm{f}\mathbf{v})^2 = 120^2 = 14\,400$ for a reading at 60 kW. The full-load standard deviation after the candidate is 60 divided by the square root of the total. For n further readings at 30 kW, $\mathbf{\Lambda}_a = n\,\mathbf{j}^\mathsf{T}\mathbf{j}$, which adds $4n$ to the (T_1, T_1) entry, and the third diagonal entry of $\mathbf{\Sigma}_a$ gives the standard deviation of the full-load temperature: 0.640 K for $n = 1$, 0.539 K for $n = 2$, 0.429 K for $n = 4$ and 0.325 K for $n = 8$. More readings resolve the question; they do not decide its answer. If the full-load temperature truly lies above 28°C, the probability tends to zero as the readings accumulate, and the decision then rests on the air path, not on the readings.

4.6.2 The candidate measurements

Table 4.12 ranks three candidates. Each row is a measurement the planner can commission before the release; the columns give the information along $\mathbf{v}$ after it, the standard deviations of the resistance and of the full-load temperature that the preposterior covariance (4.67) predicts, and the width it removes from the current full-load standard deviation of 0.8321 K. One reading at full load carries four times the information of one at the test condition, because the full-load temperature moves twice as far as the test temperature along $\mathbf{v}$, and it removes exactly the width of four test readings.

Table 4.12: Candidate readings for the hall, in the limit of exact physical rows, from the posterior after the test reading. Each reading has $\sigma_T = 0.5$ K; the width removed is taken from the current $\sqrt{\Sigma_{33}} = 0.8321$ K.

Candidate	Information along $\mathbf{v}$ [(kW/K)^2]	sd(ρ) after [K/kW]	sd(T_2) after [K]	Width removed [K]
One further reading at 30 kW	8800	0.01066	0.640	0.192
Four further readings at 30 kW	19 600	0.00714	0.429	0.403
One reading at 60 kW	19 600	0.00714	0.429	0.403

4.6.3 The recommendation

Four further readings at the test condition and one reading at full load narrow the answer equally, and the choice between them rests on what else each tests. Readings at 30 kW need no

test at the design load, but repeated readings of one sensor share its bias, and no number of them can show whether the resistance at 60 kW is the resistance at 30 kW. A reading at full load tests exactly that assumption, on which the prediction rests (Section 4.6.6), at the cost of running the hall at the load in question. The next measurement is therefore one reading at full load, taken in a supervised load-bank test before the release, the load raised in steps and held at 60 kW until the hall settles, which the test engineer can stop if the hall approaches its limit; where such a test is not acceptable, it is four further readings at the test condition, by more than one sensor where possible. Chapter 18 ranks the tests and surveys a planner can commission on the site plant in the same way, by the width each takes off the quantity that decides the build.

4.6.4 A second observation as an update

Before any test at full load, the first candidate of Table 4.12 arrives: a second, independent reading at the test condition, $y_{T,2} = 23.0^{\circ}\mathrm{C}$, with the same standard deviation $\sigma_T = 0.5\,\mathrm{K}$; the first reading is now written $y_{T,1} = 23.2^{\circ}\mathrm{C}$. Against the posterior predictive (4.66) it lies $(23.0 - 23.0769)/0.6504 = -0.12$ standard deviations from the predicted mean and passes that check. The posterior after the first reading serves as the prior for the second.

The precision form. Information from independent rows adds. The new reading is one more row of the residual vector, $(T_1 - y_{T,2})/\sigma_T$, whose line of the Jacobian is $\mathbf{j} = (0,\ 2,\ 0)$, the same as the first reading's. Its contribution to the information matrix is the outer product $\mathbf{j}^{\mathsf{T}}\mathbf{j}$, which changes the (T_1, T_1) entry alone,

$$\boldsymbol{\Lambda}_2 = \mathbf{J}^{\mathsf{T}}\mathbf{J} + \mathbf{j}^{\mathsf{T}}\mathbf{j} = \mathbf{J}^{\mathsf{T}}\mathbf{J} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_2 = \boldsymbol{\Lambda}_2^{-1}.$$

The old rows are balanced at $\hat{\mathbf{z}}$, $\mathbf{J}^{\mathsf{T}}\mathbf{r}(\hat{\mathbf{z}}) = \mathbf{0}$, so the gradient of the enlarged objective at $\hat{\mathbf{z}}$ comes from the new row alone, $\mathbf{j}^{\mathsf{T}}(23.0769 - 23.0)/0.5 = (0,\ 0.30769,\ 0)^{\mathsf{T}}$, and one Gauss–Newton step (4.29) from $\hat{\mathbf{z}}$ gives $\hat{\mathbf{z}}_2 = \hat{\mathbf{z}} - 0.30769\,(\boldsymbol{\Sigma}_2)_{\cdot 2}$, with $(\boldsymbol{\Sigma}_2)_{\cdot 2}$ the second column of $\boldsymbol{\Sigma}_2$. In the limit of exact rows this column is $(3.4091 \times 10^{-3},\ 0.10227,\ 0.20455)^{\mathsf{T}}$, and

$$\hat{\mathbf{z}}_2 = \begin{bmatrix} 0.169231 \\ 23.0769 \\ 28.1538 \end{bmatrix} - 0.30769 \begin{bmatrix} 0.0034091 \\ 0.10227 \\ 0.20455 \end{bmatrix} = \begin{bmatrix} 0.168182 \\ 23.0455 \\ 28.0909 \end{bmatrix}.$$

The first entry of $\boldsymbol{\Sigma}_2$ is $1.1364 \times 10^{-4}\,(\mathrm{K/kW})^2$, so that $\sqrt{(\boldsymbol{\Sigma}_2)_{11}} = 0.010660\,\mathrm{K/kW}$.

The rank-one form. The same result follows from the first posterior alone, without forming and inverting $\boldsymbol{\Lambda}_2$. The second reading measures $\mathbf{h}^{\mathsf{T}}\mathbf{z} = T_1$ with $\mathbf{h} = (0,\ 1,\ 0)^{\mathsf{T}}$; its predicted value under the first posterior is $\mathbf{h}^{\mathsf{T}}\hat{\mathbf{z}} = 23.0769^{\circ}\mathrm{C}$ and its predictive variance, from equation (4.66), is $S = \mathbf{h}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{h} + \sigma_T^2 = \Sigma_{22} + \sigma_T^2 = 0.42308\,\mathrm{K}^2$. Conditioning the Gaussian on the new row gives the update

$$\hat{\mathbf{z}}_2 = \hat{\mathbf{z}} + \mathbf{K}\,(y_{T,2} - \mathbf{h}^{\mathsf{T}}\hat{\mathbf{z}}), \qquad \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} - S\,\mathbf{K}\mathbf{K}^{\mathsf{T}}, \qquad \mathbf{K} = \frac{\boldsymbol{\Sigma}\mathbf{h}}{S}, \tag{4.68}$$

where $\mathbf{K}$ is the gain vector, one entry per state entry, and $y_{T,2} - \mathbf{h}^{\mathsf{T}}\hat{\mathbf{z}}$ the innovation, the part of the reading the first posterior did not predict. Because $\boldsymbol{\Sigma}_2\mathbf{h} = \boldsymbol{\Sigma}\mathbf{h} - \mathbf{K}\,\mathbf{h}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{h} = (\sigma_T^2/S)\,\boldsymbol{\Sigma}\mathbf{h}$, the gain is also $\mathbf{K} = \boldsymbol{\Sigma}_2\mathbf{h}/\sigma_T^2$, the same vector as the gain (4.59) of the posterior that includes the new reading. The gain follows from writing the posterior mean as $\hat{\mathbf{z}} + \mathrm{Cov}(\mathbf{z}, Y_{\mathrm{new}})\,(y_{T,2} - \mathbb{E}[Y_{\mathrm{new}}])/S$

for the jointly Gaussian state and new reading, with $\mathrm{Cov}(\mathbf{z}, Y_{\mathrm{new}}) = \boldsymbol{\Sigma}\mathbf{h}$, the second column of $\boldsymbol{\Sigma}$. With the limit covariance (4.35),

$$\mathbf{K} = \frac{1}{0.42308}\begin{bmatrix}0.0057692\\0.17308\\0.34615\end{bmatrix} = \begin{bmatrix}0.013636\\0.40909\\0.81818\end{bmatrix},$$

in kW^{-1} for the first entry and dimensionless for the other two; the innovation is $23.0 - 23.0769 = -0.0769\,\mathrm{K}$, and $\hat{\mathbf{z}}_2 = \hat{\mathbf{z}} - 0.0769\,\mathbf{K} = (0.168182,\ 23.0455,\ 28.0909)^{\mathsf{T}}$, the state of the precision form. The covariance update subtracts the outer product $S\mathbf{K}\mathbf{K}^{\mathsf{T}}$; on its first and last diagonal entries,

$$(\boldsymbol{\Sigma}_2)_{11} = 1.9231 \times 10^{-4} - 0.42308 \times 0.013636^2 = 1.1364 \times 10^{-4}\,(\mathrm{K/kW})^2,$$
$$(\boldsymbol{\Sigma}_2)_{33} = 0.69231 - 0.42308 \times 0.81818^2 = 0.40909\,\mathrm{K}^2,$$

the values of the precision form. Because $S\,\mathbf{K}^{\mathsf{T}} = \mathbf{h}^{\mathsf{T}}\boldsymbol{\Sigma}$, the covariance update is also

$$\boldsymbol{\Sigma}_2 = (\mathbf{I} - \mathbf{K}\mathbf{h}^{\mathsf{T}})\,\boldsymbol{\Sigma},$$

with $\mathbf{I}$ the identity matrix: each column of $\boldsymbol{\Sigma}$ loses the gain vector times that column's second entry. The correction is the outer product of one vector, which is the reason for the name rank-one update. For a linear-Gaussian model it is exact, and one more independent reading costs one product of $\boldsymbol{\Sigma}$ with a vector rather than a new inference. For a strongly nonlinear model, or after a shift of the estimate large enough to make the Jacobian at the old estimate stale, the model is linearized again or the full fit is repeated.

The updated assessment. Table 4.13 follows the assessment through the two readings. Each column is a state of knowledge, before any reading, after the first, after both; the rows are the information about the resistance, the reciprocal of the first diagonal entry of the covariance, the MAP resistance, the mean and standard deviation of the full-load temperature, the threshold probability, and the 95 % credible interval.

Table 4.13: The full-load assessment before the test, after the first reading (23.2°C) and after the second (23.0°C).

Quantity	Prior	One reading	Two readings
$1/\Sigma_{11}$ [$(\mathrm{kW/K})^2$]	1600	5200	8800
$\hat{\rho}$ [K/kW]	0.160000	0.169231	0.168182
$\mathbb{E}[T_2]$ [°C]	27.600	28.154	28.091
$\mathrm{sd}(T_2)$ [K]	1.500	0.832	0.640
$\Pr(T_2 < 28°\mathrm{C})$ [–]	0.6051	0.4267	0.4435
95 % interval [°C]	24.66, 30.54	26.52, 29.78	26.84, 29.34

After the second reading the full-load temperature is the third entry of $\hat{\mathbf{z}}_2$, 28.091°C, with standard deviation $\sqrt{(\boldsymbol{\Sigma}_2)_{33}} = \sqrt{0.40909} = 0.640\,\mathrm{K}$, and the probability of staying below the limit becomes $\Phi(-0.1421) = 0.4435$. The interval narrowed by a quarter, yet the probability moved only from 0.427 to 0.443, because two effects acted in opposite directions. The narrowing alone, with the mean held at 28.154°C, would have lowered the probability to 0.405, since the mean lies above the limit; the shift of the mean alone, with the spread held at 0.832 K, would have raised it to 0.456. The effect of actual new evidence cannot be described by its reduction of the covariance alone: the value observed matters.

4.6.5 What the analysis showed

The handbook said the hall passes at 27.6°C, and the reading scaled to full load said it fails at 28.4°C; the analysis says neither. With the reading, the handbook and the physics combined, the probability that the hall stays below 28°C at 60 kW is 0.4267, less than even, and the full-load temperature is not known well enough to call it either way. One reading at the test condition fixes the thermal resistance of the air path to about ±8 % and the full-load temperature to about ±0.8 K, one standard deviation each, although the hall has never run at 60 kW: the physics carries the reading along the free direction $\mathbf{v}$, from T_1 to ρ and from ρ to T_2. The posterior is a ridge along that direction, the three entries known almost only as one combination, and the reduced state is that ridge named by its one free entry: every question asked of the hall is a question about where it sits on the ridge. The 60 kW question in particular is a resistance question: the hall stays below 28°C exactly when the resistance stays below ρ^*, and the reading at 30 kW places the resistance close to ρ^*, which leaves the question near even odds. Further readings at the test condition narrow the ridge and resolve the question without deciding its answer. A reading at full load would act on the third column of $\mathbf{J}$ directly, and its worth is what its preposterior covariance $\boldsymbol{\Sigma}_a$ says before it is taken: as much as four readings at the test condition, with the shared resistance itself put to the test.

4.6.6 Limits of the model

The steady-state balance (4.2) neglects the heat stored in the hall and its equipment, and it describes the temperature long after a load change, not minutes after it; Exercise 4.6 formulates the transient model that a question about the first minutes of full load would need. The single effective path assumes that the resistance is the same at 60 kW as at 30 kW; a path whose resistance changes with load, for example because the unit's fans or its coil change state, would break the link drawn in Figure 4.2, and one test reading cannot detect it. The supply air is pinned at 18°C by the discharge controller, which holds it only while the unit has capacity to spare; a unit at its limit at full load lets the supply air rise, which shifts the full-load temperature up, not only widens it; an uncertain supply air adds its variance through the budget (4.54).

The probabilistic assumptions have limits of their own. The Gaussian prior on a positive resistance is harmless here, with a negative mass of 8×10^{-11}, but a wide prior on a small quantity would place real probability below zero; a logarithmic parameterization avoids this at the price of a nonlinear model and an approximate posterior. The sensor error is taken as Gaussian, unbiased and of known spread; a bias would shift every inferred quantity, and Section 4.2.2 names the comparison of models that would detect it. The physical rows are taken in the limit of zero tolerance; a real model discrepancy is a finite σ_p, which widens Σ_{11} and adds σ_p^2 to the full-load variance by equation (4.34), and the reduced state of Section 4.2.8 is exact only in that limit. The reading of 23.2°C is a stated example value, not a measured record, and the chapter has no truth to check its answers against; from Chapter 5 on the records are constructed from a stated truth and the checks compare the answers with it. Each of these is a statement about the model, and the probability 0.4267 is conditional on all of them.

Chapter 5 grows the hall into hall B of a two-hall site: three cooling units on chilled water produce the supply air, now the cold aisle, and five racks draw from it, and the question becomes how far the load can rise. The method is unchanged, and it runs on the reduced state and the scenario node of this chapter.

4.7 What to retain

Definition 4.1 (The anatomy of an inference)**.** A design probability over a plant graph is obtained in two operations. Inference sums one term per row into the objective $\mathcal{L} = -\ln p(\mathbf{z} \mid \mathbf{y})$ up to a constant, and uses only its gradient and Hessian. A Gaussian row's term is half its

squared residual; a count row's is its negative log-probability, which joins the Jacobian through its working residual and line, reproducing the row's gradient and information but not its value. The physical rows, the readings and the priors, each divided by its standard deviation, stack into one residual vector $\mathbf{r}(\mathbf{z})$ with Jacobian $\mathbf{J}$; the normal equations $(\mathbf{J}^\mathsf{T}\mathbf{J})\,\Delta\mathbf{z} = -\mathbf{J}^\mathsf{T}\mathbf{r}$ lead to the MAP state $\hat{\mathbf{z}}$, and the inverse of the information matrix, $\mathbf{\Sigma} = (\mathbf{J}^\mathsf{T}\mathbf{J})^{-1}$, is the posterior covariance, exactly for a linear-Gaussian model and under the Laplace approximation, in its Gauss–Newton form, otherwise. An exact physical relation is the limit of a row's tolerance, never a substitution; in that limit the state reduces to its free entries, with Jacobian $\mathbf{J}_{\text{rows}}\mathbf{N}$ on the null space $\mathbf{N}$ of the physical rows. Query evaluation then reads the answer off the posterior: an entry of the state with its variance Σ_{ii}, or a derived quantity $f(\mathbf{z})$ with variance $\mathbf{g}^\mathsf{T}\mathbf{\Sigma}\mathbf{g}$, and a threshold probability as the normal tail $\Phi\big((f^* - f(\hat{\mathbf{z}}))/\sqrt{\mathbf{g}^\mathsf{T}\mathbf{\Sigma}\mathbf{g}}\big)$, which for an entry is $\Phi\big((f^* - \hat{z}_i)/\sqrt{\Sigma_{ii}}\big)$. A scenario node pins an input at a value not yet seen, solves the physical rows again there, and carries $\mathbf{\Sigma}$ to the result through its gradient.

Three conclusions carry beyond the example. Physical relations connect the quantities that are measured to those that are not, which is what allows an unobserved quantity, here a full-load temperature never reached, to be inferred from indirect evidence. Uncertainty is not an error bar attached to a point estimate but a probability distribution, and it must be propagated through the physical model to answer a planner's question. And the posterior is reusable: once the joint posterior is in hand, many questions are answered without repeating the inference, and new evidence updates it rather than replacing it. A derived quantity is propagated in one of two ways: its spread by the gradient, $\mathbf{g}^\mathsf{T}\mathbf{\Sigma}\mathbf{g}$, and its density by the change of variables, which multiplies the density by the change-of-variable factor of the map and so keeps the skew and the bounds that a linearization loses.

The inference of the hall returns the posterior state $\hat{\mathbf{z}}$ of equation (4.49), its covariance $\mathbf{\Sigma}$ of equations (4.50) and (4.35), and the Jacobian $\mathbf{J}$ of equation (4.25), from which the gradients of derived quantities are formed. Table 4.14 lists what the chapter has computed from them without repeating the inference: the four answers of Table 4.2, and four further quantities that later chapters reuse. Each row is a quantity, the calculation on the posterior that gives it, and the place where it is carried out.

Table 4.14: The four answers of Table 4.2 and four further quantities reused later, all computed from one posterior of the equipment hall.

Quantity	Calculation on the posterior	Section
The answers		
Thermal resistance of the air path	the entry $\hat{z}_1$ and $\sqrt{\Sigma_{11}}$	4.3.6
Hall temperature at the test condition	the entry $\hat{z}_2$ and $\sqrt{\Sigma_{22}}$	4.3.6
Hall temperature at full load	the entry $\hat{z}_3$, $\sqrt{\Sigma_{33}}$ and the credible interval	4.3.6
Probability of staying below 28°C at 60 kW	the Gaussian tail $\Phi\big((T_{\max} - \hat{z}_3)/\sqrt{\Sigma_{33}}\big)$	4.3.7
Further quantities reused later		
Spread of the conductance	$\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}$ with $\mathbf{g} = (-1/\hat{z}_1^2, 0, 0)^{\mathsf{T}}$, and the exact push-forward	4.3.6
The entry that drives the temperature	the gradient and the variance budget	4.3.2
Hall temperature at a load not yet carried	the scenario node $T_c + Q^*\hat{\rho}$, standard deviation $Q^*\sqrt{\boldsymbol{\Sigma}_{\text{red}}}$	4.3.8
Effect of another reading	the information matrix plus the reading's $\mathbf{j}^{\mathsf{T}}\mathbf{j}$	4.6.1

The table is one inference followed by a column of cheap queries on its result, the division of Section 4.1.2. Two kinds of question fall outside it: a question that compares two models, which needs the evidence of each (Section 4.2.2), and the arrival of new evidence, which changes the posterior itself (Section 4.6.4).

Table 4.15 lists the five mathematical objects of the chapter that recur throughout the book, with the form of each and its purpose. The first three are the working objects when the posterior is well represented by a local Gaussian. The posterior distribution itself is the more fundamental object: it may be Gaussian, multimodal, discrete, a mixture or otherwise non-Gaussian, and its structure decides which computational method is appropriate.

Table 4.15: The five mathematical objects of an inference.

Object	Form	Purpose
State estimate	$\hat{\mathbf{z}}$	the most probable (or otherwise selected) central state
Posterior covariance	$\boldsymbol{\Sigma}$	local uncertainty of the state and the dependence among its entries
Output gradient	$\mathbf{g} = \nabla f(\hat{\mathbf{z}})$	connects state uncertainty to a derived quantity
Posterior distribution	$p(\mathbf{z} \mid \mathbf{y}, M)$	the uncertainty about the state given the evidence
Model evidence	$p(\mathbf{y} \mid M)$	normalizes the posterior and compares models

Table 4.16 lists the terms of the chapter with their expressions, units and the subsection that defines each.

Table 4.16: The terms of the anatomy of an inference and where each is defined. Units for the equipment hall.

Term	Expression	Units	Meaning	Defined in
Thermal resistance	$\rho = 1/G$, $T = T_c + Q\rho$	K/kW	makes every row linear in the state	4.1.4
Physical rows	$r_1 = T_1 - 18 - 30\rho$, r_2	K	conservation at each hall node	4.1.5
The question as one marginal	$\Pr(T_2 < T_{\max})$ from $\hat{z}_3$, Σ_{33}	–	one entry's mean and variance	4.1.6
Model tolerance	σ_p; exact rows as $\sigma_p \to 0$	K	model discrepancy of a physical row	4.1.5
Likelihood	$p(y_T \mid \mathbf{z})$, eq. (4.14)	K^{-1}	the reading as a function of the state	4.2.1
Residual vector	$\mathbf{r}(\mathbf{z})$, eq. (4.23)	–	four rows, each over its standard deviation	4.2.3
Objective	$\mathcal{L} = \frac{1}{2}\mathbf{r}^{\mathsf{T}}\mathbf{r}$	–	minus log posterior	4.2.3
Jacobian	$\mathbf{J} = \partial\mathbf{r}/\partial\mathbf{z}$, 4×3	mixed	one line per row, one column per entry	4.2.3
Information matrix	$\mathbf{J}^{\mathsf{T}}\mathbf{J} = \sum_k \mathbf{j}_k^{\mathsf{T}}\mathbf{j}_k = \tilde{\mathbf{J}}^{\mathsf{T}}\mathbf{W}\tilde{\mathbf{J}}$	mixed	the Hessian when $\mathbf{C} = \mathbf{0}$	4.2.3
General objective	$\mathcal{L} = \sum_k \mathcal{L}_k$, $\mathcal{L}_k = -\ln p_k$	–	one term per row, any distribution	4.2.7
Count row, working form	$r_k = (\mu_k - n_k)/\sqrt{\mu_k}$, $\mathbf{j}_k = \sqrt{\mu_k}\,\nabla \ln \mu_k^{\mathsf{T}}$	–	gradient and information, not the term	4.2.7
Hessian	$\mathbf{H} = \mathbf{J}^{\mathsf{T}}\mathbf{J} + \mathbf{C}$	mixed	the information plus a named curvature term	4.2.7
Normal equations	$(\mathbf{J}^{\mathsf{T}}\mathbf{J})\,\Delta\mathbf{z} = -\mathbf{J}^{\mathsf{T}}\mathbf{r}$	mixed	Gauss–Newton step to $\hat{\mathbf{z}}$	4.2.4
Posterior covariance	$\mathbf{\Sigma} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}$	mixed	spread is inverse information	4.2.5
Laplace approximation	$\mathcal{N}(\hat{\mathbf{z}}, \mathbf{H}^{-1})$	–	local Gaussian posterior	4.2.6
Free direction	$\mathbf{v}$, $\mathbf{J}_{\mathrm{phys}}\mathbf{v} = \mathbf{0}$	mixed	$\mathbf{\Sigma} \to \mathbf{v}\mathbf{v}^{\mathsf{T}}/5200$, rank one	4.2.4
Delta method	$\mathrm{Var}(f) \approx \mathbf{g}^{\mathsf{T}}\mathbf{\Sigma}\mathbf{g}$	varies	linearized propagation	4.3.1
Change of variables	$p_u(u) = p_x(h^{-1}(u))\,\lvert \mathrm{d}x/\mathrm{d}u \rvert$, eq. (4.55)	varies	density of a transformed quantity	4.3.3
Variance budget	eq. (4.54)	K^2	one term per entry plus cross terms	4.3.2
Threshold probability	$\Phi\big((T_{\max} - \hat{z}_3)/\sqrt{\Sigma_{33}}\big)$	–	normal tail at the limit	4.3.5
Credible interval	$\hat{z}_3 \pm 1.960\sqrt{\Sigma_{33}}$	°C	central 95 % posterior range	4.3.4
Scenario node	$T^* = T_c + Q^*\rho$, eq. (4.62)	°C	a what-if at an input not yet seen	4.3.8
Reduced state	$\mathbf{N}$, $\mathbf{J}_{\mathrm{red}} = \mathbf{J}_{\mathrm{rows}}\mathbf{N}$, $\mathbf{\Sigma}_{\mathrm{red}}$	mixed	the free entries with exact physical rows	4.2.8
Prior predictive	$\mathcal{N}(22.8, 0.9014^2)$	°C	reading predicted before use	4.4.2
Posterior predictive	$\mathcal{N}(23.077, 0.6504^2)$	°C	next reading predicted	4.4.4
Evidence	$p(y_T \mid M)$, eq. (4.18)	K^{-1}	normalizer; compares models	4.2.2
Bayes factor	$B_{21} = p(y_T \mid M_2)/p(y_T \mid M_1)$	–	evidence ratio of two models	4.2.2
Rank-one update	eq. (4.68)	–	adds one reading to a posterior	4.6.4
Preposterior covariance	$\mathbf{\Sigma}_a = (\mathbf{\Sigma}^{-1} + \mathbf{\Lambda}_a)^{-1}$, eq. (4.67)	mixed	ranks candidate readings	4.6.1

4.8 Exercises

Exercise 4.1 (Application: a complete inference)**.** A different equipment hall obeys $T = T_c + Q\rho$ with its supply air held at $T_c = 20°\mathrm{C}$ and a test load $Q = 20\,\mathrm{kW}$. The prior on the resistance is $\rho \sim \mathcal{N}(0.20, 0.04^2)$ in K/kW, and a sensor with $\sigma_T = 0.4\,\mathrm{K}$ reads $y_T = 24.5°\mathrm{C}$. The question is the temperature at a design load of 40 kW. (a) Write the state $\mathbf{z} = (\rho, T_1, T_2)^\mathsf{T}$, with T_1 the test temperature and T_2 the temperature at 40 kW, the four normalized rows, and the 4×3 Jacobian at $\sigma_p = 0.01\,\mathrm{K}$. ($\mathbf{r} = ((T_1 - 20 - 20\rho)/\sigma_p,\ (T_2 - 20 - 40\rho)/\sigma_p,\ (T_1 - 24.5)/0.4,\ (\rho - 0.20)/0.04)^\mathsf{T}$; $\mathbf{J}$ has the lines $(-2000, 100, 0)$, $(-4000, 0, 100)$, $(0, 2.5, 0)$ and $(25, 0, 0)$.) (b) Take one Gauss–Newton step from the handbook state and compute $\boldsymbol{\Sigma}$ in the limit of exact rows. (At $\mathbf{z}^{(0)} = (0.20,\ 24.0,\ 28.0)^\mathsf{T}$ only the reading row is nonzero, -1.25, and $\mathbf{J}^\mathsf{T}\mathbf{r} = (0,\ -3.125,\ 0)^\mathsf{T}$; the normal equations give $\hat{\mathbf{z}} = (0.2200,\ 24.400,\ 28.800)^\mathsf{T}$, and a second step does not move it. The free direction is $\mathbf{v} = (1,\ 20,\ 40)^\mathsf{T}$, with $\mathbf{J}\mathbf{v} = (0,\ 0,\ 50,\ 25)^\mathsf{T}$ and information $50^2 + 25^2 = 3125$, so $\boldsymbol{\Sigma} \to \mathbf{v}\mathbf{v}^\mathsf{T}/3125$: $\Sigma_{11} = 3.2 \times 10^{-4}$, $\sqrt{\Sigma_{11}} = 0.01789\,\mathrm{K/kW}$; at $\sigma_p = 0.01\,\mathrm{K}$ the fit gives $\Sigma_{11} = 3.2016 \times 10^{-4}$ and a correlation of 0.9996 between ρ and T_1.) (c) Read off the posterior mean and standard deviation of the temperature at 40 kW, and the probability that it stays below 30°C. ($\hat{z}_3 = 28.8°\mathrm{C}$ and $\sqrt{\Sigma_{33}} = \sqrt{1600/3125} = 0.7155\,\mathrm{K}$; $\Pr(T_2 < 30 \mid \mathbf{y}) = \Phi(1.2/0.7155) = \Phi(1.677) = 0.9532$, against $\Phi(2/1.6) = 0.8944$ under the prior alone.)

Exercise 4.2 (Derivation: a variance budget with correlated entries)**.** For $T = T_c + Q\rho$ at the known load $Q = 50\,\mathrm{kW}$, the posterior covariance of (T_c, ρ) is $\boldsymbol{\Sigma} = \begin{bmatrix} 0.25 & -0.003 \\ -0.003 & 0.0001 \end{bmatrix}$, in K^2, K·K/kW and $(\mathrm{K/kW})^2$. (a) Derive the temperature variance from $\mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g}$ and compute the standard deviation. ($\mathbf{g} = (1, 50)^\mathsf{T}$; $\mathrm{Var}(T) = 0.25 + 50^2 \times 0.0001 + 2 \times 50 \times (-0.003) = 0.25 + 0.25 - 0.30 = 0.20\,\mathrm{K}^2$, $\mathrm{sd}(T) = 0.4472\,\mathrm{K}$.) (b) State the result if the covariance term were ignored, and explain the difference. ($\sqrt{0.50} = 0.7071\,\mathrm{K}$, 58 % too large; the posterior correlation is $-0.003/\sqrt{0.25 \times 0.0001} = -0.6$, and with sensitivities of the same sign the negative cross term cancels part of the two diagonal terms: the evidence pins the combination $T_c + 50\rho$ better than either entry.)

Exercise 4.3 (Derivation: a nonlinear derived quantity)**.** With $\rho \mid \mathbf{y} \sim \mathcal{N}(0.169231, 0.013868^2)$ from the chapter's example, derive the first-order approximation to the variance of the conductance $G = 1/\rho$, compute its standard deviation, and explain why the result is only an approximation. (With $G(\mathbf{z}) = 1/z_1$, $\mathbf{g} = (-1/\hat{z}_1^2,\ 0,\ 0)^\mathsf{T}$, so $\mathrm{Var}(G \mid \mathbf{y}) \approx \mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g} = \Sigma_{11}/\hat{z}_1^4$; $\hat{G} = 1/0.169231 = 5.909\,\mathrm{kW/K}$ and $\mathrm{sd}(G \mid \mathbf{y}) \approx 0.013868/0.169231^2 = 0.484\,\mathrm{kW/K}$. The map $1/\rho$ is nonlinear, so G is not Gaussian: its exact 5–95 % band is $[5.207, 6.830]$, skewed, against the linearized $[5.113, 6.706]$; strictly, the moments of $1/\rho$ do not exist under a Gaussian ρ, which has positive density at zero, so positivity and the transformed distribution must be handled explicitly when exact conductance statements are needed.)

Exercise 4.4 (Formulation: observability and the plant graph)**.** In $T = T_c + Q\rho$ suppose that both T_c and ρ are unknown, the discharge controller's set point not being trusted, and that there is a single temperature reading at one known load. (a) Write the observation Jacobian with respect to (T_c, ρ) and state whether the reading can determine both unknowns without further information. ($y = T_c + Q\rho + \epsilon$, so $\mathbf{J} = [1 \quad Q]$, of rank one for two unknowns: the reading determines only the combination $T_c + Q\rho$, and every pair with the same combination explains it equally well.) (b) Name a reading or operating condition that resolves the ambiguity. (A second reading at a different known load $Q_2 \neq Q_1$ gives $\mathbf{J} = \begin{bmatrix} 1 & Q_1 \\ 1 & Q_2 \end{bmatrix}$, with determinant $Q_2 - Q_1 \neq 0$ and full rank; alternatively, a direct reading of the supply air, by a discharge sensor, a new node observed in the graph, removes the ambiguity.)

Exercise 4.5 (Application: Bayesian evidence)**.** A model predicts $Y \sim \mathcal{N}(24, 1^2)$ in °C, and the reading $y = 25$°C arrives. (a) Compute the marginal likelihood $p(y \mid M)$. ($\frac{1}{\sqrt{2\pi}} \exp[-(25-24)^2/2] = 0.39894 \times 0.60653 = 0.2420\,\mathrm{K}^{-1}$.) (b) Explain why this value is not the probability that M is correct. (It is a probability density in inverse temperature units, 2.420×10^{-4} per millikelvin for the same reading, not $\Pr(M \mid y)$; a posterior model probability requires prior probabilities and the evidence of the alternative models under comparison.)

Exercise 4.6 (Formulation: model validity and thermal storage)**.** The chapter assumed the steady-state balance $Q = G(T - T_c)$. The load now steps from 30 kW to 60 kW, and the design review asks for the temperature five minutes later. (a) Explain why the steady-state equation may be insufficient. (It neglects the energy stored in the air, structure and equipment of the hall, so it gives only the temperature reached long after the change; five minutes may be short against the thermal time constant.) (b) Starting from an energy balance, formulate a model with thermal storage and name the new parameter and the condition it needs. (With an effective thermal capacitance C_{th} in kJ/K, $C_{\mathrm{th}}\,\mathrm{d}T/\mathrm{d}t = Q - G(T - T_c)$; the new parameter is C_{th} and an initial temperature $T(0)$ is required. For constant Q, G and T_c after the step, $T(t) = T_c + Q/G + [T(0) - T_c - Q/G]\,\mathrm{e}^{-Gt/C_{\mathrm{th}}}$, with time constant C_{th}/G; the probability for the temperature at five minutes propagates the posterior of G, C_{th} and $T(0)$ through this time-dependent model rather than through the steady-state equation.)

Part II

Load-Carrying Capability

Chapter 5

Thermal Headroom

5.1 How much more load hall B can take

A design review is deciding how much more IT load hall B can be given. The hall has three cooling units and five racks, and at 19:00 its load meter reads 277.20 kW. The design sheet answers the question in two ways, and neither comes with a band. Worked from the nameplate ratings of the three coils and the commissioning airflow survey, it finds that rack B3, the rack the survey put the most recirculation at, reaches the 27°C inlet limit first, at a hall load of 410.0 kW: 132.8 kW over the meter's reading. Worked instead from the two inlet sensors the hall does have, at racks B2 and B4, each reading's rise over the chilled-water supply scaled up in proportion to the load, it finds that rack B4 reaches the limit first, at 460.6 kW, a headroom of 183.4 kW.

The floor believes the second number, because it comes from readings. But the hottest rack on the survey, B3, has no sensor, and neither number says how sure the review can be. The question of this chapter is the planner's own,

How much load can hall B take before an inlet crosses the limit, and how sure is that number?

It is answered with the method of Chapter 4, on the reduced state of hall B and with the scenario node of Section 4.3.8. What is new is the quantity asked for: not a temperature at a load chosen in advance, but the load at which a temperature meets its limit, solved for.

Jargon: headroom, binding load, design load, cold aisle, return air, chilled-water supply, discharge air temperature, air capacity rate, log-derate, recirculation fraction, airflow survey, design-day supply, implicit function theorem, inversion band, importance reweighting, observability, parameter risk (variance budget), assurance level

The *headroom* of a hall is the load it can still take before a limit binds. The *binding load* $Q_e^\dagger$ of an element e of the plant is the hall load at which that element's limit is just met, and the hall's binding load is the smallest of them, the one the *binding constraint* sets. The *design load* is the load the hall is to be released for, here 340 kW of IT load. The *cold aisle* is the space in front of the racks that the cooling units fill with cooled air; the *return air* is the warm air leaving the racks on its way back to the units. The *chilled-water supply* T_s is the temperature of the water that reaches each unit's coil. The *discharge air temperature* of a unit is the temperature of the air leaving its coil, and its *air capacity rate* C_u, in kW/K, is its air mass flow times the specific heat of air. A unit's *log-derate* $g_u = \ln(G_u/\bar{G}_u)$ compares its coil's effective conductance with the rating; it is zero at the rating. The *recirculation fraction* β_k of rack k is the share of the air at its inlet that is return air drawn back over or around the rack. An *airflow survey* is a one-time measurement campaign that estimates those fractions. The *design-day supply* is the chilled-water supply temperature the drawing assumes on the design day, here 14.0°C. The *implicit function theorem* gives the gradient of a quantity defined by an equation it must satisfy, such as the load at which a temperature meets its limit, from the gradients of that equation. The *inversion band* of a quantity that is a monotone function of a nearly normal one is the band of the normal one carried through the function, end to end. *Importance reweighting* turns draws from an approximate posterior into draws from the exact one by weighting each draw by the ratio of the two densities. *Observability* (in the framework, the rank of the Jacobian's columns and an entry's posterior spread against its prior spread) is whether the readings determine an entry of the state. The *parameter risk* of a derived quantity, its *variance budget* in the framework, splits its variance into one share per entry of the state. The *assurance level* κ is the posterior probability that the design policy requires a requirement to reach.

Key idea

Headroom is a scenario solved backward. The scenario node fixes what the planner pins, solves the plant for the load at which an element's limit binds, and carries the posterior of the state to that load by the gradient of the binding entry divided by the entry's sensitivity to the load. The binding load is then a derived quantity of the fitted state, with a band, a budget and a ranking of the measurements that would narrow it. The hall's binding load is the smallest over its racks, so it is read jointly, from draws of the posterior of the state pushed through every rack's binding load; and because each rack's binding load is a ratio, a symmetric normal on one rack misplaces its tails.

5.1.1 Headroom read as a binding load

Chapter 4 carried the calibrated posterior to a load the hall had not carried, and read the temperature there with its band. The headroom question turns that around. The limit is fixed and the load is the unknown: for each element e whose entry is constrained, here each rack inlet, the scenario is solved for the load $Q_e^\dagger$ at which the entry meets its limit, and the hall's binding load is the smallest of them, $Q^\dagger = \min_e Q_e^\dagger$. The headroom is that load less the load today.

The binding load is defined by an equation, not by a formula. Write $T_e^*(\mathbf{z}, Q^*)$ for the constrained entry of element e at the scenario node of equation (4.62), as a function of the free entries $\mathbf{z}$ and of the pinned load Q^*. The binding load is the load at which that entry equals

the limit,

$$T_e^*(\mathbf{z},\ Q_e^\dagger(\mathbf{z})) = T_{\max},$$

and it holds at every state near the calibrated one. Differentiating the identity with respect to the state, the change of the entry through the state and through the load must cancel,

$$\nabla_{\mathbf{z}} T_e^* + \frac{\partial T_e^*}{\partial Q^*}\,\nabla_{\mathbf{z}} Q_e^\dagger = \mathbf{0},$$

so that, wherever the entry moves with the load,

$$\mathbf{g}_e = \nabla_{\mathbf{z}} Q_e^\dagger = -\left.\frac{\nabla_{\mathbf{z}} T_e^*}{\partial T_e^*/\partial Q^*}\right|_{(\hat{\mathbf{z}},\ Q_e^\dagger(\hat{\mathbf{z}}))}, \qquad \mathrm{Var}(Q_e^\dagger \mid \mathbf{y}) \approx \mathbf{g}_e^{\mathsf{T}}\,\boldsymbol{\Sigma}\,\mathbf{g}_e. \tag{5.1}$$

Here $\nabla_{\mathbf{z}} T_e^*$ is the gradient of the constrained entry at the scenario with the load held at the binding load, and $\partial T_e^*/\partial Q^*$ is the entry's sensitivity to the load there, in K/kW. The first relation is the implicit function theorem; the second is the delta method of equation (4.53) applied to its result. The band of the binding load is the band of the binding entry, in kelvin, divided by how fast the entry rises with the load, so that a shallow slope turns a narrow temperature band into a wide load band. Later chapters cite equation (5.1) for every binding load they meet. The hall's binding load $Q^\dagger$ takes the gradient of the element that binds at $\hat{\mathbf{z}}$, as long as no other element ties with it; how often another element binds first over the posterior is a question of its own.

Principle 5.1. *A binding load is the scenario node solved for the load at which a limit binds. Its gradient is the binding entry's gradient divided by the entry's sensitivity to the load, and its band is that gradient pushed through the covariance of the calibrated state.*

5.1.2 Hall B and its rows (step 0)

The equipment hall of Part I grows here into hall B, one of the two halls of a plant served by one chilled-water loop and observed over a night of telemetry. Its supply air becomes the cold aisle T_c, produced by three cooling coils B1–B3 from chilled water at the supply temperature T_s; its hall air becomes the return air T_r; and five racks B1–B5 draw their inlet air from the cold aisle with recirculation fractions β_k. The inlets of racks B2 and B4 are measured.

The physical relations. The racks put the IT load Q_B into the hall and the fans and lighting add the other heat Q_o. Unit u's coil passes heat to the chilled water at the rate

$$Q_u = G_u\,(T_r - T_s), \qquad G_u = \bar{G}_u\,\mathrm{e}^{g_u},$$

the coil heat, with $\bar{G}_u$ the rated conductance. In steady state the heat the racks and the other sources put in leaves through the three coils, the hall balance; the air crossing unit u gives up to its coil what the coil passes to the water, which sets its discharge temperature $T_{d,u}$; the three discharge streams mix in the cold aisle in proportion to their capacity rates; and the inlet of rack k is cold-aisle air with a share β_k of return air mixed in. As rows that vanish when they hold, each with its own tolerance,

$$\begin{aligned}
r_{\mathrm{h}} &= \Big[\textstyle\sum_u G_u\,(T_r - T_s) - Q_B - Q_o\Big]/\sigma_{p,\mathrm{h}} && && \text{hall balance,}\\
r_{\mathrm{a},u} &= \big[C_u\,(T_r - T_{d,u}) - G_u\,(T_r - T_s)\big]/\sigma_{p,\mathrm{a}}, && u = \mathrm{B1},\mathrm{B2},\mathrm{B3} && \text{air side of unit } u,\\
r_{\mathrm{m}} &= \Big[\textstyle\sum_u C_u\,T_c - \sum_u C_u T_{d,u}\Big]/\sigma_{p,\mathrm{m}} && && \text{cold-aisle mixing,}\\
r_{\mathrm{i},k} &= \big[T_k - T_c - \beta_k\,(T_r - T_c)\big]/\sigma_{p,\mathrm{i}}, && k = \mathrm{B1},\ldots,\mathrm{B5} && \text{rack inlet } k,
\end{aligned} \tag{5.2}$$

ten physical rows, the tolerances $\sigma_{p,\mathrm{h}}$, $\sigma_{p,\mathrm{a}}$ and $\sigma_{p,\mathrm{m}}$ in kW and $\sigma_{p,\mathrm{i}}$ in K, taken to zero together in the limit written $\sigma_p \to 0$. The pinned quantities are the ratings $\bar{G}_u = 7\,\mathrm{kW/K}$, the air capacity rates $C_u = 10\,\mathrm{kW/K}$ at the units' fixed fan speed, and the other heat $Q_o = 10\,\mathrm{kW}$. The free quantities of the graph form the full state of twenty entries,

$$\mathbf{z}_{\mathrm{full}} = (T_s,\ Q_B,\ g_{B1},\ g_{B2},\ g_{B3},\ \beta_{B1},\dots,\beta_{B5} \mid T_r,\ T_{d,B1},\ T_{d,B2},\ T_{d,B3},\ T_c,\ T_{B1},\dots,T_{B5})^{\mathsf{T}}, \tag{5.3}$$

ten free entries and ten temperatures that the ten rows fix. The supply temperature is a free entry in the calibration: a supply sensor reads it and the chiller's set point is its prior.

Modeling choices at the graph. The hall is taken in steady state within the window: this is what makes every relation of equation (5.2) algebraic, so that one window's readings fix one state, and it fails in the minutes after a load step or a unit start, while the hall's air and equipment store or release heat. The heat path through each unit is one effective conductance between the return air and the chilled water: this keeps one entry per coil, and it fails if a coil's heat transfer depends on its own discharge temperature in a way one conductance cannot carry. The units' airflow is pinned and nothing controls the discharge temperature within the window: this is what lets the discharge readings inform the log-derates, since a coil that passes less heat leaves its discharge warmer, and an active controller that held $T_{d,u}$ at a set point by moving its valve would break it (Section 5.6.5).

The graph. The rows of hall B are drawn in Figure 5.1. The cold aisle on the left feeds the five racks through a common supply line; the return air in the middle receives the IT load Q_B from above and feeds the racks' recirculated air back through a second line, and feeds the three units on the right; the units pass heat to the chilled water below them and send their discharge air back to the cold aisle along the top. Each rack node is the mixing row of its inlet, fed by the cold aisle with weight $1-\beta_k$ and by the return air with weight β_k. Rack B3, the survey's hottest rack and unmeasured, is drawn in red, and beside it the scenario node at which the headroom question is asked. The dashed arrows leave the graph for the eight readings.

The reduced state. The ten physical rows fix the ten temperatures at given free entries, and in the limit $\sigma_p \to 0$ the fit runs on the reduced state of Section 4.2.8, the ten free entries,

$$\mathbf{z} = (T_s,\ Q_B,\ g_{B1},\ g_{B2},\ g_{B3},\ \beta_{B1},\ \beta_{B2},\ \beta_{B3},\ \beta_{B4},\ \beta_{B5})^{\mathsf{T}}, \tag{5.4}$$

in °C, kW and eight dimensionless entries. The dependent block $\mathbf{A}_{\mathrm{dep}}$ of equation (4.45) is triangular in the order of equation (5.3), with nonzero diagonal entries over the tolerances (the conductance sum $\sum_u G_u$, the three air capacity rates C_u, their sum $\sum_u C_u$, and ones), so it is invertible and the dependent temperatures follow in that order,

$$\begin{aligned} T_r &= T_s + \frac{Q_B + Q_o}{\sum_u G_u}, \qquad & T_{d,u} &= T_r - \frac{G_u\,(T_r - T_s)}{C_u}, \\ T_c &= \frac{\sum_u C_u T_{d,u}}{\sum_u C_u}, \qquad & T_k &= T_c + \beta_k\,(T_r - T_c), \end{aligned} \tag{5.5}$$

the point $\mathbf{z}_{\mathrm{full}}(\mathbf{z})$ of the set on which the balances hold; the rows of $\mathbf{N}_{\mathrm{b}}$ are the gradients of these ten temperatures. The air-side rows summed with the mixing row and the hall balance give the air balance of the three units together,

$$\textstyle\sum_u C_u\,(T_r - T_c) = Q_B + Q_o, \tag{5.6}$$

so that the rise from the cold aisle to the return is set by the heat and the airflow alone, not by the coils.

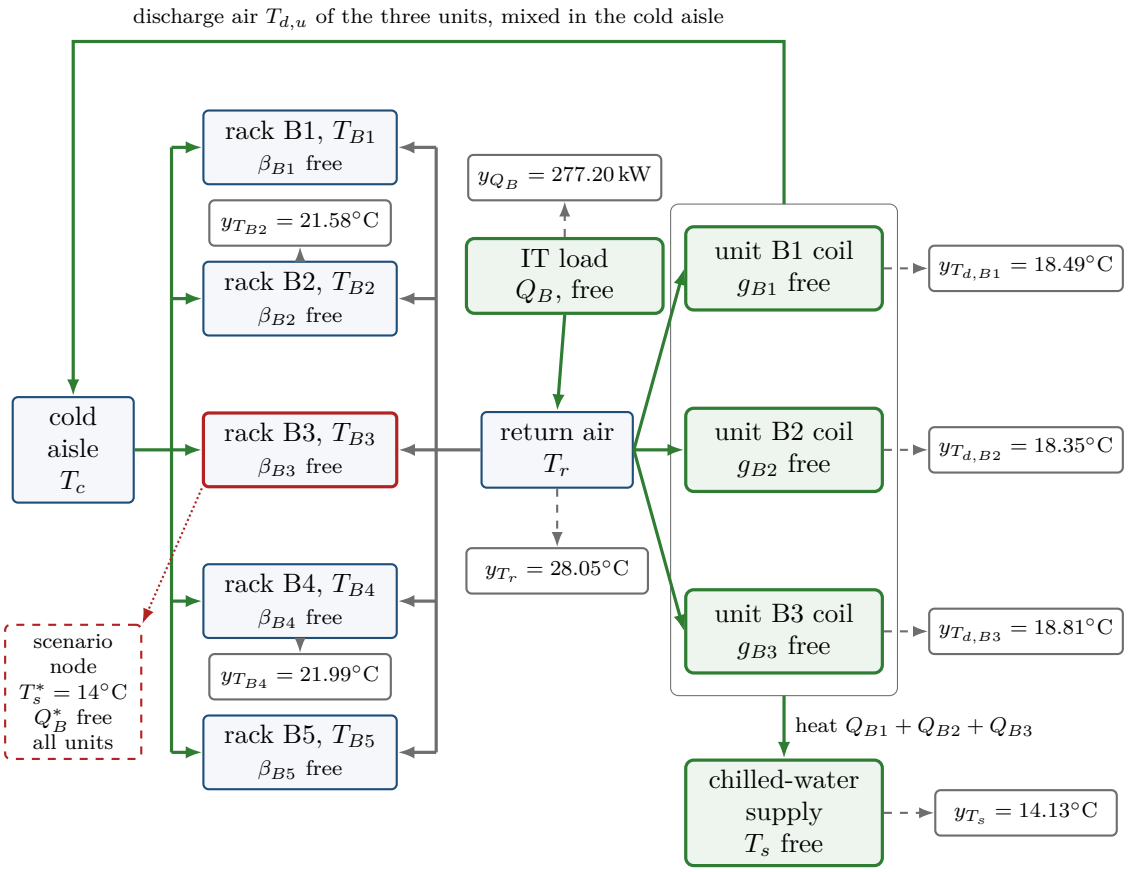


Figure 5.1: The rows of hall B in window 1 of the night record (19:00). Free (green): the chilled-water supply T_s, the IT load Q_B and the coils' log-derates g_{B1}, g_{B2}, g_{B3}; the recirculation fractions $\beta_{B1}, \ldots, \beta_{B5}$ are free entries carried by the rack nodes. Blue: temperatures fixed by the physical rows (5.2), the cold aisle T_c, the return air T_r and the rack inlets T_k, each a mix of the cold aisle (weight $1 - \beta_k$, green line) and the return air (weight β_k, gray line). Pinned, not drawn as nodes: the ratings $\bar{G}_u = 7\,\text{kW/K}$, the capacity rates $C_u = 10\,\text{kW/K}$ and the other heat $Q_o = 10\,\text{kW}$. Observed (dashed): the supply, the return air, the three discharge temperatures, the hall load meter and the inlets of racks B2 and B4. Rack B3 (red) is the survey's hottest rack; the dotted arrow marks the scenario node (red, dashed), which solves the same rows at the pinned supply T_s^* with the load Q_B^* free, for every rack. The priors on T_s, the g_u and the β_k are rows not drawn.

5.1.3 The scenario node of hall B

The headroom question is asked at the scenario node of Section 4.3.8. Three inputs are pinned there, each a modeling choice. The supply is pinned at the drawing's design-day value $T_s^* = 14.0°\mathrm{C}$: this is what makes the answer a statement about the hall on the drawing's terms rather than about window 1's supply, and it fails if the central plant cannot hold that supply at the higher load, a limit of the chillers that Part III takes up. All three units run: this is the healthy plant, and the hall with a unit out is a later chapter's question. The load Q_B^* is free, and the coils' log-derates and the racks' fractions keep their calibrated values as the load rises: this is what makes every inlet linear in the load, and it fails if the airflow pattern, and with it a fraction, changes with the load.

At these pins the chain (5.5) gives $T_r^* - T_s^* = (Q_B^* + Q_o)/\sum_u G_u$, and the air balance (5.6) gives $T_r^* - T_c^* = (Q_B^* + Q_o)/\sum_u C_u$. Writing the inlet as $T_k^* = T_r^* - (1-\beta_k)(T_r^* - T_c^*)$,

$$T_k^*(\mathbf{z},\ Q_B^*) = T_s^* + a_k(\mathbf{z})\,(Q_B^* + Q_o), \qquad a_k(\mathbf{z}) = \frac{1}{\sum_u \bar{G}_u\,\mathrm{e}^{g_u}} - \frac{1-\beta_k}{\sum_u C_u}, \tag{5.7}$$

every inlet rising in proportion to the heat in the hall, with a_k the inlet rise of rack k per kilowatt, in K/kW. The rise per kilowatt depends on the three log-derates, through the return air, and on the rack's own fraction; it does not depend on the supply or on today's load, which the scenario has pinned and freed. All five racks share the first term, so at any one state the racks bind in the order of their fractions, the largest first.

The binding load of rack k solves $T_k^* = T_{\max}$ with $T_{\max} = 27°\mathrm{C}$,

$$Q_k^\dagger(\mathbf{z}) = \frac{T_{\max} - T_s^*}{a_k(\mathbf{z})} - Q_o, \qquad Q^\dagger(\mathbf{z}) = \min_k Q_k^\dagger(\mathbf{z}), \qquad H(\mathbf{z}) = Q^\dagger(\mathbf{z}) - Q_B, \tag{5.8}$$

with H the headroom over today's load Q_B. By equation (5.7) the sensitivity of the inlet to the load is $\partial T_k^*/\partial Q_B^* = a_k$, and the gradient of the inlet with respect to the state at the binding load is $(Q_k^\dagger + Q_o)\,\nabla_{\mathbf{z}} a_k$, so equation (5.1) gives

$$\mathbf{g}_k = \nabla_{\mathbf{z}} Q_k^\dagger = -\frac{Q_k^\dagger + Q_o}{a_k}\,\nabla_{\mathbf{z}} a_k, \qquad \frac{\partial Q_k^\dagger}{\partial g_u} = \frac{(Q_k^\dagger + Q_o)\,G_u}{a_k\left(\sum_u G_u\right)^2}, \qquad \frac{\partial Q_k^\dagger}{\partial \beta_k} = -\frac{Q_k^\dagger + Q_o}{a_k \sum_u C_u}, \tag{5.9}$$

in kW, every other entry zero; the same vector follows by differentiating equation (5.8) directly, since $(T_{\max} - T_s^*)/a_k^2 = (Q_k^\dagger + Q_o)/a_k$. A coil below its rating lowers $\sum_u G_u$, raises the rise per kilowatt and lowers every binding load: the log-derate entries are positive. More recirculation at rack k lowers its own binding load and no other: the fraction's entry is negative and sits in rack k's column alone.

5.1.4 The planner's questions (step 1)

The planner asks four questions of window 1 of hall B, and step 1 writes each as a probability or a variance over the reduced state, with the unknown on the left of the bar and the readings and the priors on the right. Table 5.1 lists them. The first column is the question in the planner's words, the drill question first; the second the general form of its category in the notation of Chapter 2, for a quantity $f(\mathbf{z})$ derived from the state and a level f^* of it; the third its mathematical form for hall B; the fourth its category; the fifth the computation of Section 2.3.6 that answers it; and the last the subsection where it is answered. The first and the last are estimation: the binding load with its band, and the split of the band's variance. The second and third are predictions under a threshold: which rack binds first, and whether the binding load clears the design load with the assurance the policy requires.

Table 5.1: The questions asked of hall B in window 1. $\mathbf{y}$ is the eight readings and the nine priors; $Q^{\dagger}$ and $Q_k^{\dagger}$ are the binding loads of equation (5.8); κ is the assurance level.

Planner's question	General form	Mathematical form	Category	Computation	Answered in
How much load can hall B take before an inlet crosses the limit, and how sure is that number?	$\Pr(f(\mathbf{z}) \le f^* \mid \mathbf{y})$, quantiles	$Q^{\dagger}$ and $H = Q^{\dagger} - Q_B$: median and 90 % band	estimation	posterior	5.3.1
Which rack binds first?	$\Pr(f(\mathbf{z}) \ge f^* \mid \mathbf{y})$ over a family	$\Pr(\arg\min_k Q_k^{\dagger} = j \mid \mathbf{y})$, $j = \mathrm{B1}, \ldots, \mathrm{B5}$	prediction under a threshold	posterior	5.3.2
Can hall B carry its 340 kW design load with 95 % assurance?	$\Pr(f(\mathbf{z}) \ge f^* \mid \mathbf{y}) \ge \kappa$	$\Pr(Q^{\dagger} \ge 340\,\mathrm{kW} \mid \mathbf{y})$ against $\kappa = 0.95$	prediction under a threshold	posterior	5.3.3
What is the band made of?	$\mathrm{Var}(f(\mathbf{z}) \mid \mathbf{y})$, split by entry	the budget of $\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}$	estimation	posterior	5.3.4

5.1.5 The evidence as rows (step 2)

Step 2 attaches the evidence of window 1 of the night record, at 19:00, to the graph as rows on the reduced state. Eight readings are shown in Figure 5.1: the supply $y_{T_s} = 14.13$°C, by a sensor of standard deviation 0.1 K; the return air $y_{T_r} = 28.05$°C and the three discharge temperatures 18.49, 18.35 and 18.81°C, each of standard deviation 0.3 K; the hall load meter $y_{Q_B} = 277.20$ kW, of standard deviation 4 kW, which reads Q_B directly; and the inlets of racks B2 and B4, 21.58 and 21.99°C, of standard deviation 0.3 K. Nine priors carry what was known before the window. The chiller's set point puts the supply at 14°C to within 0.5 K, $T_s \sim \mathcal{N}(14,\ 0.5^2)$; the commissioning record puts each coil at its rating to within about 15 %, $g_u \sim \mathcal{N}(0,\ 0.15^2)$; and the airflow survey gives the fractions $\beta^{\mathrm{s}} = (0.20,\ 0.30,\ 0.50,\ 0.40,\ 0.35)$ for racks B1 to B5, each with standard deviation 0.08, $\beta_k \sim \mathcal{N}(\beta_k^{\mathrm{s}},\ 0.08^2)$. The load has no prior row: the meter reads it.

Which row acts on which entry. On the reduced state a reading row sees the free entries through the row of $\mathbf{N}_{\mathrm{b}}$ of the temperature it reads, equation (4.47). By equation (5.5) the return air and the discharge temperatures depend on $(T_s, Q_B, g_{B1}, g_{B2}, g_{B3})$; a rack inlet on those five and its own fraction; the supply sensor on T_s alone and the meter on Q_B alone. With $*$ for an entry that can be nonzero, the eight reading rows have the pattern

$$\mathbf{J}_{\mathrm{read}} = \begin{bmatrix} * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & 0 & 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & 0 & * & 0 & 0 & 0 \\ * & * & * & * & * & 0 & 0 & 0 & * & 0 \end{bmatrix}, \tag{5.10}$$

$$\mathbf{g}_{\mathrm{B3}}^{\mathsf{T}} = (0,\ 0,\ *,\ *,\ *,\ 0,\ 0,\ *,\ 0,\ 0),$$

rows in the order supply, return air, the three discharges, meter, the inlets of B2 and B4; columns in the order of equation (5.4). Each of the nine priors adds one row with a single nonzero in its own column. The second line is not a row: it is the pattern of the gradient (5.9)

of rack B3's binding load. Three facts of the graph are read here before any number. No reading has a nonzero in the columns of β_{B1}, β_{B3} or β_{B5}: each holds only its prior row, the eight readings have rank at most seven on ten entries, and read as observability, the airflow of the three unmeasured racks is known only from the survey. The gradient of rack B3's binding load has an entry in one of those three columns. And it has none in the columns of T_s and Q_B, the two entries the readings fix best, because the scenario pins the supply and frees the load.

Modeling choices at the rows. Three choices make these rows. Each reading is taken as Gaussian, unbiased and independent of the others given the state: this makes each reading one normalized row and the objective a sum of squares, and it fails if a sensor carries a bias the rows do not model. Each rack carries one recirculation fraction, constant over the window and over the height of its inlet, with an independent prior: this gives each inlet one entry of its own, and it fails if the inlet is stratified or if the racks share a cause of recirculation (Section 5.6.5). The airflow survey is the only knowledge of the three unmeasured racks' fractions: by the pattern (5.10) no reading sees them, so their prior rows are the whole of what the fit knows of them, and the answer inherits whatever the survey got wrong; the survey's width is varied in Section 5.4.3. Two further choices fix which evidence is used. One window is read, window 1 at 19:00, rather than the whole night: this keeps the state that of one steady condition, and it gives up the information that later windows would add. The window is the healthy plant: the night record shows unit B2's coil derating later in the night, and this chapter, which asks about the hall as drawn, leaves the derate aside.

5.2 The posterior of the state (step 3)

All four questions of Table 5.1 are read from one posterior: that of the reduced state of hall B under the seventeen rows of step 2, with the pinned quantities of Section 5.1.2.

5.2.1 The rows and the normal equations

The rows of the reduced state are the evidence rows at the point $\mathbf{z}_{\text{full}}(\mathbf{z})$ of the set, equation (4.46), with the temperatures of equation (5.5); stacked in the order of step 2, each divided by its standard deviation, they form the residual vector

$$\mathbf{r}(\mathbf{z}) = \begin{bmatrix} (T_s - y_{T_s})/0.1 \\ (T_r(\mathbf{z}) - y_{T_r})/0.3 \\ (T_{d,u}(\mathbf{z}) - y_{T_{d,u}})/0.3, \quad u = \mathrm{B1}, \mathrm{B2}, \mathrm{B3} \\ (Q_B - y_{Q_B})/4 \\ (T_k(\mathbf{z}) - y_{T_k})/0.3, \quad k = \mathrm{B2}, \mathrm{B4} \\ (T_s - 14)/0.5 \\ g_u/0.15, \quad u = \mathrm{B1}, \mathrm{B2}, \mathrm{B3} \\ (\beta_k - \beta_k^{\mathrm{s}})/0.08, \quad k = \mathrm{B1}, \ldots, \mathrm{B5} \end{bmatrix}, \qquad \mathcal{L}(\mathbf{z}) = \tfrac{1}{2}\,\mathbf{r}(\mathbf{z})^{\mathsf{T}}\mathbf{r}(\mathbf{z}), \tag{5.11}$$

seventeen rows on ten entries, every row dimensionless. The reduced Jacobian is $\mathbf{J} = \mathbf{J}_{\text{rows}}\mathbf{N}$ of equation (4.47), evaluated at each iterate because the rows are nonlinear in the state through e^{g_u} and through the products $\beta_k(T_r - T_c)$. The MAP state $\hat{\mathbf{z}}$ minimizes the objective; each iteration solves the normal equations of equation (4.29) on the reduced rows,

$$(\mathbf{J}^{\mathsf{T}}\mathbf{J})\,\Delta\mathbf{z} = -\mathbf{J}^{\mathsf{T}}\mathbf{r}(\mathbf{z}), \qquad \mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + \Delta\mathbf{z}, \tag{5.12}$$

with the temperatures re-solved by equation (5.5) at every iterate. The iteration starts from the design sheet's state, the supply and the load as read, the coils at their ratings and the survey's

Table 5.2: Gauss–Newton iterations of hall B in window 1, equation (5.12), from the design sheet's state with the read supply 14.13°C and load 277.20 kW.

Iteration k [–]	Objective $\mathcal{L}(\mathbf{z}^{(k)})$ [–]	$\lVert\mathbf{J}^{\mathsf{T}}\mathbf{r}\rVert$ [–]
0	3.92002	1.53×10^{2}
1	0.23953	2.80×10^{-1}
2	0.23949	3.78×10^{-4}
3	0.23949	$< 10^{-6}$
4	0.23949	$< 10^{-6}$

fractions. With every row normalized the information matrix is $\mathbf{\Lambda} = \mathbf{J}^{\mathsf{T}}\mathbf{J}$ (with unnormalized rows it would be $\mathbf{J}^{\mathsf{T}}\mathbf{W}\mathbf{J}$), and at the minimum the Laplace covariance is

$$\mathbf{\Sigma} = \left(\mathbf{J}^{\mathsf{T}}\mathbf{J}\right)^{-1}\Big|_{\hat{\mathbf{z}}}, \qquad \mathbf{z} \mid \mathbf{y} \approx \mathcal{N}(\hat{\mathbf{z}}, \mathbf{\Sigma}), \tag{5.13}$$

the Gauss–Newton form of equation (4.38), which drops the curvature term of the nonlinear rows; Section 5.4.1 sets the full Hessian beside it.

The block structure. The pattern (5.10) fixes the shape of $\mathbf{\Lambda}$ before any number. The columns of β_{B1}, β_{B3} and β_{B5} each hold one prior row, so

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_{\mathrm{o}} & \mathbf{0} \\ \mathbf{0} & 0.08^{-2}\,\mathbf{I}_3 \end{bmatrix}, \qquad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Lambda}_{\mathrm{o}}^{-1} & \mathbf{0} \\ \mathbf{0} & 0.08^{2}\,\mathbf{I}_3 \end{bmatrix}, \tag{5.14}$$

with $\mathbf{\Lambda}_{\mathrm{o}}$ the 7×7 block of the entries the readings see, $(T_s, Q_B, g_{B1}, g_{B2}, g_{B3}, \beta_{B2}, \beta_{B4})$, and the entries reordered to put the three unseen fractions last. The objective separates in those three fractions, so their MAP values are the survey's and their posterior is the prior exactly, uncorrelated with every other entry.

5.2.2 The fit of window 1

The iterations. Table 5.2 runs the normal equations (5.12). Each row is an iteration; the columns give the objective and the norm of its gradient $\mathbf{J}^{\mathsf{T}}\mathbf{r}$. The objective falls from 3.92002 to 0.23953 in the first step and to 0.23949 in the second, and the gradient's norm falls below 10^{-6} by the third.

The Jacobian at the MAP state. The eight reading rows of $\mathbf{J}(\hat{\mathbf{z}})$, in the order of equation (5.11) and with the columns of equation (5.4), are

$$\mathbf{J}_{\mathrm{read}}(\hat{\mathbf{z}}) = \begin{bmatrix} 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3.333 & 0.162 & -15.610 & -15.819 & -15.105 & 0 & 0 & 0 & 0 & 0 \\ 3.333 & 0.050 & -36.933 & -4.911 & -4.690 & 0 & 0 & 0 & 0 & 0 \\ 3.333 & 0.049 & -4.703 & -37.281 & -4.551 & 0 & 0 & 0 & 0 & 0 \\ 3.333 & 0.054 & -5.195 & -5.264 & -36.076 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.250 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3.333 & 0.086 & -15.610 & -15.819 & -15.105 & 0 & 31.884 & 0 & 0 & 0 \\ 3.333 & 0.092 & -15.610 & -15.819 & -15.105 & 0 & 0 & 0 & 31.884 & 0 \end{bmatrix}, \tag{5.15}$$

the second column in kW^{-1}, the first in K^{-1}, the others dimensionless; the nine prior rows are the reciprocals of their standard deviations in their own columns. The zeros sit where the pattern (5.10) places them, and the reading rows have rank 7 of 10. Every temperature row has $1/0.3$ in the column of T_s: a warmer supply warms every temperature of the hall alike. The two inlet rows repeat the return-air row in the log-derate columns, because by the air balance (5.6) a derate leaves the rise unchanged and moves the inlet as it moves the return air; their entry 31.884 in their own fraction's column is the rise divided by the sensor's 0.3 K.

Table 5.3: The MAP state of hall B in window 1 with the posterior and prior spreads of each entry.

Entry [unit]	$\hat{z}_i$ [entry's]	$\sqrt{\Sigma_{ii}}$ [entry's]	Prior sd [entry's]	Ratio [–]	Informed by
T_s [°C]	14.1295	0.0977	0.5	0.195	supply sensor and set point
Q_B [kW]	276.9528	3.7224	none	–	meter and the temperature readings
g_{B1} [–]	−0.0151	0.0285	0.15	0.190	discharge, return and inlet readings
g_{B2} [–]	−0.0018	0.0282	0.15	0.188	discharge, return and inlet readings
g_{B3} [–]	−0.0479	0.0293	0.15	0.195	discharge, return and inlet readings
β_{B1} [–]	0.2000	0.0800	0.08	1.000	survey only
β_{B2} [–]	0.3168	0.0320	0.08	0.400	rack-B2 inlet sensor and survey
β_{B3} [–]	0.5000	0.0800	0.08	1.000	survey only
β_{B4} [–]	0.3670	0.0320	0.08	0.400	rack-B4 inlet sensor and survey
β_{B5} [–]	0.3500	0.0800	0.08	1.000	survey only

The covariance as numbers. The block of $\mathbf{\Sigma}$ for the load and the three log-derates, $(Q_B, g_{B1}, g_{B2}, g_{B3})$, and its correlations are

$$\mathbf{\Sigma}_{Q,g} = \begin{bmatrix} 13.85629 & 0.02289 & 0.02230 & 0.02439 \\ 0.02289 & 0.00081 & -0.00012 & -0.00013 \\ 0.02230 & -0.00012 & 0.00080 & -0.00012 \\ 0.02439 & -0.00013 & -0.00012 & 0.00086 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0.216 & 0.212 & 0.224 \\ 0.216 & 1 & -0.144 & -0.150 \\ 0.212 & -0.144 & 1 & -0.148 \\ 0.224 & -0.150 & -0.148 & 1 \end{bmatrix},$$

in kW^2, kW and dimensionless units. The load is correlated positively with every log-derate, and the log-derates negatively with each other, and the rows fix those signs. The return-air reading fixes $T_r - T_s = (Q_B + Q_o)/\sum_u G_u$, the heat over the three conductances together: a larger load is explained equally well by larger conductances; and with the sum of the conductances fixed, a larger g_{B1} is explained by smaller g_{B2} and g_{B3}. The three unseen fractions keep the survey's variance 0.08^2, with no covariance, as equation (5.14) requires.

The MAP state. Table 5.3 gives the MAP state with its posterior spreads $\sqrt{\Sigma_{ii}}$, the prior spread of each entry and their ratio. Each row is an entry; the last column says which rows inform it. The supply and the log-derates are known to about a fifth of their prior spread. The fractions of racks B2 and B4 move from the survey and their spread falls to 0.032, two fifths of the survey's, the effect of their inlet sensors. The fractions of racks B1, B3 and B5 stay at the survey's values with the survey's spread, a ratio of 1.000.

By equation (5.5) the temperatures at the MAP state are $T_r = 28.090$°C and $T_c = 18.525$°C, and the five inlets 20.438, 21.555, 23.307, 22.035 and 21.872°C for racks B1 to B5: rack B3's inlet, which no sensor reads, is the warmest of the five.

The fit of the readings. The plain count of the fit's freedom, rows less entries, is $17 - 10 = 7$, and twice the objective, 0.479, sits far below it. The count is the wrong yardstick here. The three prior rows of the unseen fractions sit at their means: they add nothing to the objective, and each cancels its own entry in the count. Of the remaining count, most belongs to prior rows, whose residuals test the commissioning record and the survey, not the readings. The test of the

readings is the eight reading rows alone, against the freedom the fit leaves them, the leverage count $\sum_k (1 - h_{kk})$ with h_{kk} the diagonal of $\mathbf{J}_{\text{read}} \boldsymbol{\Sigma} \mathbf{J}_{\text{read}}^{\mathsf{T}}$. The leverages are 0.955 for the supply, 0.328 for the return air, 0.880, 0.881 and 0.879 for the discharges, 0.866 for the meter and 0.871 for each inlet: the return air is the one reading the other rows check. The count is 1.468, the reading rows' χ^2 is 0.085, and its upper tail probability is 0.894: the readings sit well inside their noise.

5.2.3 The exact posterior by reweighted draws

The Laplace normal (5.13) approximates the posterior $p(\mathbf{z} \mid \mathbf{y}) \propto \mathrm{e}^{-\mathcal{L}(\mathbf{z})}$, and the exact posterior is computed beside it by reweighting draws from it. Draw N states $\mathbf{z}_i$ from $\mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$, the Laplace draws; give each the weight of the exact posterior over the normal that produced it,

$$w_i \propto \exp\Big[-\mathcal{L}(\mathbf{z}_i) + \mathcal{L}(\hat{\mathbf{z}}) + \tfrac{1}{2}\,(\mathbf{z}_i - \hat{\mathbf{z}})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{z}_i - \hat{\mathbf{z}})\Big], \qquad \sum_i w_i = 1, \qquad N_{\text{eff}} = \frac{1}{\sum_i w_i^2}, \tag{5.16}$$

and read any posterior probability as the summed weight of the draws in which its event holds, and any quantile from the weighted distribution function. The effective sample size N_{eff} measures how far the weights are from equal: it equals N when the Laplace normal is the exact posterior. Each draw is pushed through the exact physics, the chain (5.5) and the binding loads (5.8) of all five racks. With $N = 50\,000$ draws the effective sample size is 49 638: the weights are nearly equal, and the posterior of the state is nearly the Laplace normal.

5.3 The answers (step 4)

Every answer is read from the pair $(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$ of Section 5.2.2. A derived quantity is carried by its gradient at $\hat{\mathbf{z}}$, the reported value is the state route the question needs, and the exact posterior of Section 5.2.3 stands beside it with the gap.

5.3.1 The binding load and the headroom

Planner's question. How much load can hall B take before an inlet crosses the limit, and how sure is that number? The question asks for the binding load of the hall at the drawing's supply, with its band, and for the headroom it leaves over today's load.

General form. An estimation question about a derived quantity asks for its posterior distribution, equation (2.9),

$$\Pr(f(\mathbf{z}) \le f^* \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) \le f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

summarized by a location and by a central credible interval, equation (2.11), here the central 90 % interval, between the 5 % and the 95 % points.

Specialized form. The question is equation (2.9) with $\mathbf{z}$ the reduced state (5.4), $f(\mathbf{z}) = Q^{\dagger}(\mathbf{z})$ of equation (5.8) for the binding load and $f(\mathbf{z}) = H(\mathbf{z})$ for the headroom, and $p(\mathbf{z} \mid \mathbf{y})$ the Laplace normal (5.13). The hall's binding load is the minimum over the five racks, and which rack attains it changes across the posterior, so no single gradient carries it. The reported answer is joint over the racks: the Laplace normal is integrated by draws, each pushed through the exact binding loads (5.8) of all five racks,

$$\mathbf{z}_i \sim \mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma}), \qquad Q_i^{\dagger} = \min_k Q_k^{\dagger}(\mathbf{z}_i), \qquad H_i = Q_i^{\dagger} - Q_{B,i}, \qquad \hat{F}(q) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\big[Q_i^{\dagger} \le q\big], \tag{5.17}$$

with $Q_{B,i}$ the load entry of draw i and $\hat{F}$ the distribution function of the draws, from which the median and the 5 % and 95 % points are read, and those of the headroom from the H_i in the same way. This is the state route; the exact posterior is the same draws under the weights (5.16).

Two routes on the binding rack alone stand beside the joint answer, and each shows what it corrects. Each rack's binding load is carried by its gradient (5.9) with variance $\mathbf{g}_k^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}_k$, the implicit-function form (5.1); the delta form reads the rack with the smallest value at $\hat{\mathbf{z}}$ with that rack's gradient, and the headroom with the gradient $\mathbf{g}_k - \mathbf{e}_Q$, with $\mathbf{e}_Q$ the unit vector of the load. But a rack's binding load is a ratio, $(T_{\max} - T_s^*)/a_k - Q_o$, and a symmetric normal on a ratio misplaces its tails. The inversion route keeps the ratio exact on the one rack. The rise per kilowatt a_k of equation (5.7) is linear in the fraction β_k and nearly linear in the log-derates over their posterior width, so its posterior is carried by the delta method,

$$a_k \mid \mathbf{y} \approx \mathcal{N}\big(a_k(\hat{\mathbf{z}}),\ s_a^2\big), \qquad s_a^2 = (\nabla_{\mathbf{z}} a_k)^{\mathsf{T}}\,\boldsymbol{\Sigma}\,\nabla_{\mathbf{z}} a_k, \tag{5.18}$$

and the binding load, a decreasing function of a_k, takes its quantiles from those of a_k in reverse order, as Chapter 2 states for a monotone function: the median is the binding load at the median of a_k, and the 90 % band is

$$\Big[\ \frac{T_{\max} - T_s^*}{a_k(\hat{\mathbf{z}}) + 1.645\, s_a} - Q_o,\ \frac{T_{\max} - T_s^*}{a_k(\hat{\mathbf{z}}) - 1.645\, s_a} - Q_o\ \Big], \tag{5.19}$$

with $\Phi(1.645) = 0.95$. The density of the binding load follows from that of a_k by the change of variables (4.55), with $a = (T_{\max} - T_s^*)/(q + Q_o)$ and the change-of-variable factor $|\mathrm{d}a/\mathrm{d}q| = (T_{\max} - T_s^*)/(q + Q_o)^2$,

$$p_{Q^\dagger}(q) = \frac{1}{s_a}\,\phi\!\left(\frac{a(q) - a_k(\hat{\mathbf{z}})}{s_a}\right)\frac{T_{\max} - T_s^*}{(q + Q_o)^2}, \tag{5.20}$$

with ϕ the standard normal density. The delta form is the normal $\mathcal{N}(Q^\dagger(\hat{\mathbf{z}}),\ \mathbf{g}_k^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}_k)$ with the band $Q^\dagger(\hat{\mathbf{z}}) \pm 1.645\sqrt{\mathbf{g}_k^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}_k}$. The inversion route and the delta form read the binding rack alone; the joint draws (5.17) and the exact posterior let any rack bind in any draw.

Building blocks. The MAP state $\hat{\mathbf{z}}$ and the covariance $\boldsymbol{\Sigma}$ (Section 5.2.2); the rise per kilowatt (5.7) and the binding loads (5.8) at $\hat{\mathbf{z}}$ and in each draw; the joint draws (5.17); the gradients (5.9); the inversion (5.19); the reweighted draws (5.16).

The racks at the scenario. Table 5.4 gives, for each rack, its rise per kilowatt and binding load at $\hat{\mathbf{z}}$, the standard deviation of the binding load by the delta method, and the nonzero entries of its gradient. The inlet lines of equation (5.7) are drawn in Figure 5.2: the horizontal axis is the hall load at the scenario, the vertical axis the scenario inlet temperature, one line per rack, the horizontal black line the 27°C limit and the dots the binding loads where the lines cross it; the shaded wedge is the 90 % band of rack B3's rise per kilowatt, and the dashed vertical line the 340 kW design load.

Rack B3 binds first at the MAP state, at 396.46 kW, and racks B4 and B5 follow at 461.88 and 471.78 kW. The order is the order of the fractions, since the racks share the first term of a_k. The fitted coils run slightly below their ratings, which raises the rise per kilowatt of every rack and moves rack B3's binding load down from the sheet's 410.00 kW. The spreads follow the gradients: the measured racks B2 and B4 have the narrowest bands, 21.06 and 18.68 kW, because their sensors cut their fractions' spread to 0.032; rack B1 has the widest, 74.00 kW, because its small rise per kilowatt makes its fraction's lever the longest.

Table 5.4: The five racks of hall B at the scenario node, at the MAP state: the rise per kilowatt (5.7), the binding load (5.8) with its delta-method standard deviation, and the nonzero entries of its gradient (5.9).

Rack	a_k [K/kW]	$Q_k^\dagger$ [kW]	sd$(Q_k^\dagger)$ [kW]	$\partial Q_k^\dagger/\partial g_{B1}$ [kW]	$\partial Q_k^\dagger/\partial g_{B2}$ [kW]	$\partial Q_k^\dagger/\partial g_{B3}$ [kW]	$\partial Q_k^\dagger/\partial \beta_k$ [kW]
B1	0.021983	581.36	74.00	439.0	444.9	424.8	−896.7
B2 (measured)	0.025876	492.41	21.06	316.9	321.1	306.6	−647.2
B3	0.031983	396.46	34.96	207.4	210.2	200.7	−423.6
B4 (measured)	0.027549	461.88	18.68	279.5	283.3	270.5	−571.0
B5	0.026983	471.78	49.12	291.4	295.3	282.0	−595.2

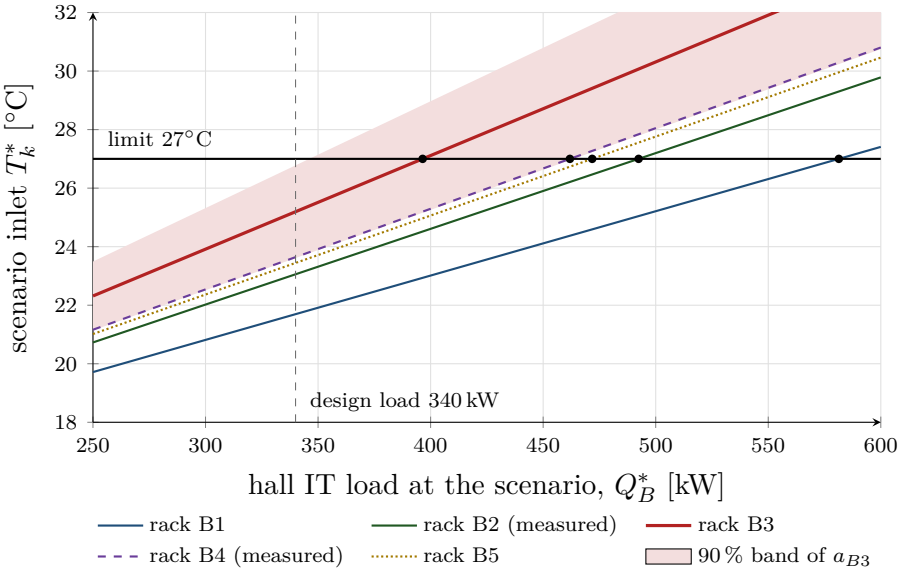


Figure 5.2: The scenario inlets of hall B at the drawing's supply $T_s^* = 14°\text{C}$, $T_k^* = 14 + a_k\,(Q_B^* + 10)$ by equation (5.7), with the rise per kilowatt a_k of each rack at the MAP state (Table 5.4). The dots are the binding loads, where each line meets the 27°C limit. Shaded: the lines of rack B3 with a_{B3} at its 5 % and 95 % points, $0.031983 \mp 1.645 \times 0.002751\,\text{K/kW}$; the wedge meets the limit at the ends of the inversion band.

With $N = 50\,000$ draws (seed 5), the joint draws (5.17) give the reported binding load of the hall,

$$\text{median } Q^{\dagger} = 395.21\,\text{kW}, \qquad Q^{\dagger} \in [345.89,\ 450.43]\,\text{kW with } 90\,\%\text{ probability}, \tag{5.21}$$

and the headroom over today's fitted load a median of 118.43 kW, with a 90 % band from 69.41 to 172.92 kW. The exact posterior, the same draws reweighted, gives the binding load a median of 394.92 kW and a band from 345.67 to 450.17 kW, and the headroom 118.35 kW, from 69.37 to 172.73 kW. Its gaps from the reported answer, exact minus reported, are -0.29 kW at the median, -0.22 kW at the 5 % point and -0.26 kW at the 95 % point: the posterior of the state is nearly the Laplace normal.

Beside the joint answer stand the two routes on rack B3 alone. Its rise per kilowatt is $a_{B3} = 0.031983\,\text{K/kW}$ with $s_a = 0.002751\,\text{K/kW}$, a relative spread of 0.086, and by equation (5.19) the inversion route gives rack B3's binding load a median of 396.46 kW and the band [346.08, 463.45] kW. The delta form puts it at 396.46 kW with standard deviation 34.96 kW and the band [338.96, 453.97] kW, and the headroom at 119.51 kW, standard deviation 34.75 kW, band [62.35, 176.67] kW. Over the planning figure of 300 kW for hall B's load today, the joint draws leave a median of 95.21 kW, band [45.89, 150.43] kW; rack B3 alone at the MAP state leaves 96.46 kW.

The routes differ in their tails, not in their centers. The density of the binding load is drawn in Figure 5.3: the horizontal axis is the binding load, the vertical axis its posterior density; the solid curve is the inversion route's density (5.20) for rack B3, the dashed curve the delta form's normal; the red area is the inversion route's probability below the 340 kW design load, the dotted line the stated truth of the record, a check marker for Section 5.4.4; and the three bars above the curves are the 90 % bands of the joint draws and of the two routes on rack B3, with their medians. The inversion route shows why the delta form fails at the left end: the binding load is a ratio over a_{B3}, skewed to the right, so its lower end, 346.08 kW, lies close to the joint 345.89 kW, where the delta form's symmetric band reaches down to 338.96 kW and overstates the left tail. At the upper end the joint band stops at 450.43 kW, short of the inversion route's 463.45 kW. The gap is rack B5, which binds first in 8 % of the draws (Section 5.3.2): where rack B3's fraction is small, another rack's binding load is lower, and the minimum over the racks cuts the upper tail that rack B3 alone would stretch. That is the difference between the joint answer and the worst marginal: the binding rack read alone, even with its ratio kept exact, places the upper end too high.

5.3.2 The rack that binds first

Planner's question. Which rack binds first? The question asks how sure the review can be that rack B3, and not another rack, sets the hall's binding load.

General form. A prediction under a threshold is a tail of a derived quantity under the posterior, equation (2.12),

$$\Pr(f(\mathbf{z}) \geq f^{*} \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) \geq f^{*}\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

asked here over a family of events, one per rack.

Specialized form. For rack j the question is equation (2.12) with $f(\mathbf{z}) = \min_{k \neq j} Q_k^{\dagger}(\mathbf{z}) - Q_j^{\dagger}(\mathbf{z})$, $f^{*} = 0$, and the binding loads of equation (5.8): the event that rack j's binding load is the smallest,

$$P_j = \Pr\Big(\arg\min_{k} Q_k^{\dagger}(\mathbf{z}) = j \;\Big|\; \mathbf{y}\Big), \qquad \sum_{j} P_j = 1. \tag{5.22}$$

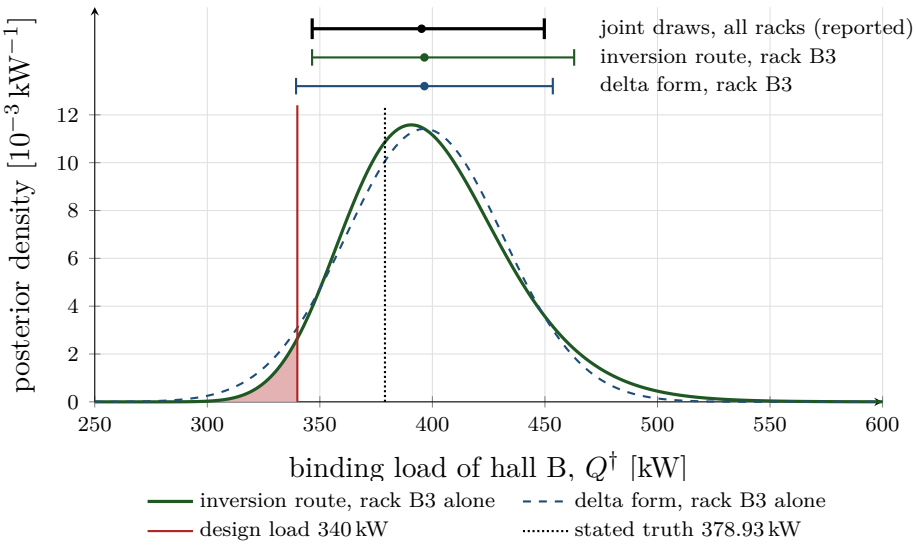


Figure 5.3: The posterior of the binding load of hall B. Solid: the inversion route's density (5.20) for rack B3 alone, from $a_{B3} \mid \mathbf{y} \approx \mathcal{N}(0.031983,\ 0.002751^2)$ K/kW with $T_{\max} - T_s^* = 13$ K and $Q_o = 10$ kW. Dashed: the delta form for rack B3 alone, $\mathcal{N}(396.46,\ 34.96^2)$ kW. Red area: the inversion route's probability below the design load. Dotted: the stated truth of the record, a check marker (Section 5.4.4). Bars: the 90 % bands, with their medians as dots; the top bar is the reported band, [345.89, 450.43] kW, that of the minimum over the five racks from the joint draws (5.17); the exact posterior's band, [345.67, 450.17] kW, lies within 0.3 kW of it at each end and is not drawn.

Table 5.5: The probability that each rack of hall B binds first, equation (5.22), from 50 000 Laplace draws and from the same draws reweighted to the exact posterior.

Rack	$Q_k^\dagger(\hat{\mathbf{z}})$ [kW]	P_j, Laplace draws [–]	P_j, exact posterior [–]
B3	396.46	0.8752	0.8751
B5	471.78	0.0813	0.0813
B4 (measured)	461.88	0.0400	0.0402
B2 (measured)	492.41	0.0022	0.0022
B1	581.36	0.0013	0.0012

The event is a region of the state bounded by the surfaces on which two binding loads tie, and no single gradient carries it; the state route is the Laplace normal (5.13) integrated over that region by its draws, each pushed through the exact binding loads of all five racks, and the exact posterior is the same count under the weights (5.16). Since the racks share the first term of a_k, the event is nearly the event that rack j has the largest fraction, and its probability is set by the five fractions' posteriors: the survey's for racks B1, B3 and B5, the narrower sensor-informed ones for racks B2 and B4.

Building blocks. The Laplace draws and the weights (5.16) (Section 5.2.3); the binding loads (5.8) in each draw.

The binding rack. Table 5.5 gives the probabilities: each row is a rack, the columns its binding load at $\hat{\mathbf{z}}$ and the probability that it binds first, by the Laplace draws and by the exact posterior.

Rack B3 binds first with probability 0.875 by both routes. The second most likely is rack B5, at 0.081, which is unmeasured too, although rack B4 has the lower binding load at $\hat{\mathbf{z}}$: rack B5's fraction keeps the survey's spread, so it overtakes rack B3 more often than the measured rack B4, whose fraction its sensor holds at 0.367 ± 0.032. The two measured racks together bind first with probability below 0.05: the inlets the hall reads are the inlets least likely to set its headroom.

5.3.3 The design load against the assurance level

Planner's question. Can hall B carry its 340 kW design load with 95 % assurance? The question asks for the posterior probability that the binding load is at least the design load, set against the policy's assurance level.

General form. A requirement is accepted when its posterior probability reaches the assurance level, equation (2.12),

$$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y}) \geq \kappa.$$

Specialized form. The question is equation (2.12) with $f(\mathbf{z}) = Q^\dagger(\mathbf{z})$, the required value $f^* = Q_d = 340\,\text{kW}$ and $\kappa = 0.95$. The reported probability is joint over the racks: the share of the joint draws (5.17) in which the minimum over the five racks is at least the design load,

$$\Pr(Q^\dagger \geq Q_d \mid \mathbf{y}) \approx \hat{P} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[Q_i^\dagger \geq Q_d], \qquad \text{se}(\hat{P}) = \sqrt{\hat{P}\,(1-\hat{P})/N}, \tag{5.23}$$

with se the Monte Carlo standard error of the share; the exact posterior sums the weights (5.16) of the same draws. On rack B3 alone, by the inversion route, the event $Q_{B3}^\dagger \geq Q_d$ is the event that rack B3's rise per kilowatt stays at or below the rise at which the limit binds at the design load,

$$\Pr(Q^\dagger \geq Q_d \mid \mathbf{y}) = \Pr(a_{B3} \leq a_d \mid \mathbf{y}) = \Phi\!\left(\frac{a_d - a_{B3}(\hat{\mathbf{z}})}{s_a}\right), \qquad a_d = \frac{T_{\max} - T_s^*}{Q_d + Q_o}, \tag{5.24}$$

by the normal (5.18). The delta form reads the same probability off its normal, $\Phi((Q^\dagger(\hat{\mathbf{z}}) - Q_d)/\sqrt{\mathbf{g}_{B3}^\top \boldsymbol{\Sigma} \mathbf{g}_{B3}})$.

Building blocks. The joint draws (5.17) and their weights (5.16); the rise per kilowatt $a_{B3}(\hat{\mathbf{z}})$ and its spread s_a (Section 5.3.1); the delta form's standard deviation of Table 5.4; the normal distribution function Φ.

The design load. By equation (5.23) the joint draws give the reported value,

$$\Pr(Q^\dagger \geq 340\,\text{kW} \mid \mathbf{y}) = 0.9698,$$

with a Monte Carlo standard error of 0.0008. The exact posterior gives 0.9692, a gap of -0.0006 taken from the unrounded values, smaller than the standard error. On rack B3 alone, with $a_d = 13/350 = 0.037143\,\text{K/kW}$, equation (5.24) gives 0.9696 by the inversion route, and the delta form gives 0.9468. Table 5.6 sets the four routes side by side for questions one and three: each row is a route, the columns the location and the 90 % band of the binding load and the probability of the design load.

The joint draws, the exact posterior and the inversion route clear the assurance level of 0.95; the delta form alone falls below it. The four are stated here as they stand, and none is chosen over another by a rule. The gap has one cause. The binding load is a constant over the rise per kilowatt, less a constant, so a symmetric uncertainty in the rise becomes a skewed one in the

Table 5.6: The binding load of hall B and the probability that it clears the 340 kW design load, by four routes. The joint draws and the exact posterior take the minimum over the five racks; the inversion route and the delta form read rack B3 alone.

Route	Median [kW]	5 % point [kW]	95 % point [kW]	$\Pr(Q^{\dagger} \geq 340\,\text{kW})$ [–]
Joint draws, all racks (reported)	395.21	345.89	450.43	0.9698
Exact posterior, all racks	394.92	345.67	450.17	0.9692
Inversion route, rack B3 alone	396.46	346.08	463.45	0.9696
Delta form, rack B3 alone	396.46	338.96	453.97	0.9468

load: compressed on the left, where the rise is large, and stretched on the right. The delta form puts a symmetric normal on a quantity whose left tail is short, and so overstates the probability that the binding load falls below the design load, $1 - 0.9468$ against $1 - 0.9698$. The decision sits close enough to the assurance level for that difference to cross it. The minimum over the racks, which moves the upper end of the band, barely moves this probability, which lives in the left tail where rack B3 binds.

5.3.4 The budget of the band

Planner's question. What is the band made of? The question asks which entries of the state, and so which rows of the evidence, stand behind the band of the binding load.

General form. The variance of a derived quantity under the posterior, equation (2.10),

$$\operatorname{Var}(f(\mathbf{z}) \mid \mathbf{y}) = \int \left(f(\mathbf{z}) - \mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}]\right)^2 p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

split into one share per entry.

Specialized form. The question is equation (2.10) with $f(\mathbf{z}) = Q^{\dagger}_{B3}$ and $p(\mathbf{z} \mid \mathbf{y})$ the Laplace normal, on which the variance is $\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}$ with $\mathbf{g} = \mathbf{g}_{B3}$ of equation (5.9). It splits into one share per entry by the covariance split

$$\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g} = \sum_i g_i\,(\boldsymbol{\Sigma}\mathbf{g})_i, \qquad s_i = \frac{g_i\,(\boldsymbol{\Sigma}\mathbf{g})_i}{\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}} = \frac{g_i^2\,\Sigma_{ii} + \sum_{j \neq i} g_i\,g_j\,\Sigma_{ij}}{\mathbf{g}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}}, \qquad \sum_i s_i = 1, \tag{5.25}$$

each entry's diagonal term with half of every cross term it takes part in. The shares are the same on the inversion route: by equation (5.9) the gradient of $Q^{\dagger}_{B3}$ is a constant times the gradient of a_{B3}, and a common factor cancels in the ratio. Three facts of the gradient and of $\boldsymbol{\Sigma}$ fix the shape of the budget before any number. The gradient has no entry in T_s or Q_B, so those shares are zero, although Q_B is correlated with the log-derates. The fraction β_{B3} has no covariance with any other entry, equation (5.14), so its share is its diagonal term alone, $(423.6 \times 0.08)^2$ over the variance. The three log-derate entries are positive and their covariances negative, so every cross term among them is negative, and the readings know the log-derates' combined effect better than their three diagonal terms suggest. With β_{B3} known exactly, the remaining variance is the quadratic form on the covariance conditioned on it, the Schur complement of Chapter 3, which by the block structure is $\boldsymbol{\Sigma}$ with the row and column of β_{B3} removed.

Building blocks. The gradient $\mathbf{g}_{B3}$ of Table 5.4; the covariance $\boldsymbol{\Sigma}$ of Section 5.2.2; the split (5.25).

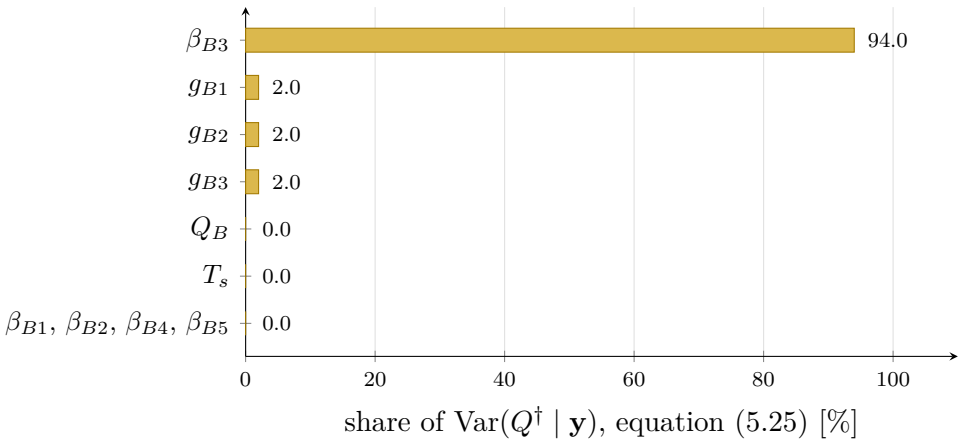


Figure 5.4: The variance budget of the binding load of hall B, equation (5.25), at the MAP state: each entry's diagonal term with half of its cross terms, in percent of $\mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g} = 34.96^2\,\mathrm{kW}^2$. The supply and the load have no share because the scenario pins the one and frees the other.

The budget. The shares are drawn in Figure 5.4, one bar per entry with the four other fractions together, the horizontal axis in percent of the variance. The survey's fraction for rack B3 carries 94.0 % of the variance, each log-derate 2.0 %, and the supply, the load and the other four fractions nothing. With β_{B3} known, the standard deviation of the binding load would be 8.59 kW, against 34.96 kW with it: the band of the hall's headroom is one airflow number.

5.4 The checks (step 5)

The answers of Section 5.3 rest on approximations and inputs that can be tested on this hall, each against the calculation it stands for.

5.4.1 The implicit-function gradient and the curvature term

The gradient (5.9) of rack B3's binding load was formed by the implicit function theorem, from the gradient of the scenario inlet and its slope with the load. Differencing the binding load itself numerically, entry by entry, gives the same vector to within 10^{-4} in every entry; at the binding load the scenario inlet is 27.0000°C and its slope 0.031983 K/kW, the rise per kilowatt of Table 5.4. The covariance (5.13) is the Gauss–Newton form, without the curvature term of the nonlinear rows. The full Hessian of the objective at $\hat{\mathbf{z}}$ gives a Newton covariance under which the standard deviation of the binding load is 34.962 kW, against 34.962 kW from the Gauss–Newton covariance: the curvature term moves the band by less than 0.005 %. The fit runs on the reduced state of Section 4.2.8, which Section 4.4.5 showed to be the limit of exact physical rows on the full state; hall B's rows are exact balances of the same kind, and that check is not repeated here.

5.4.2 The routes against the draws

The reported answers rest on the Laplace normal of the state. The joint draws (5.17), 50 000 states from $\mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$ pushed through the exact binding loads of all five racks, give the binding load a median of 395.21 kW, the band $[345.89,\ 450.43]$ kW and $\Pr(Q^\dagger \geq 340\,\mathrm{kW}) = 0.9698$, with a Monte Carlo standard error of 0.0008 on the probability. Reweighted to the exact posterior, the same draws give 394.92 kW, $[345.67,\ 450.17]$ kW and 0.9692, gaps of -0.29, -0.22 and -0.26 kW and -0.0006, exact minus reported; the headroom moves from 118.43 kW and

[69.41, 172.92] kW to 118.35 kW and [69.37, 172.73] kW. The two use the same states, so their difference is not sampling noise between independent sets: it is the departure of the posterior of the state from the Laplace normal, and with an effective sample size of 49 638 of 50 000 that departure is small, the probability's gap below its Monte Carlo standard error. The posterior of the state is nearly Gaussian; the gaps between the routes come from the derived quantity.

Two features of the derived quantity make them. The first is the ratio. Rack B3's own binding load, by the draws, has median 396.55 kW and standard deviation 35.97 kW, against 396.46 and 34.96 kW by the delta method: the skew of a ratio widens its spread a little and leaves its median where the inversion route puts it. The second is the minimum over the racks, which cuts the upper tail where another rack binds first. The gaps of the exact posterior from the delta form are −1.54 kW at the median, +6.72 kW at the 5 % point, −3.80 kW at the 95 % point and +0.0223 in the probability of the design load: the left end moves in because of the skew, and the right end moves in because the minimum cuts the tail that the skew would stretch. The gaps of the exact posterior from the inversion route on rack B3 alone are −1.54 kW at the median, −0.41 kW at the 5 % point, −13.28 kW at the 95 % point and −0.0005 in the probability: the inversion route corrects the first feature and not the second, and the probability of the design load, which lives in the left tail, is the one it gets right. Its gap of −0.0005 is smaller than the Monte Carlo standard error of 0.0008, and these draws do not resolve it.

5.4.3 Sensitivity to the airflow survey

The survey is the only knowledge of β_{B3}, so the answer is tested against its width, the fit repeated with the fractions' prior standard deviation at 0.04, 0.08 and 0.12. The binding load at the MAP state barely moves, 396.82, 396.46 and 396.34 kW; its delta-form standard deviation follows the survey's width almost in proportion, 18.93, 34.96 and 51.54 kW; and the delta form's probability of the design load falls from 0.9987 to 0.9468 to 0.8629. Only the delta form is repeated over the widths; at the survey's own width it sits below the joint draws by the gap of Table 5.6.

5.4.4 The answers against the stated truth

The record of hall B is generated from a stated truth, and the last check compares the answers with it. Table 5.7 collects the checks: each row is a route or a variation, the columns the binding load, its standard deviation, the ends of its 90 % band and the probability of the design load. The true binding load is 378.93 kW, at rack B3, inside the reported band. Rack B3's binding load at the MAP state lies 0.50 of the delta form's standard deviation above it: the survey's 0.50 for rack B3 is below the true 0.57, and the fit, which no reading of β_{B3} corrects, carries that error into the answer through the fraction's lever of −423.6 kW. The true binding load clears the 340 kW design load. The design sheet's two answers, 410.00 and 460.55 kW, both lie above the truth and above the reported median.

5.5 The report (step 6)

The design sheet overstates the headroom. Hall B can take a median of 118.4 kW more over today's fitted load, with a 90 % band of 69.4 to 172.9 kW, not the 183 kW that the measured racks suggest, and the unmeasured rack B3 sets it in most of the posterior. The 340 kW design load holds with probability 0.970, against an assurance level of 0.95, which a delta-method band on rack B3 alone would have failed.

Table 5.7: The checks of the binding load of hall B. The draws are 50 000 states from the Laplace posterior through the exact binding loads; the reported route is the joint draws; the survey variations repeat the fit; the probabilities of the variations are the delta form's. The gaps are taken from the unrounded values.

Route or variation	$Q^{\dagger}$ [kW]	sd [kW]	5 % point [kW]	95 % point [kW]	$\Pr(Q^{\dagger} \geq 340\,\text{kW})$ [–]
Joint draws, minimum over racks, reported (median)	395.21	–	345.89	450.43	0.9698
Exact posterior, minimum over racks (median)	394.92	–	345.67	450.17	0.9692
Gap, exact minus reported	−0.29	–	−0.22	−0.26	−0.0006
Inversion route, rack B3 alone (median)	396.46	–	346.08	463.45	0.9696
Delta form, rack B3 alone	396.46	34.96	338.96	453.97	0.9468
Gap, exact minus inversion route	−1.54	–	−0.41	−13.28	−0.0005
Gap, exact minus delta form	−1.54	–	6.72	−3.80	0.0223
Rack B3 alone, Laplace draws (median)	396.55	35.97	–	–	–
Newton covariance	–	34.962	–	–	–
Survey sd 0.04	396.82	18.93	–	–	0.9987
Survey sd 0.12	396.34	51.54	–	–	0.8629
Stated truth	378.93	–	–	–	–
Design sheet, nameplate and survey	410.00	–	–	–	–
Design sheet, measured inlets extrapolated	460.55	–	–	–	–
Fit of the readings	$\chi^2 = 0.085$ on the leverage count 1.468; upper tail 0.894				

5.5.1 The answers and their bands

Table 5.8 gives the answers in the order of Table 5.1. Each row is a question; the second column gives the reported answer with the exact posterior beside it, and the third its band: the central 90 % interval for the binding load and the headroom, for a probability the other routes or its range over the survey widths of Section 5.4.3, and for the budget the standard deviation that would remain.

5.5.2 The design-review entry

The design review's entry reads the binding load, not the sheet. Hall B may be released to its 340 kW design load: the probability that every inlet stays within 27°C there, at the drawing's supply with all units running, is 0.970 by the joint draws over the five racks and 0.969 by the exact posterior, both above the 0.95 the policy asks, as is the inversion route's 0.970 on rack B3 alone, while the delta-method reading of rack B3, 0.947, falls just short; the entry records all four. The headroom over today's fitted load is 118.4 kW at the median, with a 90 % band of 69.4 to 172.9 kW, the load that growth will spend; over the planning figure of 300 kW for today's load, the median headroom is 95.2 kW, band 45.9 to 150.4 kW. The entry also records what the band rests on: one airflow number, the survey's recirculation fraction for rack B3, which carries 94 % of the band's variance and which no reading in the hall checks. The second part of the entry follows from it: before any commitment beyond the design load, the review buys a measurement of rack B3.

Table 5.8: The report for hall B in window 1, at the drawing's supply of 14.0°C with all three units running. Reported values by the Laplace draws of the state, joint over the five racks; the exact posterior beside each.

Question	Answer	Band
How much load can hall B take before an inlet crosses the limit, and how sure is that number?	binding load, median 395.2 kW (exact: 394.9 kW); headroom over today's fitted load, median 118.4 kW (exact: 118.4 kW)	binding load 345.9 to 450.4 kW (exact: 345.7 to 450.2 kW); headroom 69.4 to 172.9 kW (exact: 69.4 to 172.7 kW); central 90 %
Which rack binds first?	rack B3, probability 0.875 (exact: 0.875)	rack B5 0.081, rack B4 0.040, racks B2 and B1 0.002 and 0.001
Can hall B carry its 340 kW design load with 95 % assurance?	probability 0.970 (exact: 0.969), above 0.95	rack B3 alone: inversion route 0.970; delta form 0.947, below 0.95; 0.863 to 0.999 over the survey widths (delta form)
What is the band made of?	rack B3's binding load: 94.0 % the survey's fraction for rack B3; 2.0 % each log-derate	sd 8.59 kW with that fraction known, against 34.96 kW

Common mistake

Reading the headroom off the racks the hall measures. The measured inlets give the right slope for the wrong racks; the rack that binds is read from the fit, with the band its unmeasured airflow puts on it.

5.6 The next measurement (step 7)

The report leaves a band almost all of whose variance is one prior, and a commitment beyond the design load would rest on it. A further measurement is one more row of the residual vector, and what it would remove is known before its value is.

5.6.1 The preposterior of a candidate measurement

A candidate measurement a is a row with Jacobian line $\mathbf{J}_a$, a 1×10 row in the columns of the state, and information $\mathbf{\Lambda}_a = \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a$. Its preposterior covariance is equation (4.67),

$$\mathbf{\Sigma}_a = \left(\mathbf{\Sigma}^{-1} + \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a\right)^{-1}, \qquad \mathrm{sd}_a(Q^{\dagger}) = \sqrt{\mathbf{g}_{B3}^{\mathsf{T}}\mathbf{\Sigma}_a\,\mathbf{g}_{B3}}, \tag{5.26}$$

and the width it removes is $1 - \mathrm{sd}_a(Q^{\dagger})/\mathrm{sd}(Q^{\dagger})$. Because the gradient of the binding load is a constant times that of a_{B3}, the same share comes off the inversion band's spread s_a. Four kinds of candidate are open to the review. An inlet probe at rack k of standard deviation σ_a reads today's inlet, $\mathbf{J}_a = \nabla_{\mathbf{z}} T_k^{\mathsf{T}}/\sigma_a$ at $\hat{\mathbf{z}}$; a permanent inlet sensor at rack k is the same row, added in every window. An airflow re-survey of rack k with standard deviation 0.03 is a new prior row on its fraction, $\mathbf{J}_a = \mathbf{e}_{\beta_k}^{\mathsf{T}}/0.03$. A second return-air sensor repeats the return-air row of $\mathbf{J}$. A better load meter, of standard deviation 1 kW, is a row on Q_B alone, $\mathbf{J}_a = \mathbf{e}_Q^{\mathsf{T}}/1$.

5.6.2 The candidates

Table 5.9 ranks thirteen candidates. Each row is a measurement the review can commission; the columns give its standard deviation, the standard deviation of the binding load after it by equation (5.26), the share of the current 34.96 kW that it removes, and the posterior spread of β_{B3} after it, the diagonal of $\mathbf{\Sigma}_a$. The width removed is drawn in Figure 5.5 for the seven leading candidates, ranked, the horizontal axis in percent.

Table 5.9: Candidate measurements for the binding load of hall B, from the posterior of window 1; the width removed is taken from the delta form's $\mathrm{sd}(Q^\dagger \mid \mathbf{y}) = 34.96\,\mathrm{kW}$.

Candidate	Its sd [K, kW or –]	$\mathrm{sd}_a(Q^\dagger)$ [kW]	Width removed [%]	$\mathrm{sd}_a(\beta_{B3})$ [–]
Inlet probe at rack B3	0.1	8.08	76.9	0.0182
Inlet probe at rack B3; a permanent inlet sensor there	0.3	14.07	59.8	0.0320
Airflow re-survey of rack B3	0.03	14.68	58.0	0.0281
Better load meter	1	34.76	0.6	0.0800
Airflow re-survey of rack B2	0.03	34.91	0.1	0.0800
Airflow re-survey of rack B4	0.03	34.92	0.1	0.0800
Second return-air sensor	0.3	34.92	0.1	0.0800
Inlet probe at rack B1	0.1	34.94	0.1	0.0800
Inlet probe at rack B1	0.3	34.94	0.1	0.0800
Inlet probe at rack B5	0.1	34.94	0.0	0.0800
Inlet probe at rack B5	0.3	34.95	0.0	0.0800
Airflow re-survey of rack B1	0.03	34.96	0.0	0.0800
Airflow re-survey of rack B5	0.03	34.96	0.0	0.0800

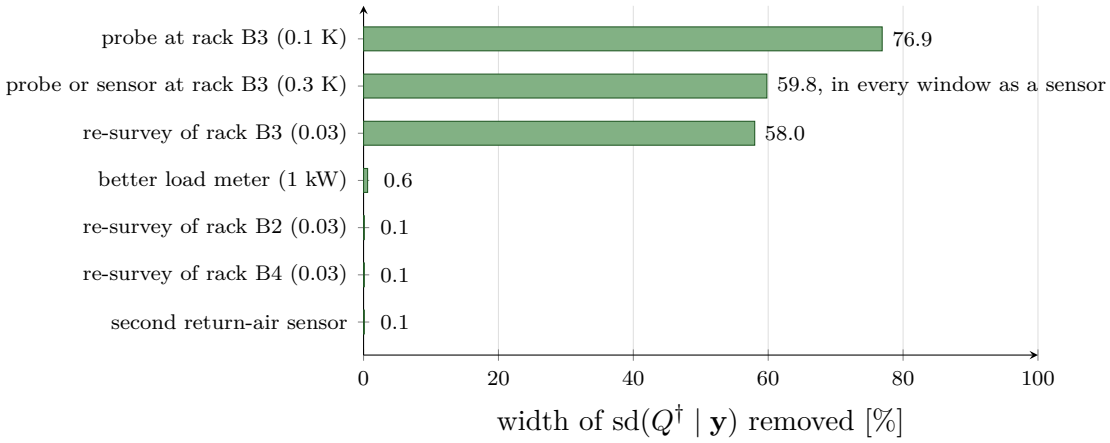


Figure 5.5: The leading candidate measurements of Table 5.9, ranked by the width they would remove from the band of the binding load of hall B, by the preposterior (5.26).

The ranking follows the budget. Every candidate that touches β_{B3}, directly or through rack B3's inlet, removes more than half the width; every candidate that does not removes almost nothing, however good the instrument, and leaves β_{B3} at the survey's 0.0800. Read on the entries, the 0.3 K probe and the re-survey do different things. The re-survey is a row on the fraction alone and narrows it the more, to 0.0281; the probe narrows it to 0.0320, the spread the sensors leave at racks B2 and B4, and yet removes more of the binding load's width, 59.8 against 58.0 %. The probe reads the inlet itself, whose rise over the supply is a_{B3} times the heat in the hall by equation (5.7), so it narrows the combination of the fraction and the log-derates on which the binding load depends, not the fraction alone. The 0.1 K probe, which takes β_{B3} to 0.0182, leaves 8.08 kW, below the 8.59 kW that knowing β_{B3} exactly would leave, for the same reason: the inlet it reads also carries the log-derates, through the cold aisle and the return air, and so it takes a little off their share too. The better load meter narrows an entry the binding load does not depend on: the scenario frees the load, and the meter reaches the binding load only through the load's correlation with the log-derates.

5.6.3 The recommendation

The measurement to buy before committing beyond the design load is an inlet reading at rack B3. A permanent inlet sensor of 0.3 K removes 59.8 % of the band in every window, and reads the rack that binds instead of inferring it; a 0.1 K probe visit removes 76.9 % once. A re-survey of rack B3's airflow removes nearly as much as the sensor, 58.0 %, and is the alternative where a sensor cannot be fitted; a re-survey of the whole hall is worth more than its row shows, since it also covers racks B1 and B5, whose fractions are as unobserved as rack B3's, and can find the stratification that one fraction per rack hides. A better load meter or a second return-air sensor is not worth buying for this question. Chapter 18 ranks the tests and surveys a planner can commission on the site plant in the same way, by the width each takes off the quantity that decides the build.

5.6.4 What the analysis showed

The design sheet put the headroom at 183 kW from the racks it measures, and the analysis puts it at a median of 118 kW, with a 90 % band of 69 to 173 kW, set in most of the posterior by a rack it does not measure. The measured racks are the wrong proxy: the extrapolation from their inlets uses the right physics, a rise in proportion to the heat, but rack B3 draws more return air than either, and the inlet sensors fix their own racks' airflow and nothing about rack B3's. The headroom's band is one airflow number. The readings fix the supply and the coils to a fifth of their prior spread and the load to 3.72 kW, and what is left is the survey's fraction for rack B3, so a sensor anywhere else in the hall cannot narrow it and a reading at rack B3 removes most of it. The hall's binding load is joint over its racks: read on rack B3 alone, even with the ratio kept exact, its band reaches 463.45 kW where the joint band stops at 450.43 kW, because rack B5 binds first in 8 % of the draws. And the delta form misreads a ratio's tail: a rack's binding load is a constant over its rise per kilowatt, its left tail is short, and a symmetric band on it puts the design load below the assurance level when every other route puts it above. The hall's sensors were placed to watch the cooling units, and they watch them well; the headroom is decided at a rack none of them watches.

5.6.5 Limits of the model

A linear scenario at a pinned supply. The scenario holds the supply at 14.0°C at every load and keeps the coils and the fractions at their calibrated values. A chilled-water plant that cannot hold its supply at the higher load would bind before the racks do; the central plant's

capacity enters in Part III. A fraction that changes with the load, as recirculation can when the racks' fans speed up, would bend the inlet lines of Figure 5.2.

One window. The fit reads window 1 alone. Later windows of the night would narrow the load and the log-derates further but leave the unseen fractions where the survey put them, and a window after unit B2's coil derates describes another plant.

One fraction per rack, and the survey's age. On a stratified inlet the one fraction describes an average, and a probe at the top of rack B3 would read warmer than T_{B3}. Racks that share a cause of recirculation would have correlated fractions. The survey was taken at another time, under another load and perhaps another arrangement of the racks, and the model has no row that would detect a survey gone stale.

The pinned airflow and the absent control. The capacity rates C_u are pinned at the units' fixed fan speed; an error in them scales the rise and every fraction's lever, and lies outside every band above. Under an active discharge control the log-derates would be read through the valve positions, rows this model does not have.

The constructed record. The readings of window 1 are generated from a stated truth with the stated sensor errors, not taken from a hall, and the comparison with that truth in Section 5.4.4 is a check of the method on its own model, not of the model against a real hall.

The next chapter. Chapter 6 lets the load grow over the years and asks when this headroom is spent.

5.7 What to retain

Definition 5.1 (Binding load and headroom)**.** The binding load $Q_e^\dagger$ of an element e is the load at which its constrained entry meets its limit at the scenario node, $T_e^*(\mathbf{z}, Q_e^\dagger) = T_{\max}$; the plant's binding load is the smallest over its elements, and the headroom is that load less the load today. Its gradient is the implicit-function form $\mathbf{g}_e = -\nabla_{\mathbf{z}} T_e^* / (\partial T_e^* / \partial Q^*)$ at the binding load, and its variance $\mathbf{g}_e^\mathsf{T} \boldsymbol{\Sigma} \mathbf{g}_e$. Where the entry is linear in the load, $T_e^* = T_s^* + a_e(\mathbf{z})(Q^* + Q_o)$, the binding load is the ratio $(T_{\max} - T_s^*)/a_e - Q_o$, whose band on one element is read by inverting the band of a_e. The plant's binding load, the minimum, is reported jointly, by draws of the posterior of the state pushed through every element's binding load, with the exact posterior beside it and the single-element routes beside both.

Four conclusions carry beyond hall B. A headroom is a scenario solved for the load, and it is known as well as the entries behind its gradient are known; the budget says which they are, and in a hall with unmeasured racks it is often one airflow number. The element that binds is a probability, not a label, and the instrumented elements may be the ones least likely to bind. The plant's binding load is joint over its elements: the binding element read alone overstates the upper tail that another element cuts. And an element's binding load is a ratio, whose band is read by inverting the band of its denominator: a symmetric band on a ratio overstates the low tail, which is the tail a release decision reads.

Table 5.10 lists the terms of the chapter with their expressions, units and the subsection that defines each.

Table 5.10: The terms of thermal headroom and where each is defined. Units for hall B.

Term	Expression	Units	Meaning	Defined in
Binding-load gradient	$\mathbf{g}_e = -\nabla_{\mathbf{z}} T_e^* / (\partial T_e^* / \partial Q^*)$	kW	implicit function theorem	5.1.1
Physical rows	eq. (5.2), ten rows	kW, K	balance, air sides, mixing, inlets	5.1.2
Air balance	$\sum_u C_u(T_r - T_c) = Q_B + Q_o$	kW	the rise set by heat and airflow	5.1.2
Reduced state	$(T_s, Q_B, g_{B1..B3}, \beta_{B1..B5})$	°C, kW, –	the ten free entries	5.1.2
Scenario inlet	$T_k^* = T_s^* + a_k(Q_B^* + Q_o)$	°C	inlet at a pinned supply	5.1.3
Rise per kilowatt	$a_k = 1/\sum_u G_u - (1-\beta_k)/\sum_u C_u$	K/kW	slope of the inlet with the load	5.1.3
Binding load	$Q_k^\dagger = (T_{\max} - T_s^*)/a_k - Q_o$; $Q^\dagger = \min_k Q_k^\dagger$	kW	load at which a limit binds	5.1.3
Headroom	$H = Q^\dagger - Q_B$	kW	load still available	5.1.3
Rows and posterior				
Residual vector	eq. (5.11), seventeen rows	–	readings and priors on the reduced state	5.2.1
Block structure	eq. (5.14)	–	unseen fractions keep their prior	5.2.1
Leverage count	$\sum_k (1 - h_{kk})$ over the reading rows	–	freedom the fit leaves the readings	5.2.2
Reweighted draws	$w_i \propto \mathrm{e}^{-\mathcal{L}(\mathbf{z}_i)} / \mathcal{N}(\mathbf{z}_i; \hat{\mathbf{z}}, \boldsymbol{\Sigma})$	–	the exact posterior	5.2.3
Questions and answers				
Joint draws	$Q_i^\dagger = \min_k Q_k^\dagger(\mathbf{z}_i)$, $\mathbf{z}_i \sim \mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$	kW	the reported binding load and headroom	5.3.1
Inversion band	quantiles of a_k carried through $Q_k^\dagger(a_k)$, reversed	kW	band of a ratio	5.3.1
Density of a ratio	eq. (5.20)	kW^{-1}	change of variables	5.3.1
Binding-rack probability	$\Pr(\arg\min_k Q_k^\dagger = j \mid \mathbf{y})$	–	which element binds first	5.3.2
Assurance test	$\Pr(a_k \le a_d \mid \mathbf{y}) \ge \kappa$	–	design load against the policy	5.3.3
Variance budget	$s_i = g_i(\boldsymbol{\Sigma}\mathbf{g})_i / \mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g}$	–	entries behind the band	5.3.4
Preposterior	$\boldsymbol{\Sigma}_a = (\boldsymbol{\Sigma}^{-1} + \mathbf{J}_a^\mathsf{T}\mathbf{J}_a)^{-1}$	mixed	band after a candidate	5.6.1
Width removed	$1 - \mathrm{sd}_a/\mathrm{sd}$	–	ranks the candidates	5.6.2

5.8 Exercises

Exercise 5.1. *Formulation.* Write the planner's question for rack B5 alone, the load at which rack B5's inlet meets the limit, as a probability over the reduced state, give its gradient, and state its answer from the fit of this chapter. ($\Pr(Q_{B5}^\dagger \le q \mid \mathbf{y})$ with $Q_{B5}^\dagger(\mathbf{z}) = 13/a_{B5}(\mathbf{z}) - 10$ and $a_{B5} = 1/\sum_u G_u - (1-\beta_{B5})/30$; by equation (5.9) its gradient has the entries 291.4, 295.3 and 282.0 kW on the three log-derates and −595.2 kW on β_{B5}; its binding load is 471.78 kW with a delta-form standard deviation of 49.12 kW, and it binds first with probability 0.0813.)

Exercise 5.2. *Derivation.* Derive $Q_k^\dagger = (T_{\max} - T_s^*)/a_k - Q_o$ from the inlet row of equation (5.2), and check rack B3's value. (At the scenario, $T_r^* - T_s^* = (Q^* + Q_o)/\sum_u G_u$ by the hall balance and $T_r^* - T_c^* = (Q^* + Q_o)/\sum_u C_u$ by the air balance (5.6); the inlet row gives $T_k^* = T_r^* - (1-\beta_k)(T_r^* - T_c^*)$, so $T_k^* - T_s^* = a_k(Q^* + Q_o)$; setting $T_k^* = T_{\max}$ and solving for the load gives the binding load. For rack B3, $13/0.031983 - 10 = 396.46$ kW.)

Exercise 5.3. *Derivation.* Show from the gradient and the covariance why the log-derates' share of the budget is 2 % each and the load's zero. (The gradient of $Q_{B3}^\dagger$ has no entry in Q_B, so Q_B's share $g_Q(\boldsymbol{\Sigma}\mathbf{g})_Q$ is zero whatever its correlations. The fraction β_{B3} is uncorrelated

with every other entry, so its share is $(423.6 \times 0.08)^2/34.96^2 = 0.940$, and the remaining 0.060 belongs to the log-derates; their gradient entries, 207.4, 210.2 and 200.7 kW, and their variances and covariances are nearly equal, so they share it nearly equally, 0.020 each.)

Exercise 5.4. *Application.* Invert the 5 % and 95 % points of a_{B3} to the band of rack B3's binding load, and say why the reported band of the hall ends lower. (Given the evidence, a_{B3} is close to $\mathcal{N}(0.031983,\ 0.002751^2)$ in K/kW; its 95 % point is $0.031983 + 1.645 \times 0.002751 = 0.036508$ K/kW and its 5 % point 0.027458 K/kW. The binding load decreases in a_{B3}, so the large rise gives the low end, $13/0.036508 - 10 = 346.08$ kW, and the small rise the high end, $13/0.027458 - 10 = 463.45$ kW. The reported band, joint over the five racks, ends at 450.43 kW: in the draws where rack B3's fraction is small, another rack, chiefly rack B5, binds first and cuts the upper tail.)

Exercise 5.5. *Application.* The facilities team proposes a better hall load meter, of standard deviation 1 kW, to tighten the headroom. Explain from the preposterior why it removes almost nothing. (Its row is $\mathbf{J}_a = \mathbf{e}_Q^{\mathsf{T}}/1$, which narrows Q_B; but the scenario frees the load, so the gradient (5.9) of the binding load has no entry in Q_B, and the meter reaches it only through the correlation of Q_B with the log-derates, 0.21 to 0.22, whose share of the budget is 6 % in all. By equation (5.26) the standard deviation falls from 34.96 to 34.76 kW, a width removed of 0.6 %.)

Chapter 6

Load Growth Scenarios

6.1 When hall B leaves its envelope

The design review that released hall B to its design load in Chapter 5 turns next to time. Hall B carries 300 kW of IT load today, and the planning brief says the load grows 15 % a year. The design sheet answers the question of time the way it answered the question of load: with a date and no band. Worked from the nameplate ratings of the three coils and the commissioning airflow survey, the hall's binding load is 410.00 kW, and the load path reaches it in year 2.24. Worked from the two inlet sensors the hall has, at racks B2 and B4, each reading's rise scaled up in proportion to the load, the binding load is 460.55 kW, reached in year 3.07. The floor believes the second, and says the hall is "inside through year three".

Neither answer says how sure it is, and the review has a date to set: the expansion or the remedy has to be planned before the hall leaves its envelope, not after. The question of this chapter is the drill question in the planner's own words,

If the IT load grows 15 % a year, what is the probability the plant is still inside its envelope in year three?

Here the plant of the question is hall B, its envelope every rack inlet at or below 27°C with all three units running and the chilled-water supply at the drawing's 14°C. It is answered with the method of Chapter 4, on the fit of hall B of Chapter 5. What is new is not the fit but the scenario: the load is pinned not at one value but at a chain of values, one per year, along a path the planner states.

Jargon: load path, growth rate, scenario chain, envelope, crossing year, Monte Carlo standard error, preposterior distribution, decision level

The *load path* Q_t is the IT load the plan expects in year t, and the *growth rate* γ, in %/yr, is its yearly relative increase. A *scenario chain* is a sequence of scenario nodes, one per point of the path, each pinning the load at that point. The *envelope* of hall B is the set of conditions under which every constrained entry is within its limit; the hall is *inside* it at a load when every rack inlet is at or below 27°C there. The *crossing year* $t^{\dagger}$ is the year, counted continuously from today, in which the load path reaches the hall's binding load. The *Monte Carlo standard error* of a probability read as a share of N draws is the spread of that share over repeated sets of draws, $\sqrt{\hat{P}(1-\hat{P})/N}$. The *preposterior distribution* of a posterior quantity is the spread of the values it may take after a measurement not yet made, over the outcomes that measurement may give. A *decision level* is a probability at which the planner acts; here a probability above 0.9 or below 0.1 is taken to decide.

Key idea

A load path asks the headroom question once per year. Each year is a scenario node with the load pinned at that year's value, and the event that the hall is inside its envelope there is the event that the hall's binding load is at least that value. The fit of the plant is not repeated: the growth question costs no new inference, only new scenario nodes, read from the same posterior of the state by the same joint draws. The year the hall leaves its envelope is a monotone function of the binding load, so its band is the binding load's band carried through that function, end to end.

6.1.1 The envelope along a load path

Chapter 5 solved one scenario for the load at which an inlet meets its limit. A plan asks the same question at every year of its horizon, with the load set by the path. The path is a known input: the planner states today's load Q_0 and the growth rate γ, and year t's load is

$$Q_t = Q_0\,(1+\gamma)^t, \qquad t = 0, 1, \ldots, \tag{6.1}$$

in kW, with $Q_0 = 300\,\mathrm{kW}$ and $\gamma = 0.15\,\mathrm{yr}^{-1}$ for hall B. Each year is a scenario node of Section 5.1.3 with the load pinned at $Q_B^* = Q_t$, the supply at T_s^* and all units running.

The hall is inside its envelope at year t when every rack inlet at that node is within the limit. By equation (5.7) every inlet rises with the load at the scenario, so rack k's inlet is within the limit at the load Q_t exactly when that load does not exceed rack k's binding load, and the hall is inside when that holds for every rack,

$$\bigcap_k \{T_k^*(\mathbf{z}, Q_t) \le T_{\max}\} = \bigcap_k \{Q_k^\dagger(\mathbf{z}) \ge Q_t\} = \{Q^\dagger(\mathbf{z}) \ge Q_t\}, \tag{6.2}$$

with $Q_k^\dagger$ and $Q^\dagger = \min_k Q_k^\dagger$ the binding loads of equation (5.8). The event on the right is the event Chapter 5 read for the design load, with the design load replaced by year t's load. Nothing in the state changes with the year: the scenario node moves along the path, and the posterior of the state stays where the fit put it.

The same identity read the other way gives the year itself. The load path rises in t, so it reaches the hall's binding load at one year, the solution of $Q_0(1+\gamma)^t = Q^\dagger$,

$$t^\dagger(\mathbf{z}) = \frac{\ln(Q^\dagger(\mathbf{z})/Q_0)}{\ln(1+\gamma)}, \qquad \{Q^\dagger \ge Q_t\} = \{t^\dagger \ge t\}, \tag{6.3}$$

in years, counted continuously from today. The crossing year is an increasing function of the binding load, so the probability that the hall is inside at year t is the probability that it has not yet crossed, and the quantiles of $t^\dagger$ are those of $Q^\dagger$, each mapped by equation (6.3), in the same order.

> **Principle 6.1.** *On a known load path the hall is inside its envelope at year t exactly when its binding load is at least that year's load. Every year's answer is a tail of the one binding load at a new scenario node, and the year the hall leaves its envelope is that binding load mapped through the path.*

6.1.2 Hall B on its load path (step 0)

The plant is hall B of Chapter 5, unchanged: three cooling units B1–B3 serving five racks B1–B5 from one cold aisle, the ten physical rows of equation (5.2), the reduced state of equation (5.4) with its ten free entries, and the scenario node of Section 5.1.3, at which every inlet is linear in

Table 6.1: The questions asked of hall B on its load path. $\mathbf{y}$ is the eight readings and the nine priors of window 1; $Q^{\dagger}$ is the binding load of equation (5.8); Q_t the path of equation (6.1); $t^{\dagger}$ the crossing year of equation (6.3).

Planner's question	General form	Mathematical form	Category	Computation	Answered in
If the IT load grows 15 % a year, what is the probability the plant is still inside its envelope in year three?	$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y})$ at a scenario node	$\Pr(Q^{\dagger} \geq Q_3 \mid \mathbf{y})$, $Q_3 = 456.26\,\text{kW}$	prediction under a threshold	posterior	6.3.1
And in each year before and after?	the same, per node	$\Pr(Q^{\dagger} \geq Q_t \mid \mathbf{y})$, $t = 0, \ldots, 5$	prediction under a threshold	posterior	6.3.2
In what year does the hall leave its envelope?	$\Pr(f(\mathbf{z}) \leq f^* \mid \mathbf{y})$, quantiles	$t^{\dagger}$: median and 90 % band	estimation	posterior	6.3.3
How much does the answer depend on the growth rate assumed?	the same at stated inputs	$\Pr(Q^{\dagger} \geq Q_3 \mid \mathbf{y})$ at $\gamma = 12$, 15, 18 %/yr	prediction under a threshold	posterior	6.3.4

the load, $T_k^* = T_s^* + a_k(\mathbf{z})\,(Q_B^* + Q_o)$, equation (5.7). The graph of Figure 5.1 is the graph of this chapter. What the chapter adds is a chain of six scenario nodes, one per year from today to year five, each pinning the load at the path's value by equation (6.1): 300.00, 345.00, 396.75, 456.26, 524.70 and 603.41 kW, or 1.0000, 1.1500, 1.3225, 1.5209, 1.7490 and 2.0114 times today's load.

Modeling choices at the scenario chain. Four choices make the chain, each with its cost. The path is a known input: the planner's note states the rate, and the answer is a statement about the hall under that rate, not about the rate; it fails where the rate is itself uncertain, a question taken up with the stated alternatives of Section 6.3.4 and, as an uncertain input, in Chapter 12. The path starts from the planning figure of 300 kW, not from the load meter's reading in window 1: the plan is written in the planner's figure, and the scenario frees the load, so the fit of the hall does not depend on which figure the path starts from. The hall is the same in every year: the coils' log-derates and the racks' fractions keep their calibrated values along the path, and this fails if a coil fouls or the racks are rearranged over five years, changes no row of this model can see. And the growth is in the IT load alone: the other heat $Q_o = 10\,\text{kW}$ stays pinned, and the supply stays at $T_s^* = 14°\text{C}$ at every node, which fails if the central plant cannot hold that supply at the higher loads (Section 6.6.5).

6.1.3 The planner's questions (step 1)

The planner asks four questions of hall B on its load path, and step 1 writes each as a probability or a quantile over the reduced state, with the unknown on the left of the bar and the readings and the priors on the right. Table 6.1 lists them, in the columns of Table 5.1: the question in the planner's words, the drill question first; the general form of its category from Chapter 2; its mathematical form for hall B; its category; the computation of Section 2.3.6 that answers it; and the subsection that answers it. Three are predictions under a threshold: the drill question at year three, the same at every year, and the same at stated growth rates. One is estimation: the year the hall leaves its envelope, with its band. All four are posterior questions, read at scenario nodes from the one fit.

6.1.4 The evidence as rows (step 2)

The evidence is that of Section 5.1.5, row for row: window 1 of the night record at 19:00, eight readings (the supply, the return air, the three discharge temperatures, the hall load meter and the inlets of racks B2 and B4) and nine priors (the chiller's set point on the supply, the commissioning record on the three log-derates, and the airflow survey on the five fractions), seventeen rows on ten entries with the pattern of equation (5.10). The load path adds no row. It is not evidence about the hall: it is a statement of the loads the plan will ask the hall to carry, and it enters only at the scenario nodes, where it pins the load. That is why the growth question costs no new inference.

Modeling choices at the rows. The choices of Section 5.1.5 carry over, and with them their limits: Gaussian, unbiased and independent readings; one fraction per rack, constant over the window, with the survey as the only knowledge of the fractions of racks B1, B3 and B5; one window, the healthy plant. One choice is this chapter's own. The path is not fitted to any record of past load: the planner's rate is taken as stated, so that the answers separate what the hall can carry from what the plan expects it to carry, and the rate's own uncertainty, which a record of past loads would inform, stays outside every band below.

6.2 The posterior of the state (step 3)

All four questions of Table 6.1 are read from one posterior, that of Chapter 5: the reduced state of hall B under the seventeen rows of window 1. No row is added and no fit is repeated; this section cites the posterior and the draws the answers use.

6.2.1 The fit of window 1, cited

The MAP state $\hat{\mathbf{z}}$ and the Laplace covariance $\boldsymbol{\Sigma} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}$ at $\hat{\mathbf{z}}$ are those of Section 5.2.2, equation (5.13), with the block structure of equation (5.14): the supply and the log-derates known to about a fifth of their prior spread, the fractions of the two measured racks to 0.032, and the fractions of racks B1, B3 and B5 left at the survey's values with the survey's spread of 0.08. At $\hat{\mathbf{z}}$ rack B3 binds first, at $Q^{\dagger}_{B3}(\hat{\mathbf{z}}) = 396.46\,\mathrm{kW}$, with the rise per kilowatt $a_{B3} = 0.031983\,\mathrm{K/kW}$ and its spread $s_a = 0.002751\,\mathrm{K/kW}$ by equation (5.18), and the delta-form standard deviation of its binding load is 34.96 kW by equation (5.1), 94 % of whose variance is the survey's fraction for rack B3.

6.2.2 The draws of the state

The answers read the posterior through the same $N = 50\,000$ Laplace draws of Section 5.2.3, seed 5, each pushed through the exact binding loads of all five racks, equation (5.17), and through the same weights, equation (5.16), for the exact posterior; their effective sample size is 49 638. Over those draws the hall's binding load has a median of 395.21 kW and a central 90 % band from 345.89 to 450.43 kW, and the rack that binds first is rack B3 in a share 0.875 of the draws, rack B5 in 0.081 and rack B4 in 0.040. Every answer of step 4 is a function of these 50 000 minima: a share of them above a year's load, or a quantile of them mapped to years.

6.3 The answers (step 4)

Every answer is read from the pair $(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$ and its draws. The reported value is the joint one, the share of the draws of the state whose minimum over the five racks clears the year's load; the exact posterior stands beside it with the gap and the Monte Carlo standard error, and the inversion route and the delta form on rack B3 alone stand beside both.

6.3.1 The envelope in year three

Planner's question. If the IT load grows 15 % a year, what is the probability the plant is still inside its envelope in year three? The question asks for the posterior probability that every rack inlet of hall B is within the limit at the third node of the chain.

General form. A prediction under a threshold is a tail of a derived quantity under the posterior, equation (2.12),

$$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) \geq f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

read here at a scenario node.

Specialized form. The question is equation (2.12) with $f(\mathbf{z}) = Q^{\dagger}(\mathbf{z})$, the binding load of equation (5.8), and $f^* = Q_3$, year three's load on the path (6.1), $300 \times 1.15^3 = 456.26\,\mathrm{kW}$. By equation (6.2) the event is the intersection of the five racks' events, and it is integrated jointly by the draws of the state,

$$P_t = \Pr(Q^{\dagger} \geq Q_t \mid \mathbf{y}) \approx \hat{P}_t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\Big[\min_k Q_k^{\dagger}(\mathbf{z}_i) \geq Q_t\Big], \qquad \mathrm{se}(\hat{P}_t) = \sqrt{\hat{P}_t\,(1-\hat{P}_t)/N}, \tag{6.4}$$

the joint share of equation (5.23) with the design load replaced by the path's load; the exact posterior sums the weights (5.16) of the same draws, $\sum_i w_i\, \mathbf{1}[Q_i^{\dagger} \geq Q_t]$. On rack B3 alone, the inversion route of equation (5.24) and the delta form read

$$P_t^{\mathrm{inv}} = \Phi\!\left(\frac{a_t - a_{B3}(\hat{\mathbf{z}})}{s_a}\right), \quad a_t = \frac{T_{\max} - T_s^*}{Q_t + Q_o}, \qquad P_t^{\mathrm{delta}} = \Phi\left(\frac{Q_{B3}^{\dagger}(\hat{\mathbf{z}}) - Q_t}{\sqrt{\mathbf{g}_{B3}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{g}_{B3}}}\right), \tag{6.5}$$

with a_t the rise per kilowatt at which rack B3 meets the limit at year t's load and Φ the standard normal distribution function.

Building blocks. The draws and the weights of Section 6.2.2; the path's load Q_3; the rise per kilowatt $a_{B3}(\hat{\mathbf{z}})$ and its spread s_a, and the delta-form standard deviation 34.96 kW (Section 6.2.1).

The answer. By equation (6.4) the joint draws give the reported probability,

$$\Pr(Q^{\dagger} \geq 456.26\,\mathrm{kW} \mid \mathbf{y}) = 0.0316,$$

with a Monte Carlo standard error of 0.0008. The exact posterior gives 0.0307, a gap of -0.0009, exact minus reported, taken from the unrounded values and about one standard error. On rack B3 alone the inversion route gives 0.0680 and the delta form 0.0436. The hall almost certainly does not make year three: the floor's "inside through year three" has a posterior probability of three in a hundred.

The joint answer is less than half the inversion route's on rack B3 alone, and the cause is the other racks. Year three's load, 456.26 kW, lies just below the binding loads of racks B4 and B5 at $\hat{\mathbf{z}}$, 461.88 and 471.78 kW (Table 5.4). For the hall to be inside there, rack B3's fraction must sit far enough below the survey's value that its binding load clears 456.26 kW, and in many of the draws in which it does, rack B5, whose fraction keeps the survey's spread, or rack B4 binds below that load instead. The inversion route on rack B3 counts those draws as inside; the joint draws, which take the minimum over the five racks, do not. This is the gap between the binding rack's band and the joint band that Chapter 5 found at the upper end of the binding load: the upper tail of rack B3's binding load is cut by racks B5 and B4, and year three's question lives in that upper tail.

Table 6.2: The probability that hall B is inside its envelope at each node of the load path, $\Pr(Q^{\dagger} \geq Q_t \mid \mathbf{y})$, by the joint draws over the five racks (reported), the exact posterior, and rack B3 alone. Gaps are exact minus reported, from the unrounded values.

Year t [yr]	Q_t [kW]	Joint, reported [–]	MC se [–]	Exact [–]	Gap [–]	Inversion, B3 alone [–]	Delta, B3 alone [–]
0	300.00	0.9999	0.0001	0.9999	0.0000	0.9999	0.9971
1	345.00	0.9536	0.0009	0.9527	−0.0009	0.9540	0.9295
2	396.75	0.4819	0.0022	0.4785	−0.0035	0.4967	0.4967
3	456.26	**0.0316**	0.0008	**0.0307**	−0.0009	0.0680	0.0436
4	524.70	0.0000	0.0000	0.0000	0.0000	0.0026	0.0001
5	603.41	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

6.3.2 The envelope year by year

Planner's question. And in each year before and after? The question asks for the same probability at every node of the chain, from today to year five.

General form. The same tail, equation (2.12), read at a family of thresholds, one per node,

$$P_t = \Pr(f(\mathbf{z}) \geq f_t^* \mid \mathbf{y}), \qquad t = 0, 1, \ldots, 5. \tag{6.6}$$

Specialized form. The family (6.6) with $f(\mathbf{z}) = Q^{\dagger}(\mathbf{z})$ and $f_t^* = Q_t$ of the path (6.1), each member computed by the joint share (6.4), the exact weights, and the two routes (6.5) on rack B3 alone. The same draws serve every node: only the threshold moves. Because the thresholds rise with t, the joint probabilities can only fall along the chain, and by equation (6.3) each is the probability that the crossing year is at least t.

Building blocks. The joint share (6.4) and its standard error; the weights (5.16); the routes (6.5); the six loads of Section 6.1.2.

The years. Table 6.2 sets out the six nodes: each row is a year, the columns its load, the reported joint probability with its Monte Carlo standard error, the exact posterior, its gap from the reported value, and the two routes on rack B3 alone. Figure 6.1 plots the same probabilities against the year: the horizontal axis is the year of the path, the vertical axis the probability of being inside the envelope; the solid black line joins the joint answers at the six nodes, the green line the inversion route on rack B3, the dashed blue line the delta form, and the dotted vertical line marks the year in which the stated truth of the record leaves the envelope, a check marker for Section 6.4.3.

The answers fall from near certainty to near zero over three years. Today and in year one the hall is inside with probability 0.9999 and 0.9536; in year two it is a coin flip, 0.4819; in year three it is 0.0316, and in years four and five the joint share rounds to 0.0000. The routes agree where one rack decides and part where another can. In year one the joint answer and the inversion route agree, 0.9536 against 0.9540, because the question lives in the lower tail of the binding load, where rack B3 binds; the delta form reads 0.9295, lower by a skew that it cannot carry. The binding load is a ratio over rack B3's rise per kilowatt, its lower tail short, and a symmetric normal on it puts too much probability below year one's load. In year two the two routes on rack B3 coincide at 0.4967, since year two's load, 396.75 kW, sits on rack B3's median, where a ratio and its symmetric normal agree; the joint answer is lower, 0.4819, by the draws in which another rack binds below that load. In year three the three part: the joint 0.0316, the delta form 0.0436, the inversion route 0.0680. No assurance level is set for these years, and no route is chosen by a rule; the reported value is the joint one because the envelope

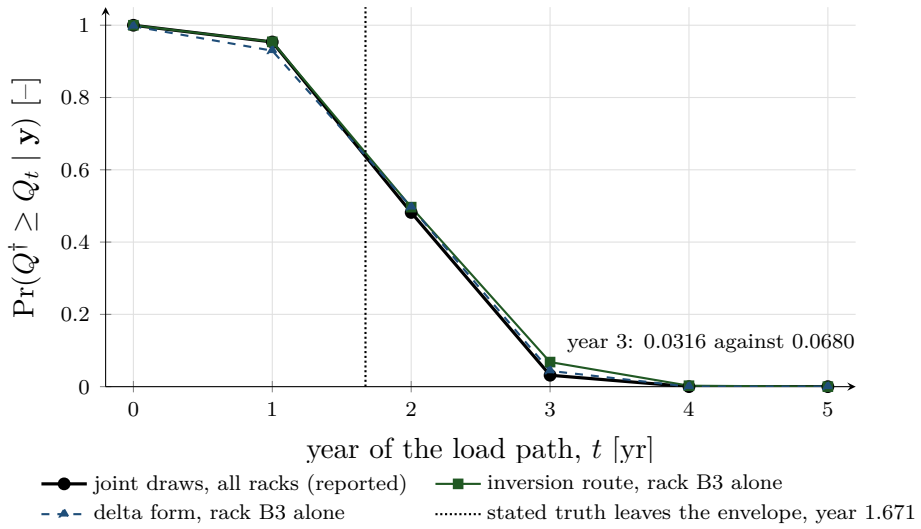


Figure 6.1: The probability that hall B is inside its envelope at each year of the load path $Q_t = 300 \times 1.15^t$ kW, from Table 6.2. The marks are the six scenario nodes; the lines join them for the eye and carry no value between the nodes. Dotted: the year in which the stated truth of the record leaves the envelope, a check marker (Section 6.4.3).

is joint over the racks, and the two routes on rack B3 stand beside it as what the binding rack alone would say.

6.3.3 The year the hall leaves its envelope

Planner's question. In what year does the hall leave its envelope? The question asks for the crossing year with its band.

General form. An estimation question about a derived quantity asks for its posterior distribution, equation (2.9),

$$\Pr(f(\mathbf{z}) \leq f^* \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) \leq f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

summarized by its median and its central 90 % credible interval, equation (2.11).

Specialized form. The question is equation (2.9) with $f(\mathbf{z}) = t^\dagger(\mathbf{z})$ of equation (6.3). The crossing year is an increasing function of the binding load, so the event $\{t^\dagger \leq \tau\}$ is the event $\{Q^\dagger \leq Q_0(1+\gamma)^\tau\}$, and every quantile of $t^\dagger$ is the quantile of $Q^\dagger$ at the same level, mapped,

$$t^\dagger_p = \frac{\ln(Q^\dagger_p / Q_0)}{\ln(1+\gamma)}, \qquad p = 0.05,\ 0.5,\ 0.95, \tag{6.7}$$

with $Q^\dagger_p$ the p-quantile of the joint draws (5.17). The map is exact on the draws: each draw's crossing year is its binding load mapped, and the quantiles of the mapped draws are the mapped quantiles. Beside it stand the two routes on rack B3 alone. The inversion route maps the same way: rack B3's binding load, $(T_{\max} - T^*_s)/a_{B3} - Q_o$, falls as its rise per kilowatt rises, so its p-quantile is the binding load at the $(1-p)$-quantile of $a_{B3} \mid \mathbf{y} \approx \mathcal{N}(a_{B3}(\hat{\mathbf{z}}), s_a^2)$, carried to years by equation (6.7). For the delta form, writing $Q^\dagger$ for rack B3's binding load, the chain rule on equation (6.3) gives the gradient of the crossing year as that of the binding load divided

Table 6.3: The year hall B leaves its envelope on the 15 %/yr load path, by four routes. The joint draws and the exact posterior take the minimum over the five racks; the inversion route and the delta form read rack B3 alone.

Route	Median [yr]	5 % point [yr]	95 % point [yr]	$\Pr(t^\dagger \le 3)$ [–]
Joint draws, all racks (reported)	1.972	1.019	2.908	0.9684
Exact posterior, all racks	1.967	1.014	2.904	0.9693
Inversion route, rack B3 alone	1.995	1.022	3.112	0.9320
Delta form, rack B3 alone	1.995	0.957	3.033	0.9444

by the binding load times $\ln(1+\gamma)$,

$$\nabla_{\mathbf{z}} t^\dagger = \left. \frac{\nabla_{\mathbf{z}} Q^\dagger_{B3}}{Q^\dagger_{B3}\,\ln(1+\gamma)} \right|_{\hat{\mathbf{z}}}, \qquad \mathrm{sd}(t^\dagger \mid \mathbf{y}) \approx \frac{\sqrt{\mathbf{g}_{B3}^{\mathsf{T}} \boldsymbol{\Sigma} \mathbf{g}_{B3}}}{Q^\dagger_{B3}(\hat{\mathbf{z}})\,\ln(1+\gamma)}, \tag{6.8}$$

in years, with $\mathbf{g}_{B3}$ the gradient of equation (5.9). The logarithm of a ratio is a difference of logarithms, so the crossing year is less skewed than the binding load, and its symmetric band misplaces less.

Building blocks. The joint draws and their weights (Section 6.2.2); the map (6.7); the rise per kilowatt $a_{B3}(\hat{\mathbf{z}})$ and its spread s_a; the delta-form standard deviation 34.96 kW and rack B3's binding load 396.46 kW at $\hat{\mathbf{z}}$; the growth rate γ.

The crossing year. The joint draws give the reported answer,

$$\text{median } t^\dagger = 1.972\,\text{yr}, \qquad t^\dagger \in [1.019,\ 2.908]\,\text{yr with 90 \% probability}, \tag{6.9}$$

or 23.7 months, with a band from month 12.2 to month 34.9. The exact posterior gives a median of 1.967 yr and a band from 1.014 to 2.904 yr, gaps of -0.0053, -0.0046 and -0.0042 yr, exact minus reported, at the median and the two ends. The probability that the hall has left its envelope by year three is 0.9684 by the joint draws and 0.9693 by the exact posterior; by year two 0.5181; by year one 0.0464. On rack B3 alone the inversion route carries the rise per kilowatt's quantiles through the binding load to years, a median of 1.995 yr and a band from 1.022 to 3.112 yr. By equation (6.8) the delta form on rack B3 puts the crossing year at 1.995 yr with standard deviation 0.631 yr, a band of $1.995 \pm 1.645 \times 0.631$, from 0.957 to 3.033 yr.

The crossing is drawn on the path itself in Figure 6.2: the horizontal axis is the year, the vertical axis the hall's IT load; the black curve is the load path with its six scenario nodes, the blue band the joint 90 % band of the binding load with its median as a dashed line, and the dotted line the stated truth's binding load. Where the path meets the median, the dashed vertical line marks the median crossing year, and the bar along the foot of the plot is the band of crossing years, the band of loads read across to the path. The two squares are the design sheet's dates.

The plot reads the band of years off the band of loads: the path is steep enough that a band of about 105 kW in load becomes a band of under two years, and the sheet's two dates sit at the path's crossing of loads in the band's upper half and above it.

Table 6.3 sets the routes side by side: each row is a route, the columns the median and the ends of the 90 % band of the crossing year and the probability of leaving the envelope by year three.

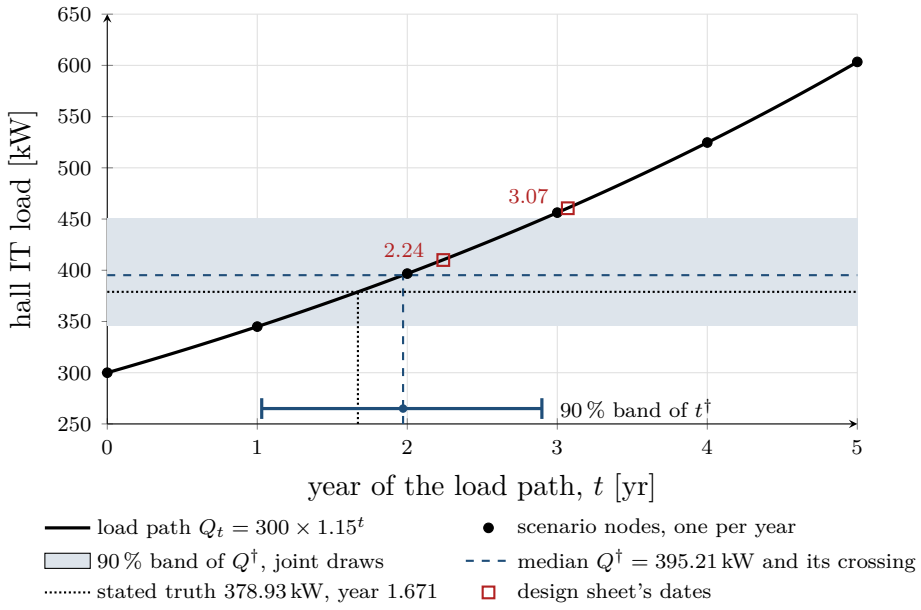


Figure 6.2: The load path of hall B against its binding load. The path $Q_t = 300 \times 1.15^t$ kW of equation (6.1), with its scenario nodes; the band is the joint 90 % band of the binding load, [345.89, 450.43] kW, with its median, 395.21 kW, dashed; the path crosses the band between years 1.019 and 2.908, the bar at the foot, and the median at year 1.972. Dotted: the stated truth of the record, a check marker (Section 6.4.3). Squares: the design sheet's binding loads, 410.00 and 460.55 kW, at the years the path reaches them.

The inversion route's probability of leaving by year three in Table 6.3 is one less its probability of being inside at year three in Table 6.2, $1 - 0.0680$, because the map from loads to years is monotone and the route maps exactly. The delta form's is not: its 0.9444 comes from a normal on the crossing year, while one less the delta form in Table 6.2 would give $1 - 0.0436 = 0.9564$, from a normal on the binding load. The two are different linearizations of one ratio, taken before and after the logarithm, and the one on the logarithm sits closer to the inversion route. The density of the crossing year is drawn in Figure 6.3: the horizontal axis is the crossing year, the vertical axis its posterior density; the solid curve is the inversion route on rack B3 alone, the density of the binding load (5.20) carried to years by the change of variables $q = Q_0(1+\gamma)^t$, with $\mathrm{d}q/\mathrm{d}t = q\ln(1+\gamma)$; the dashed curve is the delta form's normal; the shaded area is the inversion route's probability of being inside at year three; the dotted line marks the year of the stated truth; and the three bars above the curves are the 90 % bands of the joint draws, the inversion route and the delta form, with their medians.

The two curves on rack B3 nearly coincide in their body: the logarithm in the map undoes most of the ratio's skew. In the tails the delta form still sits to the left of the inversion route, its band [0.957, 3.033] yr against [1.022, 3.112] yr, and its probability of having left by year three 0.9444 against 0.9320. The difference that matters more is the one between rack B3 alone and the hall. The reported band ends at 2.908 yr, before year three, where both bands on rack B3 run past it, and the shaded tail beyond year three, 0.0680 of the inversion route's probability, shrinks to the joint 0.0316 because in those draws rack B5 or B4 sends the hall out of its envelope first. The lower end, 1.019 yr or month 12.2, is where rack B3 binds, and all routes place it near year one.

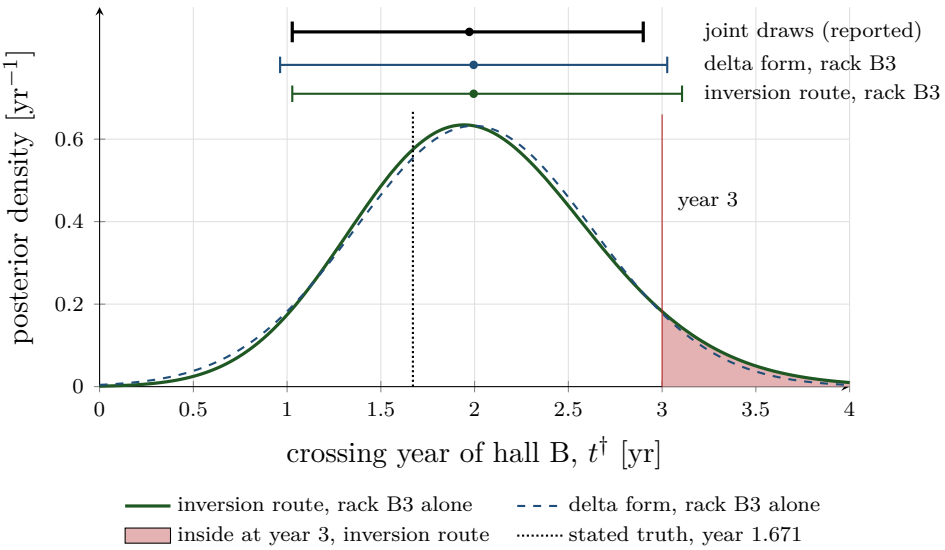


Figure 6.3: The posterior of the year hall B leaves its envelope on the 15 %/yr path. Solid: the inversion route for rack B3 alone, the density of $t^{\dagger} = \ln(Q^{\dagger}_{B3}/300)/\ln 1.15$ with $a_{B3} \mid \mathbf{y} \approx \mathcal{N}(0.031983,\ 0.002751^2)$ K/kW, $T_{\max} - T_s^* = 13$ K and $Q_o = 10$ kW. Dashed: the delta form (6.8) for rack B3, $\mathcal{N}(1.995,\ 0.631^2)$ yr. Red area: the inversion route's probability of being inside at year three, 0.0680. Dotted: the stated truth of the record, a check marker (Section 6.4.3). Bars: the 90 % bands with their medians; the top bar is the reported band, [1.019, 2.908] yr, from the joint draws over the five racks, then the delta form's [0.957, 3.033] yr and the inversion route's [1.022, 3.112] yr; the exact posterior's band lies within 0.0053 yr of the reported one and is not drawn.

6.3.4 The growth rate as a stated alternative

Planner's question. How much does the answer depend on the growth rate assumed? The question asks for the year-three answer at other stated rates, the path still taken as known at each.

General form. The tail of equation (2.12) read at a family of stated inputs, one scenario chain per rate,

$$P_3(\gamma) = \Pr(f(\mathbf{z}) \geq f^*(\gamma) \mid \mathbf{y}), \qquad \gamma \in \{0.12,\ 0.15,\ 0.18\}\,\text{yr}^{-1}. \tag{6.10}$$

Specialized form. The family (6.10) with $f(\mathbf{z}) = Q^{\dagger}(\mathbf{z})$ and $f^*(\gamma) = Q_0(1+\gamma)^3$ by equation (6.1), each member computed by the joint share (6.4); the median crossing year at each rate by the map (6.7) with that rate. The rate is a stated alternative, not an uncertain entry: the state and its draws are the same at every rate, and only the threshold and the map change.

Building blocks. The joint draws (Section 6.2.2); the path (6.1) and the map (6.7) at each rate.

The rates. Table 6.4 gives the three rates: each row is a rate, the columns year three's load on its path, the joint probability of being inside then, and the median crossing year.

At 12 %/yr the hall is inside in year three with probability 0.2213; at 18 %/yr, 0.0003. Even the slowest of the three rates leaves the hall more likely out than in by year three, and the median crossing year moves only from 2.432 to 1.665 yr across the three rates. The answer is

Table 6.4: The year-three answer for hall B at three stated growth rates, by the joint draws over the five racks; the path is known at each rate.

Growth rate γ [%/yr]	Q_3 [kW]	$\Pr(Q^\dagger \geq Q_3 \mid \mathbf{y})$ [–]	Median $t^\dagger$ [yr]
12	421.48	0.2213	2.432
15	456.26	0.0316	1.972
18	492.91	0.0003	1.665

sensitive to the rate in its probability, because year three's load moves across the upper tail of the binding load, and much less in its crossing year, because the crossing year depends on the rate only through the logarithm $\ln(1+\gamma)$. A growth rate that is itself uncertain is an input with its own distribution, and Chapter 12 takes it up; here, as the planning brief states, the path is known.

6.4 The checks (step 5)

The answers of Section 6.3 rest on the joint draws, on the map from loads to years, and on the fit of Chapter 5; each is tested here against what it stands for, and the answers against the truth the record was generated from.

6.4.1 The joint draws against the exact posterior

The reported probabilities are shares of 50 000 Laplace draws, and the exact posterior is the same draws reweighted. The gaps, exact minus reported, are -0.0009 at year three, against a Monte Carlo standard error of 0.0008, and -0.0035 at year two, against 0.0022, so that the year-two gap is $0.0035/0.0022 = 1.6$ standard errors. The two estimates use the same states, so their difference is not sampling noise between independent sets of draws: it is the departure of the posterior of the state from the Laplace normal, which an effective sample size of 49 638 of 50 000 shows to be small. The standard error measures how far either share may sit from its own limit with more draws, and both gaps are of its order: neither moves a year's answer across 0.5, 0.9 or 0.1.

6.4.2 The map from loads to years

The crossing year was read as the binding load's quantiles mapped by equation (6.7). Mapping the three quantiles of the joint binding load directly gives $\ln(395.21/300)/\ln 1.15 = 1.972$ yr, $\ln(345.89/300)/\ln 1.15 = 1.019$ yr and $\ln(450.43/300)/\ln 1.15 = 2.908$ yr, the median and the band of equation (6.9) to the last digit. The year table and the crossing year agree too: the probability of being inside at year t is one less the probability of having crossed by t, $1 - 0.0464 = 0.9536$ at year one, $1 - 0.5181 = 0.4819$ at year two and $1 - 0.9684 = 0.0316$ at year three, the joint entries of Table 6.2.

6.4.3 The answers against the stated truth

The record of hall B is generated from a stated truth, and the last check compares the answers with it. Table 6.5 collects the checks: each row is a route, a check or the truth, and the columns give the binding load, the crossing year with its band, and the probabilities of being inside at years two and three. The true binding load is 378.93 kW, so the true hall leaves its envelope in year 1.671, month 20.1: it is inside at year one and outside at years two and three. The true crossing year lies inside the reported band, below its median of 1.972 yr, for the reason

Table 6.5: The checks of the crossing year of hall B on the 15 %/yr path. The reported route is the joint draws; the gaps are from the unrounded values.

Route or check	$Q^{\dagger}$ [kW]	$t^{\dagger}$ [yr]	5 % point [yr]	95 % point [yr]	Inside, year 2 [–]	Inside, year 3 [–]
Joint draws, reported (median)	395.21	1.972	1.019	2.908	0.4819	0.0316
Exact posterior (median)	–	1.967	1.014	2.904	0.4785	0.0307
Gap, exact minus reported	–	−0.0053	−0.0046	−0.0042	−0.0035	−0.0009
Monte Carlo standard error	–	–	–	–	0.0022	0.0008
Inversion route, rack B3 alone	396.46	1.995	1.022	3.112	0.4967	0.0680
Delta form, rack B3 alone	396.46	1.995	0.957	3.033	0.4967	0.0436
Quantiles of $Q^{\dagger}$ mapped, eq. (6.7)	395.21	1.972	1.019	2.908	–	–
Stated truth	378.93	1.671	–	–	outside	outside
Design sheet, nameplate and survey	410.00	2.24	–	–	–	–
Design sheet, measured inlets extrapolated	460.55	3.07	–	–	–	–

Section 5.4.4 gave: the survey's fraction for rack B3 is below the true one, and no reading corrects it. The design sheet's two dates, years 2.24 and 3.07, both lie after the true crossing, and the second, the one the floor believes, after the reported band's upper end.

6.5 The report (step 6)

With the IT load growing 15 % a year, hall B is inside its envelope in year three with probability 0.03, not "through year three" as the floor reads its sensors. It leaves its envelope at a median of year 1.97, with a 90 % band from year 1.02 to year 2.91, and in year two whether it is still inside is a coin flip, 0.48.

6.5.1 The answers and their bands

Table 6.6 gives the answers in the order of Table 6.1. Each row is a question; the second column gives the reported answer with the exact posterior beside it, and the third its band: the routes on rack B3 alone for a probability, the central 90 % interval for the crossing year, and the range over the stated rates for the last question.

6.5.2 The design-review entry

The design review's entry reads the crossing year, not the sheet's date. On the planned 15 %/yr path hall B leaves its envelope, at the drawing's supply with all units running, at a median of month 23.7, with a 90 % band from month 12.2 to month 34.9; the probability that it is still inside in year three is 0.032 by the joint draws and 0.031 by the exact posterior, and the rack-B3-alone routes, 0.068 and 0.044, are recorded beside them. The entry plans the expansion or the remedy to be in service before month 12, the band's lower end, by which the probability of having left the envelope is still 0.046; a plan keyed to the sheet's year three would be late with probability 0.968. The entry also records what the date rests on: the same airflow number as the headroom, the survey's fraction for rack B3, whose measurement is taken up in step 7. Chapter 11 turns the in-service month into an order date, with the lead time of the equipment the remedy needs.

Table 6.6: The report for hall B on the 15 %/yr load path, at the drawing's supply of 14.0°C with all three units running. Reported values by the Laplace draws of the state, joint over the five racks; the exact posterior beside each.

Question	Answer	Band
If the IT load grows 15 % a year, what is the probability the plant is still inside its envelope in year three?	probability 0.032 (exact: 0.031)	rack B3 alone: inversion route 0.068, delta form 0.044; Monte Carlo standard error 0.0008
And in each year before and after?	years 0 to 5: 1.000, 0.954, 0.482, 0.032, 0.000, 0.000 (exact: 1.000, 0.953, 0.479, 0.031, 0.000, 0.000)	year two on rack B3 alone 0.497 by both routes; year one by the delta form 0.930
In what year does the hall leave its envelope?	median year 1.97, month 23.7 (exact: 1.97)	1.02 to 2.91 yr, months 12.2 to 34.9 (exact: 1.01 to 2.90 yr); rack B3 alone: inversion route 1.02 to 3.11 yr, delta form 0.96 to 3.03 yr
How much does the answer depend on the growth rate assumed?	year three at 12 %/yr: 0.221; at 18 %/yr: 0.0003	median crossing year 2.43 to 1.67 yr over 12 to 18 %/yr

Common mistake

Extrapolating the measured inlets along the load path to a date. The measured racks give the right slope for the wrong racks, and a date without a band hides that the hall is a coin flip a year before the date the sheet gives.

6.6 The next measurement (step 7)

The report leaves a decision for year two that is a coin flip, and a date whose band is one airflow number. A measurement is one more row, and what it would do to the decision is known before its value is.

6.6.1 The preposterior of the year-two decision

Chapter 5 ranked the candidate measurements by the width each would remove from the band of the binding load, and found an inlet probe at rack B3 of standard deviation 0.1 K at the top of the ranking. The preposterior covariance $\mathbf{\Sigma}_a$ of equation (5.26) gives the spread the binding load would keep after the probe, and the same covariance carries the crossing year by equation (6.8). The year-two decision asks more: not how wide the band would be, but where the probability of being inside at year two would land once the probe is read.

The question is answered on rack B3's delta form, the linear route on which the probe's effect has a closed form. On that route the binding load is normal, and the probe moves its posterior mean by an amount not known until the probe is read. Before the reading, that mean is itself normal around today's value: the variance the probe removes from the binding load is the variance of the mean it leaves, by the law of total variance,

$$\hat{Q}_a^{\dagger} \sim \mathcal{N}(Q_{B3}^{\dagger}(\hat{\mathbf{z}}),\; s_{\Delta}^2), \qquad s_{\Delta}^2 = \mathbf{g}_{B3}^{\mathsf{T}} \mathbf{\Sigma} \mathbf{g}_{B3} - \mathbf{g}_{B3}^{\mathsf{T}} \mathbf{\Sigma}_a \mathbf{g}_{B3}, \tag{6.11}$$

with $\hat{Q}_a^{\dagger}$ the posterior mean after the probe and s_{Δ} the spread of the shift the probe can cause,

Table 6.7: The inlet probe at rack B3, of standard deviation 0.1 K, before the year-two decision; delta form on rack B3, by the preposterior covariance and draws of the shift of the posterior mean.

Quantity [unit]	Today [row's unit]	After the probe [row's unit]
sd of the binding load $Q^\dagger$ [kW]	34.96	8.08
sd of the crossing year $t^\dagger$ [yr]	0.631	0.146
Spread of the shift of the posterior mean, s_Δ [kW]	–	34.02
Pr(inside at year 2), its 5 %, 50 %, 95 % points [–]	0.4967	0.0000, 0.4796, 1.0000
Probability that $P_2^a > 0.9$ [–]	–	0.3764
Probability that $P_2^a < 0.1$ [–]	–	0.3843

in kW. Each outcome gives a posterior probability of being inside at year two,

$$P_2^a = \Phi\left(\frac{\hat{Q}_a^\dagger - Q_2}{\sqrt{\mathbf{g}_{B3}^\mathsf{T}\boldsymbol{\Sigma}_a\mathbf{g}_{B3}}}\right), \qquad \Pr(P_2^a > 0.9), \quad \Pr(P_2^a < 0.1), \tag{6.12}$$

and the preposterior distribution of P_2^a is the distribution of that function of the shift. The two probabilities on the right are the chances that the probe decides the question, inside or out, at the decision levels 0.9 and 0.1. They are read by draws of the shift from the normal (6.11), each pushed through equation (6.12).

6.6.2 The probe at rack B3

Table 6.7 gives the probe's effect: each row is a quantity, the columns its value today and after the probe, from the preposterior covariance (5.26) and the draws (6.12).

The probe takes the standard deviation of the binding load from 34.96 to 8.08 kW and that of the crossing year from 0.631 to 0.146 yr, a standard deviation of under two months. Read as a decision, it does more than narrow. Today the probability of being inside at year two is 0.4967 on this route; after the probe it is spread almost from 0 to 1, its 5 % and 95 % points at 0.0000 and 1.0000, because a shift of spread 34.02 kW moves the mean across year two's load in most outcomes while the remaining spread of 8.08 kW is too narrow to keep the probability near one half. With probability 0.3764 the probe leaves the probability of being inside above 0.9, and with probability 0.3843 below 0.1: in $0.3764 + 0.3843 = 0.7607$ of its outcomes, about three times out of four, the probe turns the coin flip into a decision. The numbers are the delta form's preposterior on rack B3: today's value on that route, 0.4967, sits above the joint 0.4819, and the joint preposterior, which would let rack B5 or B4 bind in some draws, is not computed here.

6.6.3 The recommendation

The measurement to make before the year-two decision is the inlet probe at rack B3. It costs one visit, and in about three outcomes in four it settles whether the hall makes year two, which no other reading in the hall can do, since every other candidate of Table 5.9 leaves the band of the binding load nearly as it is. A permanent inlet sensor at rack B3 would carry the same information into every later window, and would watch the rack that sets the date as the load grows. Until the probe is read, the planning date of the entry stands: the band's lower end is set by where rack B3 binds, and only a reading of rack B3 can move it. Chapter 18 ranks the

tests and surveys a planner can commission by the decision each would settle, of which this is the first instance.

6.6.4 What the analysis showed

The design sheet put hall B's date at year 3.07 from the racks it measures, and the analysis puts the probability of being inside in year three at 0.032, with the crossing year at a median of 1.97 and a band from 1.02 to 2.91. The growth question cost no new inference: the path is a known input, the hall is inside at a year exactly when its binding load clears that year's load, and every year's answer is a tail of the one posterior of Chapter 5, read at a new scenario node. The joint reading mattered more here than in the headroom question. Year three's load falls in the upper tail of the binding load, where racks B5 and B4 cut what rack B3 alone would leave, and the joint answer is less than half the inversion route's on rack B3. The delta form departed in year one, where the ratio's short lower tail decides, and agreed in year two, where the load sits on the median. The crossing year carried the binding load's band through a logarithm that undoes most of its skew, so that the date is less sensitive to the route than the probabilities are. And the year-two decision, a coin flip today, is the decision one probe at the unmeasured rack would settle three times in four.

6.6.5 Limits of the model

A known path. The growth rate is a stated input, and the bands carry the hall's uncertainty, not the plan's. The stated alternatives of Section 6.3.4 show how far the answer moves with the rate; an uncertain rate, with a distribution of its own, is Chapter 12's question. The path is smooth, while real load arrives in steps as rows of racks are installed; a step past the binding load moves the date to the step.

The supply pinned, the plant fixed. Every node holds the supply at the drawing's 14.0°C and runs all three units. A chilled-water plant that cannot hold that supply at 450 kW in the hall would bind before the racks do; the central plant's capacity enters in Part III, with the site. The coils and the fractions keep their calibrated values over five years, and a coil that fouls or a row of racks rearranged would move every binding load.

One window, one fit. The posterior is that of window 1, the healthy plant, with the limits of Section 5.6.5: the linear scenario, one fraction per rack, the survey's age, the pinned airflow.

The preposterior on one rack. The probe's effect on the year-two decision is computed on rack B3's delta form, not jointly over the racks; the joint preposterior would count the draws in which another rack binds first.

The constructed record. The readings of window 1 are generated from a stated truth with the stated sensor errors, not taken from a hall, and the comparison with that truth in Section 6.4.3 is a check of the method on its own model, not of the model against a real hall.

The next chapter. Chapter 7 holds the load and moves the weather: it asks whether the hall stays inside its envelope on the design day.

6.7 What to retain

Definition 6.1 (Envelope on a load path and crossing year)**.** On a known load path $Q_t = Q_0(1+\gamma)^t$ the plant is inside its envelope at year t exactly when its binding load is at least

that year's load, $\{Q^\dagger \geq Q_t\}$, so that each year's probability is a tail of the one binding load at a scenario node, reported jointly by draws of the posterior of the state with the exact posterior beside it. The crossing year $t^\dagger = \ln(Q^\dagger/Q_0)/\ln(1+\gamma)$ is an increasing function of the binding load: its quantiles are the binding load's, mapped, and the probability of being inside at year t is the probability that $t^\dagger \geq t$.

Four conclusions carry beyond hall B. A load-growth question is a chain of scenario nodes on the posterior of the plant already fitted: it costs no new inference, and its uncertainty is the plant's. Where the year's load falls in the upper tail of the binding load, the other elements cut that tail, and the binding element read alone overstates the chance of staying inside. A date read from a monotone map of a binding load takes its band from the binding load's band, end to end, and a logarithm in the map undoes most of a ratio's skew. And a measurement can be judged by the decision it would settle, the preposterior probability that it moves a coin flip past a decision level, not only by the width it removes.

Table 6.8 lists the terms of the chapter with their expressions, units and the subsection that defines each.

Table 6.8: The terms of load growth scenarios and where each is defined. Units for hall B.

Term	Expression	Units	Meaning	Defined in
Load path	$Q_t = Q_0(1+\gamma)^t$	kW	known input, one node per year	6.1.1
Inside at year t	$\{Q^\dagger \geq Q_t\}$, eq. (6.2)	–	every inlet within the limit	6.1.1
Crossing year	$t^\dagger = \ln(Q^\dagger/Q_0)/\ln(1+\gamma)$	yr	year the path meets the binding load	6.1.1
Scenario chain	$Q^*_B = Q_t$, $t = 0, \ldots, 5$	kW	the nodes of the path	6.1.2
Questions and answers				
Joint tail at a node	$\hat{P}_t = N^{-1}\sum_i \mathbf{1}[Q^\dagger_i \geq Q_t]$	–	reported probability	6.3.1
Monte Carlo standard error	$\sqrt{\hat{P}_t(1-\hat{P}_t)/N}$	–	spread of a share of draws	6.3.1
Routes at a node	eq. (6.5)	–	rack B3 alone, inversion and delta	6.3.1
Quantile map	$t^\dagger_p = \ln(Q^\dagger_p/Q_0)/\ln(1+\gamma)$	yr	band of the crossing year	6.3.3
Delta form of the year	$\mathrm{sd}(Q^\dagger)/(Q^\dagger \ln(1+\gamma))$	yr	beside-route on one rack	6.3.3
Stated rates	$P_3(\gamma)$ at $\gamma = 0.12, 0.15, 0.18$	–	sensitivity to the plan	6.3.4
Shift of the mean	$s^2_\Delta = \mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g} - \mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}_a\mathbf{g}$	kW	what the probe can move	6.6.1
Decision resolved	$\Pr(P^a_2 > 0.9)$, $\Pr(P^a_2 < 0.1)$	–	preposterior of a probability	6.6.1

6.8 Exercises

Exercise 6.1. *Formulation.* Write the planner's question for year four as a probability over the reduced state, give the event in terms of the racks' binding loads, and state its answer by the joint draws and by rack B3's inversion route. ($\Pr(Q^\dagger \geq Q_4 \mid \mathbf{y})$ with $Q_4 = 300 \times 1.15^4 = 524.70\,\mathrm{kW}$; by equation (6.2) the event is $\bigcap_k \{Q^\dagger_k(\mathbf{z}) \geq 524.70\,\mathrm{kW}\}$; the joint share rounds to 0.0000, while rack B3 alone by the inversion route gives 0.0026.)

Exercise 6.2. *Derivation.* Derive the delta-form standard deviation of the crossing year from that of rack B3's binding load, and check its value. (By the chain rule on equation (6.3), $\nabla_\mathbf{z} t^\dagger = \nabla_\mathbf{z} Q^\dagger_{B3}/(Q^\dagger_{B3} \ln 1.15)$, so the standard deviation is $34.96/(396.46 \times \ln 1.15)$, with $\ln 1.15 = 0.1398$; the delta form gives 0.631 yr.)

Exercise 6.3. *Derivation.* Show that the probability of being inside at year t is one less the probability that the crossing year is at most t, and check it at year two. (The path rises in t, so $Q^{\dagger} \geq Q_t$ exactly when $t^{\dagger} \geq t$, equation (6.3); on the draws, $\Pr(t^{\dagger} \leq 2) = 0.5181$, and $1 - 0.5181$ is the joint year-two probability, 0.4819.)

Exercise 6.4. *Application.* The planner revises the growth rate to 12 %/yr. Without refitting, give year three's load, the probability that hall B is inside then, and the median crossing year. ($Q_3 = 300 \times 1.12^3 = 421.48\,\mathrm{kW}$; the same draws give $\Pr(Q^{\dagger} \geq 421.48\,\mathrm{kW} \mid \mathbf{y}) = 0.2213$, and the median binding load mapped by equation (6.7) with $\ln 1.12$ gives a median crossing year of 2.432 yr.)

Exercise 6.5. *Application.* The review asks whether the rack-B3 probe is worth a visit before the year-two decision. Give the spread of the shift it can cause and the probability that it settles the decision. (By equation (6.11), $s_{\Delta}^2 = 34.96^2 - 8.08^2$, so the shift has a spread of 34.02 kW; by the draws (6.12) the probe settles the decision with probability $0.3764 + 0.3843 = 0.7607$, leaving the probability of being inside at year two above 0.9 with probability 0.3764 and below 0.1 with probability 0.3843.)

Chapter 7

Design-Day Assessment

7.1 Hall B on the design day

The design sheet of hall B carries a design-day check. It sets the hall at its 340 kW design load; it raises the other heat in the hall from the 10 kW of the fans and lighting to 25 kW, adding the 15 kW that the design day's envelope and fresh air bring in, a planning input stated on the sheet; it runs every cooling unit; and it holds the chilled-water supply at the drawing's 14°C. Worked from the nameplate ratings of the three coils and the commissioning airflow survey, the five rack inlets come out at 21.648, 22.864, 25.298, 24.081 and 23.473°C for racks B1 to B5. The hottest is rack B3 at 25.30°C, 1.70 K inside the 27°C limit, and the sheet writes "pass".

The sheet's number is one inlet at one state, the nameplate state, and it says nothing about how far the hall's real coils and real airflow may sit from it. Nor does it say anything about the four racks it did not pick as the hottest. The question of this chapter is the planner's own,

On the design-day ambient, with every unit running, what is the probability of a violation anywhere?

It is answered with the posterior of hall B that Chapter 5 fitted, carried to a new scenario node, the design day. What is new is the event asked about: not one inlet against its limit, and not one load at which a limit binds, but the event that any of the five inlets crosses the limit, a tail of the five inlets taken together.

Jargon: design day, envelope and fresh-air gain, violation anywhere, joint tail, marginal tail, worst marginal, union (Bonferroni) bound, independence product, correlation coefficient, Slepian's inequality, shared input, Monte Carlo standard error, baffle, blanking panel

The *design day* is the hot day the drawing sizes the cooling for: its outdoor temperature and humidity set what the chilled-water plant must deliver and what heat reaches the hall from outside. The *envelope and fresh-air gain* is that outside heat: the heat conducted through the hall's walls and roof, and the heat of the ventilation air brought in from outdoors, which on a hot day enters warmer than the hall. A *violation anywhere* is the event that at least one of the constrained entries, here the five rack inlets, is beyond its limit. Its probability is a *joint tail*, a tail of the entries taken together; the probability that one entry alone is beyond its limit is that entry's *marginal tail*, and the largest of them is the *worst marginal*. The *union bound*, or *Bonferroni bound*, is the sum of the marginal tails, which no union of events can exceed. The *independence product* is the probability of a violation anywhere computed as if the entries were independent, one minus the product of the probabilities that each stays inside. The *correlation coefficient* of two entries is their covariance divided by the product of their standard deviations, between -1 and 1. *Slepian's inequality* states that for jointly normal quantities whose correlations are all nonnegative, the probability that all stay below their limits is at least the product of the probabilities that each does. A *shared input* is a quantity that moves several entries at once, such as the chilled-water supply for every inlet of a hall; in reliability engineering a shared cause of several failures is a common-mode cause. The *Monte Carlo standard error* of a probability estimated as the share of N independent draws in which an event holds is $\sqrt{p(1-p)/N}$. A *baffle* is a panel that blocks the path by which return air reaches a rack's inlet; a *blanking panel* closes an empty slot of a rack for the same purpose. Both lower the rack's recirculation fraction.

Key idea

A violation anywhere is one event over all the constrained entries, and its probability is read from the posterior of the state, by draws pushed through every entry at the scenario. The worst marginal and the union bound bracket it from below and above. Where one entry carries nearly all the risk, the bracket is narrow and the worst marginal is nearly the answer; where several entries sit near the limit and share an uncertain input, the bracket opens and the worst marginal understates the risk. Which case a hall is in is read from the correlations of its entries, which the posterior supplies.

7.1.1 A violation anywhere as a joint tail

Write A_k for the event that the inlet of rack k is above the limit at the scenario, $T_k^* > T_{\max}$, and V for the event that at least one is. The violation anywhere is the union of the five events, and it is the same as the largest inlet being above the limit,

$$V = \bigcup_k A_k = \Big\{ \max_k T_k^*(\mathbf{z}) > T_{\max} \Big\}, \qquad \Pr(V \mid \mathbf{y}) = 1 - \Pr\Big(T_k^*(\mathbf{z}) \le T_{\max} \text{ for every } k \;\Big|\; \mathbf{y} \Big), \tag{7.1}$$

with $T_k^*(\mathbf{z})$ the inlet of rack k at the scenario as a function of the state and $\mathbf{y}$ the evidence. The second form says what has to be computed: the probability that every inlet stays inside at once, a probability over the five inlets together. It depends on how the five move together, which no single inlet's distribution carries.

Two bounds hold whatever that dependence is. The union contains each of its events, so its probability is at least the largest of theirs; and the probability of a union is at most the sum of the probabilities of its events, the overlaps being counted more than once in the sum. With

$p_k = \Pr(A_k \mid \mathbf{y})$ the marginal tail of rack k,

$$\max_k p_k \;\le\; \Pr(V \mid \mathbf{y}) \;\le\; \min\Big(1,\; \sum_k p_k\Big), \tag{7.2}$$

the worst marginal on the left and the union bound on the right. The left bound is reached when the other events lie inside the worst one, the right when no two of them can happen together. Between the two sits the independence product,

$$P_{\text{ind}} = 1 - \prod_k (1 - p_k), \tag{7.3}$$

the joint tail if the entries were independent. For entries that are jointly normal with correlations that are all nonnegative, the probability that all stay inside is at least the product of the probabilities that each does (Slepian's inequality), so that the joint tail lies at or below P_{ind}. The worst marginal is therefore the joint answer when one entry carries the risk alone, and it falls short of the joint answer by as much as the other entries add to the risk that the worst one does not already cover. The bounds cost only the marginals; the answer itself needs the dependence.

Principle 7.1. *A violation anywhere is the union of the entries' exceedances. Its probability lies between the worst marginal and the sum of the marginals, and is read jointly from draws of the state; the bracket's width says how much the dependence of the entries matters.*

7.1.2 Hall B at the design-day scenario (step 0)

The plant is hall B of Chapter 5, with its rows, equation (5.2), its reduced state of ten free entries, equation (5.4), and the graph of Figure 5.1: three cooling units B1–B3 whose coils carry the log-derates g_u, five racks B1–B5 whose inlets carry the recirculation fractions β_k, and the inlets of racks B2 and B4 measured. The design day changes nothing in the graph. It changes the inputs pinned at the scenario node of Section 5.1.3.

The scenario node of the design day. Four inputs are pinned, each a modeling choice. The IT load is the design load, $Q^* = 340\,\text{kW}$. The other heat is raised to

$$Q_o^* = Q_o + \Delta Q_{\text{dd}} = 10 + 15 = 25\,\text{kW},$$

with $Q_o = 10\,\text{kW}$ the fans and lighting of Chapter 5 and $\Delta Q_{\text{dd}} = 15\,\text{kW}$ the design day's envelope and fresh-air gain. The gain belongs to the design day and to no other scenario: it is heat from outside, conducted through the walls and roof and carried in by the ventilation air, and it is largest on the hottest day. It is a stated planning input, not a quantity the night record can inform, and it enters the hall balance of equation (5.2) as the fans' heat does, as heat added to the hall air; its size is varied in Section 7.4.2. Every unit runs. The supply is pinned at the drawing's design-day value, $T_s^* = 14.0°\text{C}$: this makes the check a statement about the hall on the drawing's terms, and it fails if the chilled-water plant cannot hold its supply on the design day, a limit of the central plant that Part III takes up; Section 7.3.5 lets the supply float as a stated harder case.

At these pins the scenario inlets follow from equation (5.7) with the design day's heat in place of the ordinary day's,

$$T_k^*(\mathbf{z}) = T_s^* + a_k(\mathbf{z})\,(Q^* + Q_o^*), \qquad a_k(\mathbf{z}) = \frac{1}{\sum_u \bar{G}_u\, e^{g_u}} - \frac{1-\beta_k}{\sum_u C_u}, \tag{7.4}$$

with $Q^* + Q_o^* = 365\,\text{kW}$ the heat in the hall and a_k the inlet rise of rack k per kilowatt of Chapter 5, in K/kW. The design day moves every inlet by the same lever on its rise per kilowatt; it does not change the order in which the racks run warm.

The event as a binding load. By equation (7.4) rack k's inlet is above the limit exactly when $a_k(Q^* + Q_o^*) > T_{\max} - T_s^*$, and dividing by a_k and subtracting Q_o turns that into a statement about the binding load $Q_k^\dagger = (T_{\max} - T_s^*)/a_k - Q_o$ of equation (5.8),

$$T_k^* > T_{\max} \iff Q_k^\dagger < Q^* + Q_o^* - Q_o = 355\,\text{kW}, \qquad V = \{\min_k Q_k^\dagger < 355\,\text{kW}\}. \tag{7.5}$$

The violation anywhere on the design day is the event that the hall's binding load of Chapter 5, the minimum over its racks, falls below 355 kW: the design day's 15 kW of gain acts on every inlet as 15 kW more of IT load would. The joint tail of the five inlets and the lower tail of the minimum over the five binding loads are one event.

The gradients of the inlets. Each inlet at the scenario is a derived quantity of the state, and by equation (7.4) its gradient is the lever $Q^* + Q_o^*$ times the gradient of the rise per kilowatt,

$$\nabla_{\mathbf{z}} T_k^* = (Q^* + Q_o^*)\,\nabla_{\mathbf{z}} a_k, \qquad \frac{\partial a_k}{\partial g_u} = -\frac{G_u}{\left(\sum_u G_u\right)^2}, \qquad \frac{\partial a_k}{\partial \beta_k} = \frac{1}{\sum_u C_u}, \tag{7.6}$$

in K per unit of each entry, every other entry zero. The log-derate entries are the same for all five racks: a coil below its rating warms the return air and with it every inlet. The fraction's entry sits in rack k's own column alone. Stacking the five gradients as the rows of a 5×10 matrix $\mathbf{D}$, the delta method of equation (4.53), applied to the five inlets at once, gives their covariance and their correlations,

$$\mathbf{C} = \mathbf{D}\,\boldsymbol{\Sigma}\,\mathbf{D}^\mathsf{T}, \qquad \rho_{jk} = \frac{C_{jk}}{\sqrt{C_{jj}\,C_{kk}}}, \tag{7.7}$$

with $\boldsymbol{\Sigma}$ the covariance of the calibrated state and ρ_{jk} the correlation coefficient of the inlets of racks j and k.

The structure of $\boldsymbol{\Sigma}$ fixes the shape of $\mathbf{C}$ before any number. Write $\mathbf{c}$ for the common part of the gradients, the log-derate entries, so that $\nabla_{\mathbf{z}} T_k^* = (Q^* + Q_o^*)(\mathbf{c} + \mathbf{e}_{\beta_k}/\sum_u C_u)$. For the three unmeasured racks B1, B3 and B5 the fraction is uncorrelated with every other entry, equation (5.14), and the covariance splits into a common term and the rack's own,

$$C_{jk} = s_c^2 \quad (j \neq k), \qquad C_{kk} = s_c^2 + \left(\frac{Q^* + Q_o^*}{\sum_u C_u}\right)^2 \text{Var}(\beta_k \mid \mathbf{y}), \qquad s_c^2 = (Q^* + Q_o^*)^2\,\mathbf{c}^\mathsf{T}\boldsymbol{\Sigma}\,\mathbf{c}, \tag{7.8}$$

with s_c the spread every inlet takes from the coils, in K. Two unmeasured racks are correlated only through the coils, and their correlation is s_c^2/C_{kk}, the coils' share of an inlet's variance. The measured racks B2 and B4 have their fractions read through the same rows as the coils' log-derates, and their correlations carry those cross terms too.

7.1.3 The planner's questions (step 1)

The planner asks four questions of hall B on the design day, and step 1 writes each as a probability or a moment over the reduced state, with the unknown on the left of the bar and the evidence on the right. Table 7.1 lists them. The first column is the question in the planner's words, the drill question first; the second the general form of its category in the notation of Chapter 2; the third its mathematical form for hall B; the fourth its category; the fifth the computation of Section 2.3.6 that answers it; and the last the subsection where it is answered. The first, third and fourth are predictions under a threshold: the joint tail, the rack that carries it, and the joint tail set against its bounds. The second is estimation, the five inlets with their spreads. The fourth is answered in two subsections, on the hall after a remedy and with a shared input, because each shows a different reason for the joint tail to depart from the worst marginal.

Table 7.1: The questions asked of hall B on the design day. $\mathbf{y}$ is the eight readings and the nine priors of window 1; T_k^* is the inlet of rack k at the design-day scenario (7.4); V is the event of a violation anywhere (7.1).

Planner's question	General form	Mathematical form	Category	Computation	Answered in
On the design-day ambient, with every unit running, what is the probability of a violation anywhere?	$\Pr(\max_k f_k(\mathbf{z}) > f^* \mid \mathbf{y})$	$\Pr(\max_k T_k^* > 27^\circ\mathrm{C} \mid \mathbf{y})$	prediction under a threshold	posterior	7.3.2
How warm is each inlet on the design day, and how sure?	moments, $\mathbb{E}[f]$ and $\mathrm{Var}(f)$	T_k^*, its sd, and $\Pr(T_k^* > 27^\circ\mathrm{C} \mid \mathbf{y})$ per rack	estimation	posterior	7.3.1
Which rack violates, if one does?	$\Pr(A \mid B)$	$\Pr(T_k^* > 27^\circ\mathrm{C} \mid V,\ \mathbf{y})$	prediction under a threshold	posterior	7.3.3
When does the joint answer differ from the worst rack alone?	the joint tail against its bounds	joint, worst marginal, union bound, independence product; the inlets' correlations	prediction under a threshold	posterior	7.3.4, 7.3.5

7.1.4 The evidence as rows (step 2)

The evidence is that of Chapter 5, attached to the graph in Section 5.1.5: window 1 of the night record, at 19:00, with eight readings (the supply, the return air, the three discharge temperatures, the hall load meter and the inlets of racks B2 and B4) and nine priors (the chiller's set point on the supply, the commissioning record on the three log-derates and the airflow survey on the five fractions). No row is added, and the fit is not repeated.

Two choices make this the evidence of the design-day check. The design day is a scenario, not a reading: nothing in a night window observes the design day's heat, and the gain ΔQ_{dd} is pinned at the scenario rather than fitted. And the coils and the fractions are taken to be the same on the design day as in window 1: this is what lets one night's fit answer a question about the hottest day, and it fails if the airflow pattern of the hall, and with it a fraction, changes with the outdoor conditions or with the load. The pattern of the reading rows, equation (5.10), carries straight over to the question: no reading sees the fractions of racks B1, B3 and B5, and every one of those racks' design-day inlets has an entry in its own fraction's column, equation (7.6). Three of the five entries of the joint tail are known on the design day only through the survey.

7.2 The posterior of the state (step 3)

All four questions of Table 7.1 are read from one posterior, that of Chapter 5: the MAP state $\hat{\mathbf{z}}$ and the Laplace covariance $\boldsymbol{\Sigma}$ of the reduced state under the seventeen rows of window 1, with the exact posterior beside it by reweighted draws.

7.2.1 The fit of window 1, carried to the design day

The rows, the normal equations and the fit are those of Section 5.2: the residual vector of equation (5.11), the Gauss–Newton iteration of equation (5.12), and the Laplace posterior

$$\mathbf{z} \mid \mathbf{y} \approx \mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma}), \qquad \boldsymbol{\Sigma} = \left(\mathbf{J}^{\mathsf{T}}\mathbf{J}\right)^{-1}\Big|_{\hat{\mathbf{z}}},$$

of equation (5.13), with the MAP state and the spreads of Table 5.3. The fitted coils sit slightly below their ratings, the fractions of racks B2 and B4 are narrowed by their sensors to a spread of 0.032, and the fractions of racks B1, B3 and B5 keep the survey's values and its spread of 0.08. The exact posterior is represented by the 50 000 Laplace draws of Section 5.2.3 (seed 5) under the importance weights (5.16), with an effective sample size of 49 638. The design day adds nothing to the posterior; it adds a scenario through which each draw, and the MAP state, are pushed.

7.2.2 The covariance of the design-day inlets

The gradients (7.6) at $\hat{\mathbf{z}}$, pushed through $\boldsymbol{\Sigma}$ by equation (7.7), give the correlation matrix of the five design-day inlets, rows and columns in the order B1 to B5,

$$[\rho_{jk}] = \begin{bmatrix} 1 & 0.083 & 0.060 & 0.084 & 0.060 \\ 0.083 & 1 & 0.083 & 0.199 & 0.083 \\ 0.060 & 0.083 & 1 & 0.084 & 0.060 \\ 0.084 & 0.199 & 0.084 & 1 & 0.084 \\ 0.060 & 0.083 & 0.060 & 0.084 & 1 \end{bmatrix}. \tag{7.9}$$

This matrix decides how far the joint tail can depart from the worst marginal, and it is low everywhere. The off-diagonal entries run from 0.060 to 0.199. The three unmeasured racks are correlated with each other at 0.060, the value equation (7.8) gives for inlets that share only the coils: each such inlet takes a spread of $(Q^* + Q_o^*)/\sum_u C_u \times 0.08 = 365/30 \times 0.08 = 0.973\,\mathrm{K}$ from its own fraction, and only the much smaller common spread s_c from the coils, which is the 0.247 K that rack B3's inlet keeps once its fraction is pinned (Section 7.3.4). The largest entry, 0.199, joins the two measured racks, B2 and B4, whose fractions are read through the same coils' entries. Each rack's uncertainty on the design day is its own airflow fraction, and the five inlets are close to independent.

The exact posterior is represented by the reweighted draws, each pushed through the five inlets (7.4). The joint tail is read from them as a share of the draws,

$$\begin{gathered} \mathbf{z}_i \sim \mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma}), \qquad \hat{P}_V = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\Big[\max_k T_k^*(\mathbf{z}_i) > T_{\max}\Big], \\ \hat{P}_V^{\text{exact}} = \sum_{i=1}^{N} w_i\, \mathbf{1}\Big[\max_k T_k^*(\mathbf{z}_i) > T_{\max}\Big], \end{gathered} \tag{7.10}$$

with $N = 50\,000$, w_i the weights of equation (5.16), and the Monte Carlo standard error $\sqrt{\hat{P}_V(1-\hat{P}_V)/N}$ of the reported share. The first is the state route through the Laplace normal; the second the exact posterior, the same draws reweighted.

7.3 The answers (step 4)

Every answer is read from the pair $(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$ of Section 7.2.1 and from the draws of the state. A single inlet is carried by its gradient at $\hat{\mathbf{z}}$; the joint tail is read from the draws (7.10), and the exact posterior stands beside it with the gap.

7.3.1 The design-day inlets

Planner's question. How warm is each inlet on the design day, and how sure? The question asks for each rack's inlet at the design-day scenario, with its spread and its own probability of crossing the limit.

Table 7.2: The five inlets of hall B at the design-day scenario: the design sheet's value at nameplate and survey, the posterior mean and standard deviation by the delta method, and the marginal tail beyond 27°C by equation (7.11) and by 50 000 Laplace draws.

Rack	Sheet [°C]	$T_k^*(\hat{\mathbf{z}})$ [°C]	sd [K]	p_k, delta [–]	p_k, draws [–]
B1	21.648	22.024	1.004	0.0000	0.0000
B2 (measured)	22.864	23.445	0.396	0.0000	0.0000
B3	25.298	25.674	1.004	0.0933	0.0919
B4 (measured)	24.081	24.055	0.398	0.0000	0.0000
B5	23.473	23.849	1.004	0.0008	0.0009

General form. An estimation question about a derived quantity asks for its posterior moments, equation (2.10),

$$\mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}] = \int f(\mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \qquad \mathrm{Var}(f(\mathbf{z}) \mid \mathbf{y}) = \int \big(f(\mathbf{z}) - \mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}]\big)^2\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

and each rack's tail is the threshold form (2.12) with the limit as the level.

Specialized form. The question is equation (2.10) with $f(\mathbf{z}) = T_k^*(\mathbf{z})$ of equation (7.4) for each rack and $p(\mathbf{z} \mid \mathbf{y})$ the Laplace normal. By the scenario node of equation (4.62) the inlet is nearly normal, with its mean at $T_k^*(\hat{\mathbf{z}})$ and its variance C_{kk} of equation (7.7), and its marginal tail is

$$p_k = \Pr(T_k^* > T_{\max} \mid \mathbf{y}) \approx 1 - \Phi\left(\frac{T_{\max} - T_k^*(\hat{\mathbf{z}})}{\sqrt{C_{kk}}}\right), \tag{7.11}$$

with Φ the standard normal distribution function. Each inlet is linear in its own fraction and nearly linear in the log-derates over their posterior width, so the normal is close; its departure is the difference between this delta route and the share of the draws in which the inlet is above the limit, both reported.

Building blocks. The MAP state and the covariance (Section 7.2.1); the scenario inlets (7.4) and their gradients (7.6); the covariance (7.7); the normal distribution function.

The inlets. Table 7.2 gives the five inlets. Each row is a rack; the columns give the design sheet's inlet at nameplate and survey, the inlet at the MAP state with its delta-method standard deviation, and its marginal tail by equation (7.11) and by the draws. The five densities are drawn in Figure 7.1: the horizontal axis is the design-day inlet, the vertical axis its posterior density, one normal curve per rack from its mean and standard deviation; the vertical black line is the 27°C limit, the red area rack B3's tail beyond it, the dashed line the sheet's 25.30°C for rack B3, and the dotted line rack B3's inlet at the stated truth of the record, a check marker for Section 7.4.3.

The table and the figure show two kinds of inlet. The measured racks B2 and B4 are narrow, 0.396 and 0.398 K, because their sensors fix their fractions; the unmeasured racks are wide, each 1.004 K, and almost all of that width is the survey's fraction. Rack B3 runs warmest, at 25.674°C: by equation (7.4) that is $14 + 0.031983 \times 365$, with the rise per kilowatt of Table 5.4, and its standard deviation is $365 \times 0.002751 = 1.004$ K, the spread s_a of Chapter 5 carried by the lever. The posterior puts rack B3 0.38 K warmer than the sheet, because the fitted coils sit slightly below their ratings and raise every rack's rise per kilowatt. Its tail beyond the limit is 0.0933: the limit is $(27 - 25.674)/1.004 = 1.32$ standard deviations above the mean. Rack B5 follows at 0.0008, and the other three are below 0.00005. The delta route and the draws agree to 0.0014 on rack B3, about the Monte Carlo standard error of the share.

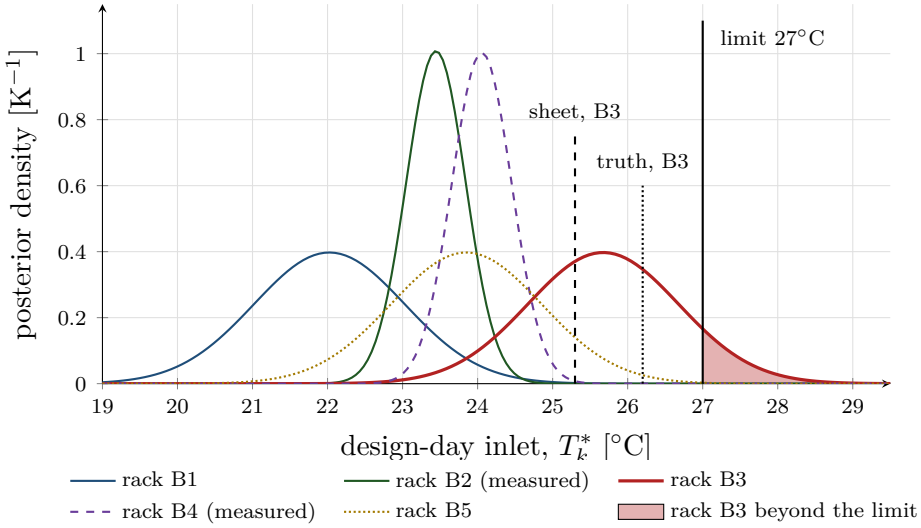


Figure 7.1: The posterior densities of the five design-day inlets of hall B, each the normal with the mean and standard deviation of Table 7.2, at $Q^* = 340\,\text{kW}$, $Q_o^* = 25\,\text{kW}$ and $T_s^* = 14°\text{C}$. Red area: rack B3's tail beyond the 27°C limit, 0.0933 by equation (7.11). Dashed: the design sheet's 25.298°C for rack B3 at nameplate and survey. Dotted: the stated truth of the record for rack B3, 26.200°C.

7.3.2 The probability of a violation anywhere

Planner's question. On the design-day ambient, with every unit running, what is the probability of a violation anywhere? The question asks for the probability that at least one of the five inlets is above 27°C at the design-day scenario.

General form. A prediction under a threshold is a tail of a derived quantity under the posterior, equation (2.12),

$$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y}) = \int \mathbf{1}[f(\mathbf{z}) \geq f^*]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

with an upper limit on a temperature written by $f = -T$, and here taken over the largest of several entries.

Specialized form. The question is equation (2.12) with $f(\mathbf{z}) = -\max_k T_k^*(\mathbf{z})$, the inlets of equation (7.4), and the level $f^* = -T_{\max}$, taken in its complement: the probability of the event V of equation (7.1). The maximum over the racks has no single gradient, since which rack is warmest changes across the posterior, so the reported value is the joint share of the draws (7.10) and the exact posterior is the weighted share. Beside them stand the bounds (7.2) and the independence product (7.3), from the marginal tails of Section 7.3.1, and the worst marginal by the delta route and by the draws. By equation (7.5) the same event is the hall's binding load below 355 kW.

Building blocks. The Laplace draws and the weights (5.16); the scenario inlets (7.4) in each draw; the marginal tails (7.11); the bounds (7.2) and the product (7.3).

The joint tail. By the draws (7.10) the reported answer is

$$\Pr(\max_k T_k^* > 27°\text{C} \mid \mathbf{y}) = 0.0927, \tag{7.12}$$

with a Monte Carlo standard error of 0.0013. The exact posterior, the same draws reweighted, gives 0.0942, a gap of 0.0015, exact minus reported, a little above the standard error: it is the departure of the posterior of the state from the Laplace normal, read in a tail that holds about one draw in eleven. Beside them, the worst marginal is rack B3's, 0.0933 by the delta route and 0.0919 by the draws; the union bound of the delta marginals is 0.0942, and their independence product 0.0941. On the draws alone the bracket (7.2) reads $0.0919 \le 0.0927 \le 0.0919 + 0.0009 = 0.0928$: the joint tail sits 1.009 times the worst marginal, and the bracket is narrower than the Monte Carlo standard error. Read through equation (7.5), the share of the draws in which the hall's binding load falls below 355 kW is 0.0927 as well, the same event counted the other way.

The design day triples the probability of the ordinary day. At the same design load without the design day's gain, the probability of a violation anywhere is Chapter 5's probability that the binding load falls below 340 kW, $1 - 0.9698 = 0.0302$; the 15 kW of gain moves the threshold of equation (7.5) from 340 to 355 kW, further into the binding load's band, and the probability to 0.0927.

7.3.3 The rack that violates

Planner's question. Which rack violates, if one does? The question asks, given that the design day produces a violation, how the probability of it is shared among the racks, and how often more than one rack is over at once.

General form. A conditional probability is the probability of the joint event over that of the conditioning event, equation (3.4), each a tail of the posterior by equation (2.12),

$$\Pr(A \mid B,\ \mathbf{y}) = \frac{\Pr(A \cap B \mid \mathbf{y})}{\Pr(B \mid \mathbf{y})}.$$

Specialized form. With $A = A_k$, rack k over the limit, and $B = V$, a violation anywhere, the event A_k lies inside V, so the joint event is A_k itself and, by equation (2.12) for each tail,

$$\Pr(A_k \mid V,\ \mathbf{y}) = \frac{p_k}{\Pr(V \mid \mathbf{y})}, \qquad \sum_k \Pr(A_k \mid V,\ \mathbf{y}) = \sum_{m \ge 1} \Pr(N \ge m \mid V,\ \mathbf{y}), \tag{7.13}$$

with N the number of racks over the limit, read on the draws as the share of the violating draws in which rack k is over. Given a violation $N \ge 1$ always, so the sum over the racks exceeds one by the probability that two or more racks are over together, and by that of three or more, and so on.

Building blocks. The draws in which V holds, from Section 7.3.2; the count of racks over the limit in each.

The violating rack. Given a violation, the rack over the limit is rack B3 with probability 0.9909 and rack B5 with probability 0.0097; racks B1, B2 and B4 are over in none of the violating draws, a share below 0.00005. Two or more racks are over with probability 0.0001, and with probability 0.0006 given a violation, which is the excess of the two shares over one, $0.9909 + 0.0097 = 1.0006$. Almost every design-day violation is rack B3 alone, the rack the sheet named; what the sheet missed is not the rack but the probability.

7.3.4 The joint answer after the remedy at rack B3

Planner's question. When does the joint answer differ from the worst rack alone? On the design day it barely does: the joint tail is 1.009 times the worst marginal, because one rack,

unmeasured, carries the risk, and the inlets are close to independent. That is not a reason to read the worst rack alone and stop. The question is asked again of the hall the review would build next: rack B3 baffled, and the load raised.

General form. The joint tail is the threshold form (2.12) over the largest entry, set against its bounds (7.2),

$$\max_k p_k \;\le\; \Pr\big(\max_k f_k(\mathbf{z}) > f^* \mid \mathbf{y}\big) \;\le\; \sum_k p_k.$$

Specialized form. The obvious remedy for rack B3 is to cut its recirculation with baffles and blanking panels. In the scenario it is a design change, pinned: rack B3's fraction is set to a stated design value $\beta^*_{B3} = 0.40$, in place of its free entry, while every other entry keeps its posterior. The inlets after the remedy are equation (7.4) at the state with that one entry replaced,

$$T^*_k(\mathbf{z};\ \beta^*_{B3}) = T^*_s + a_k\big(\mathbf{z} \text{ with } \beta_{B3} = \beta^*_{B3}\big)\,(Q^* + Q^*_o), \tag{7.14}$$

and the question is equation (2.12) over their maximum, read from the draws (7.10) pushed through equation (7.14), at the design load and at $Q^* = 380$ and 400 kW. Rack B3's gradient loses its fraction entry and keeps only the common part $\mathbf{c}$, so by equation (7.8) its standard deviation becomes s_c and its correlation with each unmeasured rack rises to $s_c/\sqrt{C_{kk}}$.

Building blocks. The Laplace draws and their weights; the remedied inlets (7.14); their gradients and covariance (7.7); the bounds (7.2) and the product (7.3).

The remedied hall. At the design load the remedy takes rack B3's inlet to 24.457°C with a standard deviation of 0.247 K, and its tail beyond the limit to below 0.00005. The probability of a violation anywhere falls to 0.0009 by the draws and by the exact posterior, and the risk moves to rack B5, at 23.849°C with 1.004 K, whose airflow is a survey number too: its marginal is 0.0009, and the joint tail, the union bound and the independence product all equal it at four decimals. The correlation of rack B3's inlet with the others rises, to 0.246 with racks B1 and B5 and 0.338 and 0.344 with racks B2 and B4, because its remaining spread is the coils' alone; the other entries of (7.9) are unchanged.

The remedy buys load, and as the load rises the joint tail moves away from the worst marginal. Table 7.3 sets the cases side by side. Each row is a case; the columns give the joint tail by the draws and by the exact posterior, the worst marginal by the draws with its rack, the union bound, the independence product and the ratio of the joint tail to the worst marginal. At 380 kW the joint tail is 0.0312 against a worst marginal of 0.0310, ratio 1.007, with a Monte Carlo standard error of 0.0008, the exact posterior at 0.0317 (a gap of 0.0005) and the independence product at 0.0312; at 400 kW it is 0.0974 against 0.0932, ratio 1.044, with a standard error of 0.0013, the exact posterior at 0.0991 (a gap of 0.0017), the independence product at 0.0980 and the union bound at 0.0985. The worst marginal is rack B5's in both, but at 400 kW the other racks, rack B3 among them, now near enough to the limit to add, contribute 0.0042 that rack B5 alone does not carry.

7.3.5 The design-day supply floating

Planner's question. When does the joint answer differ from the worst rack alone, if the design-day supply is not held? The remedied hall at 380 kW is asked again with one input shared by every inlet left uncertain: the chilled-water supply.

Table 7.3: The probability of a violation anywhere in hall B and its bounds, by case. Joint and marginals by 50 000 Laplace draws; the exact posterior by the same draws reweighted, with the floating supply's draws a stated input. Union and independence from the marginals by the draws, except on the design day, where they are from the delta marginals of Table 7.2. A dash marks a value the computation does not produce.

Case	Joint, draws [–]	Joint, exact [–]	Worst marginal [–]	Union bound [–]	Indep. product [–]	Joint over worst [–]
Ordinary day, 340 kW (Chapter 5)	0.0302	0.0308	–	–	–	–
Design day, 340 kW (reported)	0.0927	0.0942	0.0919 (B3)	0.0942	0.0941	1.009
Remedy, 340 kW	0.0009	0.0009	0.0009 (B5)	0.0009	0.0009	1.000
Remedy, 380 kW	0.0312	0.0317	0.0310 (B5)	0.0312	0.0312	1.007
Remedy, 400 kW	0.0974	0.0991	0.0932 (B5)	0.0985	0.0980	1.044
Remedy, 380 kW, supply floating	0.0503	0.0513	0.0447 (B5)	0.0542	0.0538	1.126

General form. As in Section 7.3.4, the threshold form (2.12) over the largest entry, set against its bounds; here the entries share an input outside the state,

$$\Pr\big(\max_k f_k(\mathbf{z}, u) > f^* \mid \mathbf{y}\big) = \iint \mathbf{1}\big[\max_k f_k(\mathbf{z}, u) > f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, p(u)\, \mathrm{d}\mathbf{z}\, \mathrm{d}u,$$

with u the shared input and $p(u)$ its stated distribution.

Specialized form. The question is equation (2.12) with the remedied inlets (7.14) at $Q^* = 380\,\mathrm{kW}$ and the supply no longer pinned at 14.0°C but drawn from a stated distribution, independent of the state,

$$T_s^* \sim \mathcal{N}(14.0,\ 0.5^2)\,^\circ\mathrm{C}, \qquad \mathbf{C}_{\text{float}} = \mathbf{C} + 0.5^2\, \mathbf{1}\mathbf{1}^\mathsf{T}, \tag{7.15}$$

with $\mathbf{1}$ the vector of ones: a warmer supply warms every inlet by the same amount, so it adds the same $0.25\,\mathrm{K}^2$ to every entry of the covariance (7.7), diagonal and off-diagonal alike. The spread stands for a central plant near its capacity on the design day, whose supply drifts above and below its set point. It is a stated sensitivity, not a model of the plant: the plant that sets the supply is the subject of Part III. Each draw of the state is paired with a draw of the supply and pushed through the five inlets; the joint tail, the marginals and the correlations are read from the pairs.

Building blocks. The Laplace draws and their weights; independent draws of the supply (7.15); the remedied inlets (7.14) at 380 kW; the bounds (7.2) and the product (7.3).

The shared input. The correlations with the supply floating are drawn beside those of the design day in Figure 7.2: each panel is a 5×5 grid, rows and columns the racks B1 to B5, each cell shaded by its correlation and labeled with it; the left panel is the design day of equation (7.9), the right the remedied hall at 380 kW with the supply floating, by the draws. The correlations rise from 0.060–0.199 to 0.217–0.765. The largest now join rack B3 to the measured racks, 0.765 with each: those three inlets have the smallest spreads of their own, so the supply's common 0.5 K is the larger part of them. The smallest, 0.217, joins racks B1 and B5, whose own fractions still dominate.

With the supply floating, the joint tail is 0.0503 against a worst marginal of 0.0447, rack B5's, a ratio of 1.126, with a Monte Carlo standard error of 0.0010; the exact posterior, the state's

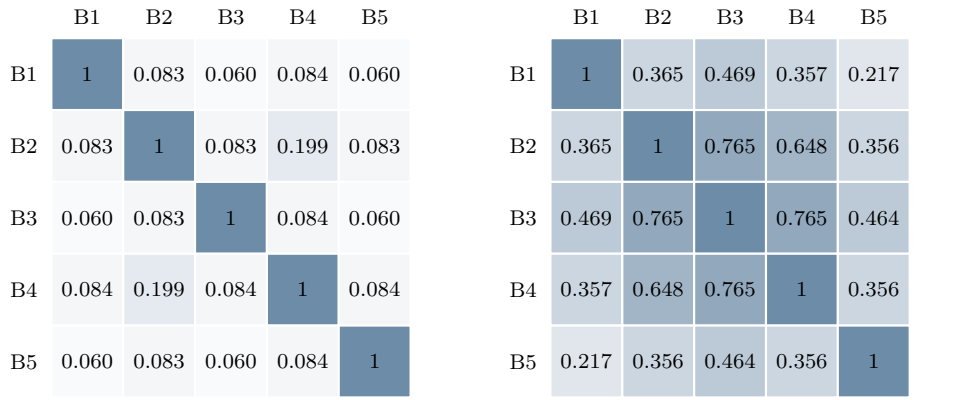


Figure 7.2: The correlations of the five inlets of hall B. Left: the design day at the pinned supply, by the delta method (7.7), equation (7.9). Right: rack B3 baffled to $\beta^*_{B3} = 0.40$, the load at 380 kW and the supply floating as in equation (7.15), by 50 000 draws. The shading deepens with the correlation.

draws reweighted with the supply's draws held as the stated input, gives 0.0513, a gap of 0.0010. The union bound is 0.0542 and the independence product 0.0538. The marginals are 0.0004, 0.0001, 0.0063, 0.0028 and 0.0447 for racks B1 to B5: the warm draws of the supply push racks B3 and B4 toward the limit too, and they add risk that rack B5 alone does not carry. The joint tail sits inside the bracket (7.2) and below the independence product, as Slepian's inequality requires for positively correlated entries, and the shared input has pulled it toward the worst marginal from where independence would put it, without closing the gap.

Figure 7.3 gathers the cases. The horizontal axis lists the five cases of Table 7.3 after the ordinary day; for each, three bars give the worst marginal, the joint tail (the reported answer) and the union bound, all by the draws, the vertical axis in probability. On the design day and on the remedied hall at the design load the three bars are one height. They separate at 400 kW, where several racks near the limit, and separate most with the supply floating, where several racks near the limit share an uncertain input. Read on the worst *measured* rack, B4, whose tail at 340 kW is below 0.00005 on the design day and after the remedy alike, the hall would have shown no risk at all.

7.4 The checks (step 5)

The answers of Section 7.3 rest on the posterior of Chapter 5, on the stated gain of the design day and on the constructed record. Each is tested here against the calculation it stands for. Table 7.4 collects the checks: each row is a check, the columns the quantity, the value and what it is set against.

7.4.1 The ordinary day against Chapter 5

With the gain set to zero, the scenario is Chapter 5's at the design load, and by equation (7.5) the probability of a violation anywhere must equal the probability that the hall's binding load falls below 340 kW. By the same draws the joint tail is 0.0302, and Chapter 5's $1 - \Pr(Q^\dagger \geq 340\,\text{kW})$ is 0.0302. The inlet route of this chapter and the binding-load route of Chapter 5 count the same draws.

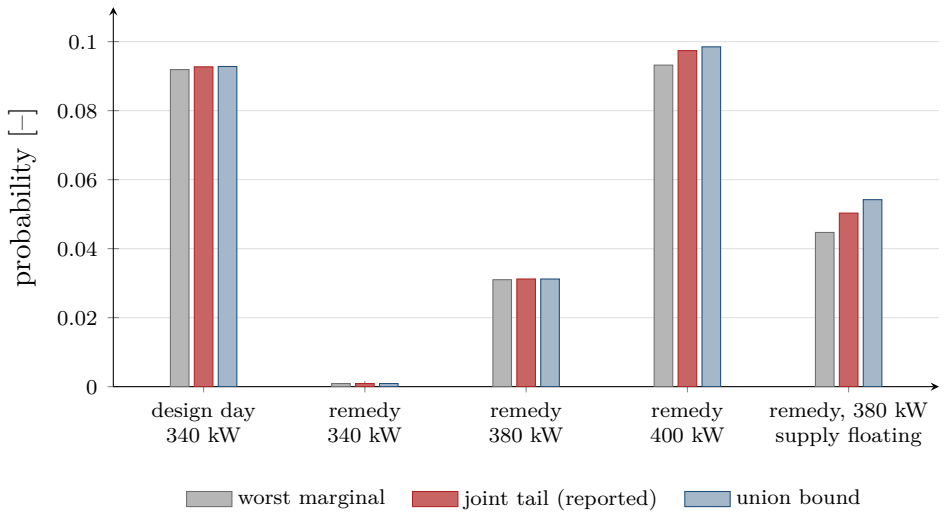


Figure 7.3: The probability of a violation anywhere in hall B against its bounds (7.2), by case, all by 50 000 Laplace draws. The union bound on the design day is the sum of the marginals by the draws, $0.0919 + 0.0009 = 0.0928$; the other values are those of Table 7.3.

7.4.2 Sensitivity to the design day's gain

The gain is a stated planning input, so the answer is tested against it, the joint tail repeated with ΔQ_{dd} at 0, 10, 15, 20 and 30 kW. The probability of a violation anywhere is 0.0302, 0.0673, 0.0927, 0.1259 and 0.2107. Each 5 kW of gain moves the threshold of equation (7.5) by 5 kW into the lower tail of the binding load, and the probability grows faster than in proportion as the threshold nears the body of the band: doubling the gain from 15 to 30 kW more than doubles it. An envelope and fresh-air gain known only to within 5 kW moves the answer from 0.0927 to 0.0673 or 0.1259, by −27 % and +36 %, more than any other approximation of this chapter.

7.4.3 The answers against the stated truth

The record of hall B is generated from a stated truth, and the last check compares the answers with it. At the true state the design-day inlets are 21.333, 23.402, 26.200, 23.888 and 24.132°C for racks B1 to B5: no rack is over, and the warmest, rack B3, sits at 26.200°C, 0.80 K inside the limit. That agrees with Chapter 5's true binding load of 378.93 kW, above the 355 kW of equation (7.5). The true rack B3 is 0.53 K warmer than the posterior mean and 0.90 K warmer than the sheet: the survey's fraction for rack B3 is low, and no reading corrects it. The true hall does not violate on the design day, an outcome the posterior gives probability $1 - 0.0927$; the truth is one hall, and a probability of 0.09 is not refuted by one hall that passes.

7.5 The report (step 6)

The design sheet's "pass" is a pass of the median. On the design day, at the 340 kW design load with every unit running and the supply at 14.0°C, hall B violates somewhere with probability 0.093, three times the ordinary day's, and almost all of it at rack B3, which no sensor reads.

7.5.1 The answers and their bands

Table 7.5 gives the answers in the order of Table 7.1. Each row is a question; the second column gives the reported answer with the exact posterior beside it where it is computed, and the third the routes and bounds that stand beside it.

Table 7.4: The checks of the design-day assessment of hall B. Joint tails by the 50 000 Laplace draws of Chapter 5.

Check	Value [–, °C]	Set against
Ordinary day, gain 0 kW, joint tail	0.0302	$1 - \Pr(Q^{\dagger} \geq 340\,\text{kW}) = 0.0302$, Chapter 5, same draws
Design day, joint tail as a binding load	0.0927	$\Pr(Q^{\dagger} < 355\,\text{kW}) = 0.0927$, same draws
Gain 10 kW	0.0673	reported 0.0927 at 15 kW
Gain 20 kW	0.1259	reported 0.0927 at 15 kW
Gain 30 kW	0.2107	reported 0.0927 at 15 kW
Exact posterior	0.0942	reported 0.0927; gap 0.0015, Monte Carlo se 0.0013
Stated truth, inlets B1 to B5	21.333, 23.402, 26.200, 23.888, 24.132	no violation; rack B3 0.80 K inside the limit
Stated truth, rack B3	26.200	posterior 25.674 ± 1.004; sheet 25.298

Table 7.5: The report for hall B on the design day: 340 kW of IT load, 25 kW of other heat, every unit running, the supply pinned at 14.0°C. Reported values by the Laplace draws of the state, joint over the five racks; the exact posterior beside them.

Question	Answer	Beside it
On the design-day ambient, with every unit running, what is the probability of a violation anywhere?	0.0927 (exact: 0.0942), Monte Carlo se 0.0013	worst marginal 0.0933 (delta), 0.0919 (draws); union bound 0.0942; independence 0.0941; ordinary day 0.0302
How warm is each inlet on the design day, and how sure?	rack B3 25.674 ± 1.004°C, tail 0.0933; rack B5 23.849 ± 1.004°C, tail 0.0008	racks B1, B2, B4: 22.024 ± 1.004, 23.445 ± 0.396, 24.055 ± 0.398°C, tails below 0.00005
Which rack violates, if one does?	rack B3, 0.9909 given a violation	rack B5 0.0097; two or more racks 0.0001
When does the joint answer differ from the worst rack alone?	on the design day, barely: ratio 1.009	after the remedy, 1.044 at 400 kW; with the supply floating at 380 kW, 0.0503 against 0.0447, ratio 1.126

7.5.2 The design-review entry

The design review's entry replaces the sheet's "pass". On the design day, at the design load, with every unit running and the supply held at the drawing's 14.0°C, the probability that some inlet of hall B exceeds 27°C is 0.093 by the joint draws over the five racks and 0.094 by the exact posterior, against 0.030 on an ordinary day at the same load. It is rack B3 in 99 % of the violations. The sheet's margin of 1.70 K at rack B3 is the margin of the median, and the band of rack B3's inlet is one airflow number, the survey's fraction, which no reading in the hall checks.

The entry then records the remedy and its order. Baffle rack B3 and fit blanking panels: at a fraction of 0.40, a stated design value, the design-day probability falls to 0.0009. The risk then moves to rack B5, whose airflow is also a survey number, and it grows with the load: 0.031 at 380 kW, 0.097 at 400 kW. So re-survey rack B5 before releasing the hall past 380 kW, and verify rack B3's fraction after the baffling, since the 0.40 is a design value, not a measurement. The entry also records the condition the whole assessment rests on, the supply held at 14.0°C; if the central plant lets it drift by 0.5 K, the remedied hall at 380 kW violates with probability 0.050, not 0.031, and that is the question of Part III.

Common mistake

Reading the design-day check off the worst rack at the nameplate state. The nameplate inlet is one state; the check is a probability over the states the evidence allows, joint over every rack. Where one rack carries the risk, the worst marginal is close to the joint answer, but only the correlations say so, and the measured racks are not the ones to read.

7.6 The next measurement (step 7)

The design-day probability rests on two survey numbers in turn: the fraction of rack B3 now, and that of rack B5 once rack B3 is baffled. A measurement is one more row of the residual vector, and what it removes from a design-day inlet's spread is known before its value is, by the preposterior of equation (5.26).

7.6.1 The probe at rack B3

The leading candidate of Chapter 5, Table 5.9, is an inlet probe at rack B3 of standard deviation 0.1 K. Its row reads today's inlet of rack B3, and its preposterior covariance $\mathbf{\Sigma}_a = (\mathbf{\Sigma}^{-1} + \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a)^{-1}$ is pushed through the gradients (7.6) of the five design-day inlets. It takes rack B3's design-day standard deviation from 1.004 to 0.232 K, and leaves the others at 1.004, 0.396, 0.398 and 1.004 K for racks B1, B2, B4 and B5. The width it removes from rack B3's inlet, $1 - 0.232/1.004$, is the 76.9 % it removes from the binding load in Table 5.9, since both quantities are carried by the same rise per kilowatt. After the probe the design-day probability of rack B3 is decided by where the reading puts its fraction, within a quarter of a kelvin, instead of by the survey's full kelvin.

7.6.2 After the remedy, rack B5's airflow

The probe leaves rack B5's design-day spread at 1.004 K: its survey is untouched, and nothing at rack B3 informs it. Once rack B3 is baffled, rack B5 carries the design-day risk, and its spread is almost all its own fraction, $365/30 \times 0.08 = 0.973$ K of the 1.004 K by equation (7.8). The next measurement after the remedy is therefore rack B5's airflow, by a re-survey of its fraction or a probe at its inlet, and with it a re-survey of rack B3 after the baffling, which replaces the design value 0.40 by a measured one. Chapter 18 ranks the tests and surveys a planner can commission on the site plant in the same way, by the width each takes off the quantity that decides the build.

7.6.3 What the analysis showed

The design sheet passed hall B on the design day by 1.70 K at rack B3, and the analysis puts the probability of a violation anywhere at 0.093: the posterior puts rack B3's inlet 0.38 K warmer than the sheet, and its mean only 1.32 standard deviations below the limit, for an inlet no sensor reads. The design day is the ordinary day with 15 kW more heat, and it triples the probability, because the binding load's band reaches below 355 kW three times as often as below 340 kW. On this hall the joint answer is almost the worst marginal, 1.009 times it, and the reason is read from the posterior, not assumed: one unmeasured rack carries the risk, and the five inlets are correlated at 0.06 to 0.20, because each rack's uncertainty is its own airflow fraction. That did not make the joint answer unnecessary; it made it easy to check. Once rack B3 is baffled, the risk moves to rack B5 and the gap grows with the load; once the supply is left uncertain, the inlets' correlations rise to 0.22 to 0.77 and the joint answer is 1.126 times the worst marginal. The joint answer departs from the worst marginal when several entries are near the limit and their uncertainty is shared, and the worst marginal and the union bound bracket it. The worst measured rack would have reported no risk at all.

7.6.4 Limits of the model

The supply pinned. The design day is assessed at the drawing's 14.0°C supply. Whether the central plant can hold that supply on the design day, at the load of both halls, is a capacity question of the chillers, and Part III takes it up; the floating supply of Section 7.3.5 is a stated spread, independent of the hall's state, not a model of the plant.

The gain stated, steady and in the hall air. The design day's 15 kW is a planning input, not estimated; the answer moves by a third for 5 kW either way (Section 7.4.2). It enters the hall balance as heat added to the hall air at a steady peak; a gain that reaches the cold aisle directly, or a peak too short for the hall to reach steady state, would change the inlets differently.

One night's airflow on the hottest day. The coils and the fractions are those fitted on window 1 of a night. A fraction that changes with the outdoor conditions, the load or the racks' fan speeds would move the inlets in a way this scenario does not carry, and the survey that sets three of the five fractions was taken at another time.

The remedy as a design value. The baffled fraction $\beta^*_{B3} = 0.40$ is pinned at a stated value with no spread. A remedy that does less than designed leaves rack B3 in the joint tail, and only a measurement after the work shows which.

The constructed record. The readings of window 1 are generated from a stated truth with the stated sensor errors, not taken from a hall, and the comparison with that truth in Section 7.4.3 is a check of the method on its own model, not of the model against a real hall.

The next chapter. Chapter 8 opens the IT load into its racks, whose peaks do not coincide, and asks what load the hall actually sees at its peak.

7.7 What to retain

Definition 7.1 (Design-day check)**.** The design-day check is the probability of a violation anywhere at the design-day scenario: the IT load at its design value, the other heat raised by the design day's envelope and fresh-air gain, every unit running and the supply pinned. The event is the union of the constrained entries' exceedances, $V = \{\max_k T^*_k > T_{\max}\}$; its

probability is read jointly, from draws of the posterior of the state pushed through every entry, with the exact posterior beside it, and is bracketed by the worst marginal and the union bound, $\max_k p_k \le \Pr(V \mid \mathbf{y}) \le \sum_k p_k$.

Four conclusions carry beyond hall B. A design-day check is a probability, not a margin at the nameplate state; the margin of the sheet is the margin of the median. A heat gain on the design day acts on every inlet as the same amount of load would, so the design-day check is the binding load's tail read at the load plus the gain. The joint answer departs from the worst marginal only when several entries are near the limit and share their uncertainty, and the correlation matrix of the entries, read from the posterior, says which case holds. And a remedy at the worst entry moves the risk to the next unmeasured one, so a remedy is assessed jointly too.

Table 7.6 lists the terms of the chapter with their expressions, units and the subsection that defines each.

Table 7.6: The terms of the design-day assessment and where each is defined. Units for hall B.

Term	Expression	Units	Meaning	Defined in
Violation anywhere	$V = \{\max_k T_k^* > T_{\max}\}$	–	union of the exceedances	7.1.1
Bounds	$\max_k p_k \le \Pr(V) \le \sum_k p_k$	–	worst marginal, union bound	7.1.1
Independence product	$1 - \prod_k (1 - p_k)$	–	joint tail if independent	7.1.1
Design-day inlet	$T_k^* = T_s^* + a_k(Q^* + Q_o^*)$	°C	inlet at the design day	7.1.2
Event as binding load	$V = \{\min_k Q_k^\dagger < Q^* + Q_o^* - Q_o\}$	kW	gain acts as load	7.1.2
Inlet covariance	$\mathbf{C} = \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\mathsf{T}$, $\rho_{jk} = C_{jk}/\sqrt{C_{jj}C_{kk}}$	K^2, –	how the inlets move together	7.1.2
Common spread	$s_c^2 = (Q^* + Q_o^*)^2\, \mathbf{c}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{c}$	K^2	the coils' part of each inlet	7.1.2
Posterior and answers				
Joint draws	share of $\mathbf{z}_i$ with $\max_k T_k^*(\mathbf{z}_i) > T_{\max}$	–	the reported joint tail	7.2.2
Marginal tail	$1 - \Phi\big((T_{\max} - T_k^*(\hat{\mathbf{z}}))/\sqrt{C_{kk}}\big)$	–	one inlet alone	7.3.1
Violating rack	$\Pr(A_k \mid V) = p_k / \Pr(V)$	–	which rack, given a violation	7.3.3
Remedy	β_{B3} pinned at β_{B3}^*	–	a design change in the scenario	7.3.4
Shared input	$\mathbf{C} + 0.5^2\,\mathbf{1}\mathbf{1}^\mathsf{T}$	K^2	a floating supply	7.3.5

7.8 Exercises

Exercise 7.1. *Formulation.* Write the planner's question "what is the probability that rack B5 alone is over the limit on the design day" as a probability over the reduced state, give the gradient of the quantity, and state its answer from the fit of this chapter. ($\Pr(T_{B5}^* > 27°\mathrm{C} \mid \mathbf{y})$ with $T_{B5}^* = 14 + a_{B5}(\mathbf{z}) \times 365$ and $a_{B5} = 1/\sum_u G_u - (1 - \beta_{B5})/30$; by equation (7.6) the gradient is 365 times that of a_{B5}, with the coils' entries common to every rack and 365/30 in the column of β_{B5}. The inlet is $23.849 \pm 1.004°\mathrm{C}$, and its tail by equation (7.11) is 0.0008.)

Exercise 7.2. *Derivation.* Show that a violation anywhere on the design day is the event that the hall's binding load of Chapter 5 falls below 355 kW, and give its probability. (By equation (7.4), $T_k^* > 27$ exactly when $a_k \times 365 > 13$, that is when $13/a_k - 10 < 355$, which is $Q_k^\dagger < 355\,\mathrm{kW}$ by equation (5.8); a violation at some rack is the minimum over the racks below 355 kW. The share of the draws in which it holds is 0.0927.)

Exercise 7.3. *Derivation.* Explain from equation (7.8) why two unmeasured racks' design-day inlets are correlated at about 0.06. (Their fractions are uncorrelated with every other entry, so their covariance is the coils' common term s_c^2, with $s_c = 0.247\,\mathrm{K}$, and each inlet's variance is 1.004^2; the correlation is $0.247^2/1.004^2 = 0.061$, and the matrix prints 0.060.)

Exercise 7.4. *Application.* Compute rack B3's marginal tail on the design day from its mean and standard deviation, and say why the joint tail differs from it so little. (By equation (7.11), $(27 - 25.674)/1.004 = 1.32$ and $1 - \Phi(1.32)$ gives the delta route's 0.0933. The joint tail adds only what the other racks contribute outside rack B3's event: rack B5's marginal is 0.0008 and the correlations are low, so the joint tail by the draws is 0.0927 against rack B3's 0.0919 by the same draws.)

Exercise 7.5. *Application.* With rack B3 baffled, the load at 380 kW and the supply floating as $\mathcal{N}(14.0,\ 0.5^2)\,^\circ\mathrm{C}$, the marginal tails of racks B1 to B5 are 0.0004, 0.0001, 0.0063, 0.0028 and 0.0447. Give the bracket on the joint tail, and say where the joint tail lies in it and why. (By equation (7.2) the bracket is from the worst marginal, 0.0447, to the union bound, 0.0542, and the independence product (7.3) is 0.0538. The supply adds the same $0.25\,\mathrm{K}^2$ to every entry of the covariance, so the inlets are positively correlated, 0.217 to 0.765, and the joint tail lies below the independence product and above the worst marginal: 0.0503.)

Chapter 8

Load Diversity and Coincident Peaks

8.1 The coincident peak of hall B at its design population

Hall B is to be filled to its design population. Its IT load today is about 300 kW, its design load 340 kW, and the plan fills it by giving every rack the same share of the growth: each rack's load is scaled by $340/300 = 1.1333$. The design sheet checks the filled hall the customary way. The rack power meters have read each rack's load at the hall's daily peak hour for sixty days, and the sheet takes each rack's largest reading, 72.21, 66.13, 86.73, 75.62 and 65.08 kW for racks B1 to B5, 365.76 kW together. Scaled to the design population the sum of the maxima is 414.53 kW. The racks never all peak on the same day, so the sheet applies the customary diversity factor of 0.8 and arrives at a diversified peak of 331.62 kW. It sets that against the hall's capacity of 410 kW, the binding load the sheet worked from the coils' nameplate ratings and the airflow survey, and writes: pass, with 78 kW to spare.

The factor of 0.8 is a habit, not a measurement, and the 410 kW is a single number where Chapter 5 found a band. The racks of hall B also run one workload: when it is busy, they are all busy. The planner's question is the drill question, verbatim,

The racks never all peak at once; what is the probability the coincident peak exceeds the plant's capacity?

It is answered with the method of Chapter 4 on two states: the state of the rack record, fitted in this chapter, and the state of hall B, fitted in Chapter 5, whose binding load is the capacity. What is new is that both sides of the comparison are uncertain, and that the load side is a sum of correlated entries.

Jargon: coincident peak, diversity, coincidence (diversity) factor, design population, rack power meter, shared workload factor, one-factor covariance, correlation, spread, exceedance probability, plug-in, aleatory and epistemic uncertainty

The *coincident peak* of a group of loads is the largest value of their sum, as distinct from the sum of their individual largest values. *Diversity* is the fact that the loads do not all reach their largest values together. The *coincidence factor* is the coincident peak over the sum of the individual peaks, at most one; the design sheet calls it the *diversity factor* and applies it as a fixed multiplier. The *design population* of a hall is the set of racks it will hold at its design load. A *rack power meter* (a power distribution unit with metering) reads one rack's electrical load, which leaves the rack as heat. A *shared workload factor* is one day-to-day variation common to every rack, such as the busyness of the workload they all serve; the covariance it produces, with one correlation ρ for every pair of racks, is the *one-factor covariance*. The *spread* s_k of a rack's load is its standard deviation from day to day. An *exceedance probability* is the probability that a load exceeds a capacity. A *plug-in* value is a probability computed with every uncertain entry set to a single best value. *Aleatory* uncertainty is the variation that remains when the state is known, here the day-to-day variation of the loads; *epistemic* uncertainty is what is not known about the state (Section 3.1.3).

Key idea

Diversity is a covariance. The coincident peak is a sum of rack loads, and the variance of a sum is the sum of every entry of the loads' covariance matrix: the diagonal is each rack's own spread, and the off-diagonal terms are the diversity, small when the racks vary independently and as large as the diagonal allows when they move together. A fixed diversity factor asserts a covariance without measuring it. Fitted to the rack record, the covariance sets the level of the probability that the coincident peak exceeds the capacity; the capacity's own band sets how sure that probability is.

8.1.1 Diversity read as a covariance

On one day the five racks draw loads $L_1, \dots, L_5$, and the hall draws their sum. If the racks varied independently, a day on which one rack is high would usually find another low, and the sum would vary much less than the sum of the racks' variations. If they all followed one workload, a high day for one would be a high day for all, and the sum would vary as much as the racks' spreads added together. The coincident peak lies between these two cases, and where it lies is set by how the racks' loads move together, their covariance.

The question therefore has three parts. The load side is the sum of correlated entries, whose distribution follows from the racks' means, spreads and correlation; these are entries of a state, fitted to the rack record. The capacity side is the hall's binding load, whose posterior Chapter 5 computed from the hall's own readings. And the answer is the probability that one exceeds the other, which averages the day-to-day variation of the load, known in closed form once the state is known, over what is not known about either state.

> **Principle 8.1.** *A coincident peak is a sum of correlated loads. Its mean is the sum of the means; its variance is the sum of all entries of the loads' covariance, and the off-diagonal entries are the diversity. A diversity factor is a statement about those entries and is tested against them.*

8.1.2 Hall B's racks and its capacity (step 0)

Hall B is the hall of Chapter 5: three cooling units, five racks B1–B5, the inlets of racks B2 and B4 measured. This chapter adds the racks' electrical loads. Each rack has a power meter, and the meters have read each rack's load at the hall's daily peak hour for sixty days, $d = 1, \ldots, 60$.

The model of a day's rack loads. The load of rack k at the peak hour of day d is its mean, plus its spread times a variation made of a part shared by every rack and a part of its own,

$$L_{kd} = \mu_k + s_k \left(\sqrt{\rho}\, F_d + \sqrt{1-\rho}\, e_{kd}\right), \qquad F_d,\ e_{kd} \sim \mathcal{N}(0,\ 1) \text{ independent}, \tag{8.1}$$

with μ_k the rack's mean load and s_k its spread, both in kW, F_d the shared workload factor of day d, e_{kd} the rack's own variation, and ρ the weight of the shared part, dimensionless. Each meter reads its rack's load with an error of its own,

$$y_{kd} = L_{kd} + n_{kd}, \qquad n_{kd} \sim \mathcal{N}(0,\ \sigma_m^2), \qquad \sigma_m = 0.5\,\text{kW}, \tag{8.2}$$

independent of every other meter and day.

The state of the record. The record is described by eleven entries,

$$\boldsymbol{\theta} = (\mu_1, \ldots, \mu_5,\ \ln s_1, \ldots, \ln s_5,\ \operatorname{atanh}\rho)^{\mathsf{T}}, \tag{8.3}$$

five means in kW and six dimensionless entries. The spreads enter by their logarithms and the correlation by its inverse hyperbolic tangent, so that every entry ranges over the whole real line: a normal on these entries never puts a spread below zero or a correlation outside $(-1,\ 1)$, and the Laplace normal of step 3 is closer to the posterior on these scales than on the raw ones.

The capacity. The capacity is the binding load of hall B, $Q^{\dagger}(\mathbf{z}) = \min_k Q_k^{\dagger}(\mathbf{z})$ of equation (5.8): the hall IT load at which the first rack inlet meets the 27°C limit, at the drawing's design-day supply $T_s^* = 14$°C with all three units running, as a function of the hall's reduced state $\mathbf{z}$ of equation (5.4). It is a hall IT load in kW, the quantity the racks' sum is measured in.

The scenario. The question is asked at the design population, a scenario not yet seen. Every rack's load is scaled by $f = 340/300 = 1.1333$, the design IT load over today's, and the coincident peak of a day at the design population is

$$S^* = f \sum_{k=1}^{5} L_k = f\, \mathbf{1}^{\mathsf{T}}\mathbf{L}, \tag{8.4}$$

with $\mathbf{L}$ the vector of the five rack loads at the peak hour of a day and $\mathbf{1}$ the vector of ones; S^* is a load, so the meters' errors do not enter it. The event the planner asks about is $S^* > Q^{\dagger}(\mathbf{z})$ on a day at the design population.

Modeling choices at the graph. The racks enter the hall only through their sum: by equation (5.7) every inlet of hall B rises with the heat of the whole hall, not with its own rack's load, so a rack's own load does not move its own inlet in this model. This is what lets the capacity be one number of the hall's state, and it fails where a rack's inlet is warmed by its own exhaust in proportion to its own load (Section 8.6.4). The racks share one workload factor with the same weight ρ for every pair: this gives the covariance one entry, the one the diversity factor asserts, and it fails if the racks serve two workloads, which would correlate some pairs more than others. Each rack's load is normal at the peak hour, and the days are independent:

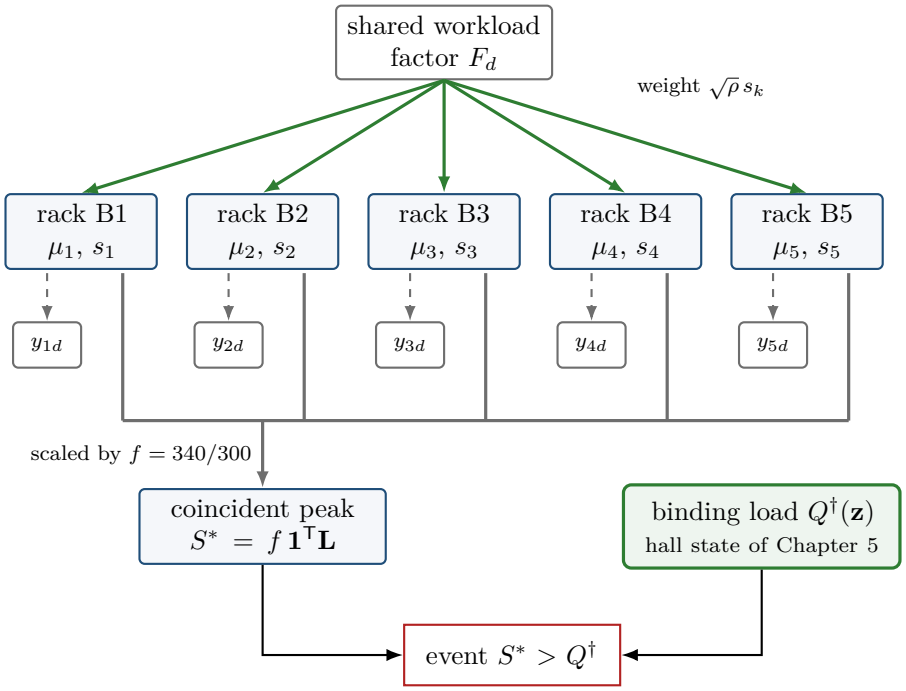


Figure 8.1: The rack record of hall B and the question asked of it. The shared workload factor F_d of day d feeds every rack with the weight $\sqrt{\rho}\, s_k$ (green arrows); each rack adds its own variation $\sqrt{1-\rho}\, s_k\, e_{kd}$, not drawn, around its mean μ_k, by equation (8.1). Dashed: the sixty days of meter readings y_{kd}, equation (8.2). Gray: the five loads summed and scaled to the design population, the coincident peak S^* of equation (8.4). Green box: the binding load of hall B, from the hall's state and its own readings (Chapter 5). Red: the event whose probability is asked.

this makes the coincident peak normal once the state is known, and it fails if the workload has a heavier upper tail than a normal or if busy days come in runs. The design population is today's racks scaled by f: means and spreads grow in proportion and the correlation is kept, which fails if the added load is a different workload. And the rack record and the hall's readings share no entry: the coils' derates and the racks' recirculation do not depend on the workload, so the two posteriors are independent.

The graph. The rows of the record and the comparison are drawn in Figure 8.1. The shared workload factor at the top feeds the five rack loads with the weight $\sqrt{\rho}\, s_k$; each rack load is read by its meter, the dashed arrows; the rack loads are summed along the line below them and scaled by f to the coincident peak S^*; and the capacity $Q^\dagger$ comes from the hall's state of Chapter 5, which this chapter does not refit. The event at the bottom is the one asked about.

8.1.3 The planner's questions (step 1)

The planner asks four questions of the rack record and the hall, and step 1 writes each as a probability, a quantile or a spread over the two states, with the unknown on the left of the bar and the evidence on the right. Table 8.1 lists them. The first column is the question in the planner's words, the drill question first; the second the general form of its category in the notation of Chapter 2; the third its mathematical form for hall B; the fourth its category; the fifth the computation of Section 2.3.6 that answers it; and the last the subsection where it is answered. The first is a prediction under a threshold, and of a kind not met before in Part II: the event compares a future load, random even when the state is known, with a capacity that

Table 8.1: The questions asked of hall B at its design population. $\mathbf{y}$ is the rack record, the hall's readings of Chapter 5 and the priors of both; S^* is the coincident peak of equation (8.4) and $Q^\dagger$ the binding load of equation (5.8).

Planner's question	General form	Mathematical form	Category	Computation	Answered in
The racks never all peak at once; what is the probability the coincident peak exceeds the plant's capacity?	$\Pr(y_{\text{new}} > f(\mathbf{z}) \mid \mathbf{y})$, a future value against a derived quantity	$\Pr(S^* > Q^\dagger \mid \mathbf{y})$	prediction under a threshold	posterior	8.3.1
How correlated are the racks?	$\Pr(z_i \le z^* \mid \mathbf{y})$, quantiles	ρ: MAP and 90 % band	estimation	posterior	8.3.2
What is the coincidence factor at the design population?	a derived quantity of the state	99th-percentile peak over the sum of the racks' 99th percentiles	estimation	posterior	8.3.3
How much does the covariance move the answer, and how much the capacity band?	conditional probabilities; the spread of a probability	$\Pr(S^* > Q^\dagger \mid \rho, \mathbf{y})$ for $\rho = 0$ to 1; the spread of $\Pr(S^* > Q^\dagger \mid \boldsymbol{\theta}, \mathbf{z})$ with ρ or $Q^\dagger$ known	estimation	posterior	8.3.4

is a derived quantity of another state. The other three are estimation: the correlation, the coincidence factor, and what the answer to the first is made of.

8.1.4 The evidence as rows (step 2)

Step 2 attaches two bodies of evidence. The first is the rack record: sixty days of five readings each, three hundred readings, each a row of equation (8.2). The record is summarized in Figure 8.2. The left panel gives, for each rack, the sample mean of its sixty readings as a dot, one sample standard deviation either side as a bar, and its largest reading as a triangle. The right panel gives the hall's sum over the record: the sum of the five sample means, the largest daily sum, the record's coincident peak, and the sum of the five largest readings, the design sheet's starting point.

The sample means are 60.15, 54.01, 65.99, 59.63 and 51.47 kW, and the sample standard deviations 6.45, 4.95, 7.39, 6.43 and 5.75 kW. Each rack's largest reading lies about two to three of its standard deviations above its mean, rack B1's the nearest and rack B3's the furthest. The ten sample correlations between pairs of racks average 0.399 and run from 0.344 to 0.470: every pair moves together, by about the same amount, the pattern the one-factor model (8.1) asks for. The largest day of the record put 354.10 kW on the hall, against a sum of maxima of 365.76 kW, so the record's own coincidence factor is 0.9681, not 0.8: on the record's busiest day the racks were almost all at their largest together.

The priors. Weak priors complete the rows: each mean $\mu_k \sim \mathcal{N}(60,\ 30^2)$ kW, each log-spread $\ln s_k \sim \mathcal{N}(\ln 6,\ 0.5^2)$, and $\operatorname{atanh}\rho \sim \mathcal{N}(0,\ 0.5^2)$, which centers the correlation on zero and leaves it free to take any value the record supports.

The capacity's evidence. The second body of evidence is that of Chapter 5: the eight readings of hall B in window 1 of the night record and the nine priors on its coils and its racks' recirculation, which fit the hall's state $\mathbf{z}$. This chapter takes that fit as it stands, the one fit of the hall.

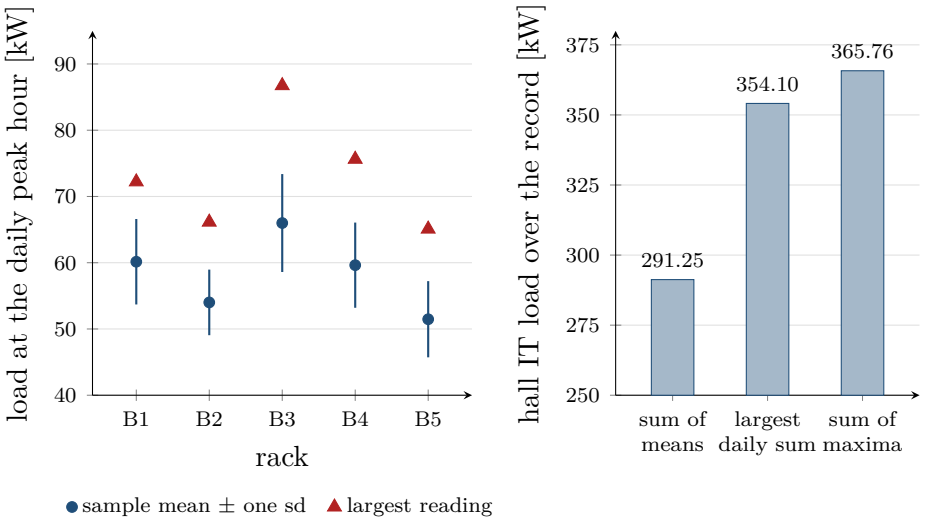


Figure 8.2: The rack record of hall B, sixty days at the daily peak hour, at today's population. Left: each rack's sample mean (dot), one sample standard deviation either side (bar) and largest reading (triangle). Right: the sum of the five sample means; the largest daily sum of the five racks, the record's coincident peak; and the sum of the five largest readings. The ratio of the last two, $354.10/365.76 = 0.9681$, is the record's own coincidence factor.

Which row acts on which entry. Each day's five readings inform all eleven entries of $\boldsymbol{\theta}$ together: the means through where the readings lie, the spreads through how far they scatter, and the correlation through how they scatter together. No reading of the rack record touches $\mathbf{z}$, and no reading of the hall touches $\boldsymbol{\theta}$.

Modeling choices at the rows. The meters are unbiased and their errors independent of each other and of the loads, with the stated 0.5 kW: this puts the meters' share on the diagonal of the readings' covariance only, where it cannot be taken for diversity, and it fails if the meters share an error, which the fit would read as correlation. The reading at the daily peak hour stands for the day: this makes one row block per day, and it fails if the racks' own peaks fall at other hours, a diversity in time that this record cannot see. And the hall's capacity is the fit of window 1, the healthy plant, as in Chapter 5.

8.2 The posterior of the two states (step 3)

The question of step 1 owes a probability over two states, and this section derives what each state must supply and computes it. The derivation starts from the probability and works down to the rows of the rack record and the curvature of its objective; the hall's posterior is Chapter 5's.

8.2.1 What the answer needs from the two states

The probability that the coincident peak exceeds the capacity is the probability of the event given the two states, averaged over their posterior, the law of total probability of equation (3.2),

$$\Pr(S^* > Q^\dagger \mid \mathbf{y}) = \iint \pi(\boldsymbol{\theta}, \mathbf{z})\, p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{y})\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\mathbf{z}, \qquad \pi(\boldsymbol{\theta}, \mathbf{z}) = \Pr(S^* > Q^\dagger(\mathbf{z}) \mid \boldsymbol{\theta}, \mathbf{z}), \tag{8.5}$$

with $\pi(\boldsymbol{\theta}, \mathbf{z})$ the exceedance probability given the states: the day-to-day variation of the load, aleatory, with every entry known. The two bodies of evidence share no entry, so the posterior

factors,

$$p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{y}) = p(\boldsymbol{\theta} \mid \mathbf{y}_{\mathrm{R}})\; p(\mathbf{z} \mid \mathbf{y}_{\mathrm{H}}), \tag{8.6}$$

with $\mathbf{y}_{\mathrm{R}}$ the rack record and $\mathbf{y}_{\mathrm{H}}$ the hall's readings. Given $\boldsymbol{\theta}$, the coincident peak (8.4) is a fixed multiple of a sum of normal loads, so it is normal, and the exceedance probability is its upper tail at the capacity,

$$\pi(\boldsymbol{\theta}, \mathbf{z}) = 1 - \Phi\!\left(\frac{Q^{\dagger}(\mathbf{z}) - m(\boldsymbol{\theta})}{s_S(\boldsymbol{\theta})}\right), \tag{8.7}$$

with Φ the standard normal distribution function and m and s_S the mean and the standard deviation of S^* given $\boldsymbol{\theta}$. The mean of a sum is the sum of the means,

$$m(\boldsymbol{\theta}) = f\, \mathbf{1}^{\mathsf{T}} \boldsymbol{\mu} = f \sum_k \mu_k, \tag{8.8}$$

with $\boldsymbol{\mu}$ the vector of the five means. The variance of the sum is the quadratic form of equation (4.53) with the gradient $f\,\mathbf{1}$, exact because the sum is linear, the five-entry form of equation (3.35),

$$s_S^2(\boldsymbol{\theta}) = f^2\, \mathbf{1}^{\mathsf{T}} \mathbf{C}(\boldsymbol{\theta})\, \mathbf{1} = f^2 \Big(\sum_k C_{kk} + \sum_{k \neq j} C_{kj} \Big), \tag{8.9}$$

with $\mathbf{C}(\boldsymbol{\theta})$ the covariance of the five rack loads of one day, in kW^2. The diagonal sum is the racks' own variances; the off-diagonal sum, twenty terms, is the diversity. The covariance follows from the model (8.1). A rack's load varies through the shared factor and its own variation, each of unit variance, with weights whose squares add to one, so $C_{kk} = s_k^2(\rho + 1 - \rho) = s_k^2$; two racks share only the factor, so $C_{kj} = s_k s_j \sqrt{\rho}\sqrt{\rho} = \rho\, s_k s_j$,

$$C_{kk} = s_k^2, \qquad C_{kj} = \rho\, s_k\, s_j \quad (k \neq j), \tag{8.10}$$

and the off-diagonal sum is ρ times the square of the sum of the spreads less the sum of their squares,

$$\mathbf{1}^{\mathsf{T}} \mathbf{C}\, \mathbf{1} = \sum_k s_k^2 + \rho \Big[\Big(\sum_k s_k \Big)^2 - \sum_k s_k^2 \Big]. \tag{8.11}$$

The two ends of the correlation are the two cases of Section 8.1.1. At $\rho = 0$ the variances add, and the spread of the sum is the root of the sum of the squared spreads; at $\rho = 1$ the spreads themselves add, and the sum varies as one rack $\sum_k s_k$ wide. The correlation interpolates linearly in the variance between them. The form is a covariance for any ρ above $-1/4$, the bound at which five equally correlated loads can still vary; the record puts the correlation far above it.

Equations (8.7)–(8.11) are what the answer needs from $\boldsymbol{\theta}$: its means, spreads and correlation. What it needs from $\mathbf{z}$ is the binding load $Q^{\dagger}(\mathbf{z})$. The remaining two subsections compute the two posteriors.

8.2.2 The record's rows and the curvature

The posterior of $\boldsymbol{\theta}$ is the product of the rows of the record and the priors, and its negative logarithm is the objective of equation (4.39), one term per row. By equations (8.1) and (8.2) a day's five readings $\mathbf{y}_d$ are jointly normal about the means, with the loads' covariance plus the meters',

$$\mathbf{y}_d \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu},\ \mathbf{C}_y(\boldsymbol{\theta})), \qquad \mathbf{C}_y(\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta}) + \sigma_m^2\, \mathbf{I}, \tag{8.12}$$

and the day's term of the objective is its negative log-density, up to a constant,

$$\mathcal{L}_d(\boldsymbol{\theta}) = \tfrac{1}{2} \ln \det \mathbf{C}_y(\boldsymbol{\theta}) + \tfrac{1}{2}\, (\mathbf{y}_d - \boldsymbol{\mu})^{\mathsf{T}} \mathbf{C}_y(\boldsymbol{\theta})^{-1} (\mathbf{y}_d - \boldsymbol{\mu}). \tag{8.13}$$

Table 8.2: The MAP state of the rack record of hall B, sixty days, with the posterior spread of each entry from the curvature (8.15) and the stated truth of the record.

Entry [unit]	MAP [entry's]	$\sqrt{\Sigma_{\theta,ii}}$ [entry's]	Truth [entry's]	Natural scale at the MAP [kW or –]
μ_1 [kW]	60.1565	0.8243	60.0000	60.16
μ_2 [kW]	54.0140	0.6328	55.0000	54.01
μ_3 [kW]	65.9903	0.9422	66.0000	65.99
μ_4 [kW]	59.6355	0.8071	60.0000	59.64
μ_5 [kW]	51.4731	0.7380	52.0000	51.47
$\ln s_1$ [–]	1.8514	0.0883	1.7918	$s_1 = 6.37$
$\ln s_2$ [–]	1.5849	0.0889	1.6094	$s_2 = 4.88$
$\ln s_3$ [–]	1.9860	0.0880	1.9459	$s_3 = 7.29$
$\ln s_4$ [–]	1.8303	0.0872	1.7918	$s_4 = 6.24$
$\ln s_5$ [–]	1.7401	0.0888	1.6094	$s_5 = 5.70$
$\operatorname{atanh}\rho$ [–]	0.4165	0.0734	0.3654	$\rho = 0.3940$

The quadratic term is the squared length of the whitened residual $\mathbf{C}_y^{-1/2}(\mathbf{y}_d - \boldsymbol{\mu})$, a residual vector of five rows as in every chapter so far. The determinant term is new: the covariance of the row is itself a function of the state, and the term charges a covariance for its size, so that the fit cannot explain the scatter of the readings away by making every spread large. With the prior rows of step 2 the objective is

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{d=1}^{60} \mathcal{L}_d(\boldsymbol{\theta}) + \tfrac{1}{2}\sum_k \Big(\frac{\mu_k - 60}{30}\Big)^2 + \tfrac{1}{2}\sum_k \Big(\frac{\ln s_k - \ln 6}{0.5}\Big)^2 + \tfrac{1}{2}\Big(\frac{\operatorname{atanh}\rho}{0.5}\Big)^2, \tag{8.14}$$

sixty five-row blocks and eleven prior rows on eleven entries. The MAP state $\hat{\boldsymbol{\theta}}$ minimizes it. Because the rows' covariance moves with the state, the Hessian of a day's term is not the product of a Jacobian line with itself, and the Gauss–Newton form of Chapter 5 would drop a term that is not small here. The information is the curvature of the objective itself, equation (4.40),

$$\boldsymbol{\Lambda}_\theta = \mathbf{H}(\hat{\boldsymbol{\theta}}) = \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathsf{T}}}\bigg|_{\hat{\boldsymbol{\theta}}}, \qquad \boldsymbol{\Sigma}_\theta = \boldsymbol{\Lambda}_\theta^{-1}, \qquad \boldsymbol{\theta} \mid \mathbf{y}_{\mathrm{R}} \approx \mathcal{N}(\hat{\boldsymbol{\theta}},\ \boldsymbol{\Sigma}_\theta), \tag{8.15}$$

the Laplace normal of equation (4.38) with the full Hessian. The minimum is found by a quasi-Newton iteration from the sample means, the logarithms of the sample standard deviations and $\rho = 0$, and the Hessian by central differences of the objective at the minimum.

8.2.3 The fit of the sixty days

Table 8.2 gives the MAP state with the posterior spread $\sqrt{\Sigma_{\theta,ii}}$ of each entry and the stated truth from which the record was generated, a check marker for Section 8.4.2; each row is an entry of equation (8.3), and the last column gives the entry on its natural scale at the MAP.

The means are known to under 1 kW each, as sixty days of a 5 to 7 kW spread allow. The log-spreads carry a posterior spread of about 0.09, a spread known to about nine percent, the same for every rack because each is read from the same sixty days. The fitted spreads lie a little below the sample standard deviations of step 2, since the fit removes the meters' share and divides by the number of days rather than one less. The correlation entry has a spread of 0.0734 on the inverse-hyperbolic-tangent scale; its MAP value carries the correlation 0.3940, beside the sample correlations' average of 0.399. The sum of the five MAP means and the spreads give the moments of the coincident peak at the MAP state by equations (8.8)–(8.11): at the design population, mean 330.11 kW and standard deviation 24.83 kW.

8.2.4 The capacity's posterior

The posterior of the hall's state is that of Chapter 5, the Laplace normal (5.13) and its 50 000 draws pushed through the binding loads of all five racks, the joint draws (5.17). This chapter uses those same draws. The binding load of hall B has the median 395.21 kW and the 90 % band [345.89, 450.43] kW, as in equation (5.21), and its value at the MAP state is 396.46 kW, rack B3's.

8.2.5 The exact posterior of both states

The exact posterior of $(\boldsymbol{\theta}, \mathbf{z})$ is computed beside the Laplace normals by reweighting, as in Section 5.2.3. Draw $\boldsymbol{\theta}_i$ from $\mathcal{N}(\hat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_\theta)$ and pair it with the ith draw $\mathbf{z}_i$ of Chapter 5; by the factorization (8.6) the weight of the pair is the product of the two weights,

$$w_i \propto \exp\Big[-\mathcal{L}(\boldsymbol{\theta}_i) + \mathcal{L}(\hat{\boldsymbol{\theta}}) + \tfrac{1}{2}\,(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}})^{\mathsf{T}} \boldsymbol{\Sigma}_\theta^{-1} (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}})\Big]\; w_{\mathbf{z},i}, \qquad \sum_i w_i = 1, \tag{8.16}$$

with $\mathcal{L}$ the objective (8.14) and $w_{\mathbf{z},i}$ the weight of equation (5.16) of the hall's draw. With $N = 50\,000$ pairs the effective sample size is 30 798. The hall's draws alone had 49 638 in Chapter 5, so most of the loss is the rack record's: the posterior of the spreads and the correlation departs from its Laplace normal more than the hall's state does, as a posterior on a variance does when sixty days fix it.

8.3 The answers (step 4)

Every answer is read from the two posteriors of Section 8.2. The reported value is the state route: paired draws of both states, each carried through the exact exceedance probability (8.7). The exact posterior stands beside it with the gap, and the single-value routes beside both.

8.3.1 The probability that the coincident peak exceeds the capacity

Planner's question. The racks never all peak at once; what is the probability the coincident peak exceeds the plant's capacity? The question asks for the probability that, on a day at the design population, the peak-hour load of the five racks together exceeds the load at which hall B's first inlet meets its limit.

General form. A future value against a requirement is a tail of the predictive distribution, equation (2.13), and the probability of the event is the posterior average of its probability given the state, as in equation (2.12),

$$\Pr(y_{\text{new}} > f(\mathbf{z}) \mid \mathbf{y}) = \int \Pr(y_{\text{new}} > f(\mathbf{z}) \mid \mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

with y_{new} a future value, random even when the state is known, and $f(\mathbf{z})$ a derived quantity of the state.

Specialized form. The question is equation (2.13) with the future value the coincident peak S^* of equation (8.4), the derived quantity the binding load $Q^\dagger(\mathbf{z})$ of equation (5.8), the state the pair $(\boldsymbol{\theta}, \mathbf{z})$, and the probability given the state the exceedance probability (8.7): it is equation (8.5). The reported answer integrates the two Laplace normals by paired draws,

$$\hat{P} = \frac{1}{N} \sum_{i=1}^{N} \pi(\boldsymbol{\theta}_i, \mathbf{z}_i), \qquad \mathrm{se}(\hat{P}) = \frac{\mathrm{sd}_i\, \pi(\boldsymbol{\theta}_i, \mathbf{z}_i)}{\sqrt{N}}, \tag{8.17}$$

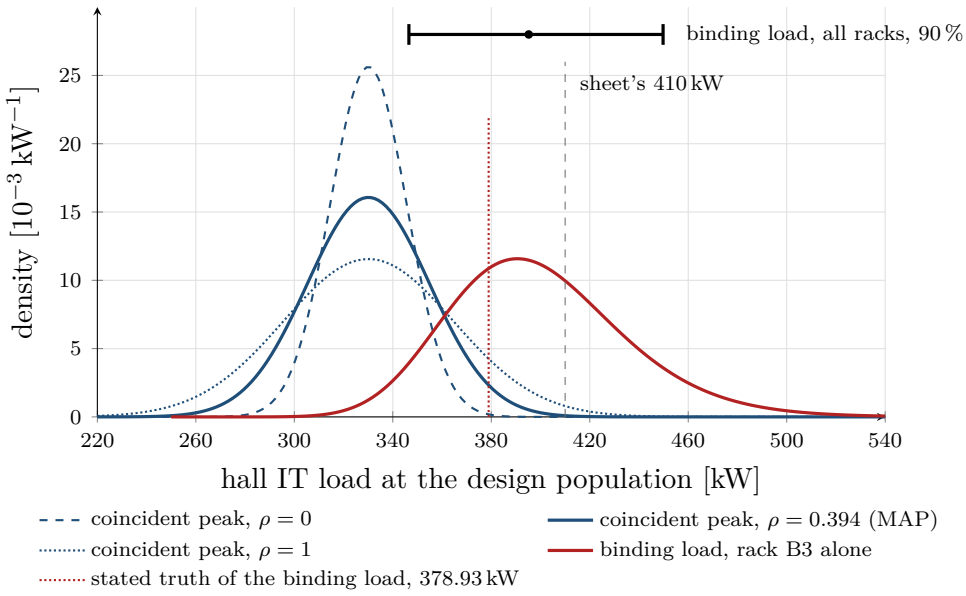


Figure 8.3: The coincident peak of hall B at its design population against its capacity. Blue: the normal density of S^* at the MAP state of the rack record, mean 330.11 kW, with standard deviation 15.57, 24.83 and 34.53 kW for $\rho = 0$, the MAP 0.394 and 1 (Table 8.4). Red: the posterior density of rack B3's binding load by the inversion route of equation (5.20), $a_{B3} \mid \mathbf{y} \approx \mathcal{N}(0.031983,\ 0.002751^2)$ K/kW. Bar: the 90 % band of the binding load over all five racks, [345.89, 450.43] kW, with its median; the probabilities are computed with these joint draws, not with the red curve. Dashed: the design sheet's capacity. Dotted: the stated truth of the binding load, a check marker for Section 8.4.2.

with $\boldsymbol{\theta}_i \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_\theta)$, $\mathbf{z}_i$ the joint draws of equation (5.17), each pushed through the binding loads of all five racks, and se the Monte Carlo standard error. The day's variation of the load is integrated in closed form inside π; the draws carry only what is not known about the two states. The exact posterior is $\sum_i w_i\, \pi(\boldsymbol{\theta}_i, \mathbf{z}_i)$ with the weights (8.16). Four routes stand beside these. The plug-in reads $\pi(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$, every entry at its MAP value. Two routes pin the correlation, replacing ρ by 0 (the racks independent, full diversity) or by 1 (the racks peaking together, no diversity) in every draw and averaging as in equation (8.17). And the sheet's route reads π at the MAP $\hat{\boldsymbol{\theta}}$ against the sheet's capacity of 410 kW taken as exact.

Building blocks. The rack record's MAP state and covariance (Section 8.2.3); the hall's joint draws and weights (Section 8.2.4); the moments (8.8) and (8.11); the exceedance probability (8.7); the paired draws (8.17); the weights (8.16).

The coincident peak against the capacity. The two sides of the comparison are drawn in Figure 8.3. The horizontal axis is the hall IT load, the vertical axis a density; the three bell curves are the coincident peak at the MAP state, with ρ at 0, at its MAP value and at 1; the right-skewed curve is the binding load of rack B3 alone by Chapter 5's inversion route (5.20), the capacity's shape; and the bar above the curves is the reported 90 % band of the binding load over all racks; the dashed line is the sheet's capacity and the dotted line the stated truth of the binding load, a check marker for Section 8.4.2. The probability asked for is how often a draw from the load side lands above a draw from the capacity side.

The bell curves sit well to the left of the capacity: the coincident peak's mean, 330.11 kW, is 65 kW below the capacity's median. The overlap is in two tails at once, the load's upper tail

Table 8.3: The probability that the coincident peak of hall B at its design population exceeds its capacity, by six routes. The gaps are the route less the reported value.

Route	$\Pr(S^* > Q^\dagger)$ [–]	Gap from reported [–]
Paired draws of both states (reported), se 0.0005	0.0481	–
Exact posterior, the same pairs reweighted	0.0508	+0.0027
Plug-in, $\hat{\boldsymbol{\theta}}$ and $Q^\dagger(\hat{\mathbf{z}}) = 396.46\,\text{kW}$	0.0038	−0.0443
$\rho = 0$ in every draw, independent racks (full diversity)	0.0263	−0.0218
$\rho = 1$ in every draw, racks peak together (no diversity)	0.0784	+0.0303
Sheet's capacity 410 kW taken as exact, $\hat{\boldsymbol{\theta}}$	0.0006	−0.0475

and the capacity's lower tail, and the correlation decides how far the load's tail reaches. With ρ at 1 the load's curve is more than twice as wide as with ρ at 0.

By equation (8.17), with $N = 50\,000$ paired draws, the reported probability is

$$\Pr(S^* > Q^\dagger \mid \mathbf{y}) = 0.0481, \qquad \text{se} = 0.0005, \tag{8.18}$$

and the exact posterior, the same pairs reweighted, gives 0.0508, a gap of +0.0027, exact minus reported. The gap is more than five times the Monte Carlo standard error of the draws, so it is not sampling noise between the two numbers; it is the departure of the rack record's posterior from its Laplace normal, carried into a tail probability (Section 8.4.1). Both are stated, and neither is chosen. Table 8.3 sets the routes side by side: each row is a route, the columns its probability and its gap from the reported value.

The plug-in is thirteen times too small. It sets the capacity at its MAP value, 396.46 kW, more than two and a half of the load's standard deviations above the load's mean, where the normal tail is thin. But the capacity is not known to that precision, and the probability of equation (8.5) averages π over the capacity's band. Across the paired draws the exceedance probability given the states has its 5 %, 50 % and 95 % points at 0.0000, 0.0045 and 0.2668: in most draws the capacity is high and the exceedance negligible, in a few it is low and the exceedance large, and the average, 0.0481, is set by those few. The plug-in sits near the median draw, 0.0038 against 0.0045, and misses the tail that makes the answer. The sheet's route misses it twice, with the capacity taken as exact and at a value above the capacity's median: 0.0006.

8.3.2 The correlation of the racks

Planner's question. How correlated are the racks? The question asks for the correlation ρ of the one-factor model (8.1), with its band.

General form. An estimation question about an entry asks for its posterior distribution, equation (2.9), summarized by a location and a central credible interval, equation (2.11), here the central 90 % interval.

Specialized form. The question is equation (2.9) for the entry $\alpha = \operatorname{atanh}\rho$ of the state (8.3) under the Laplace normal (8.15), read on the correlation's own scale. The correlation is an increasing function of the entry, $\rho = \tanh\alpha$, so its quantiles are the images of the entry's, as Chapter 2 states for a monotone function: the location is $\tanh\hat{\alpha}$ and the band is

$$\left[\ \tanh(\hat{\alpha} - 1.645\,s_\alpha),\ \tanh(\hat{\alpha} + 1.645\,s_\alpha)\ \right], \tag{8.19}$$

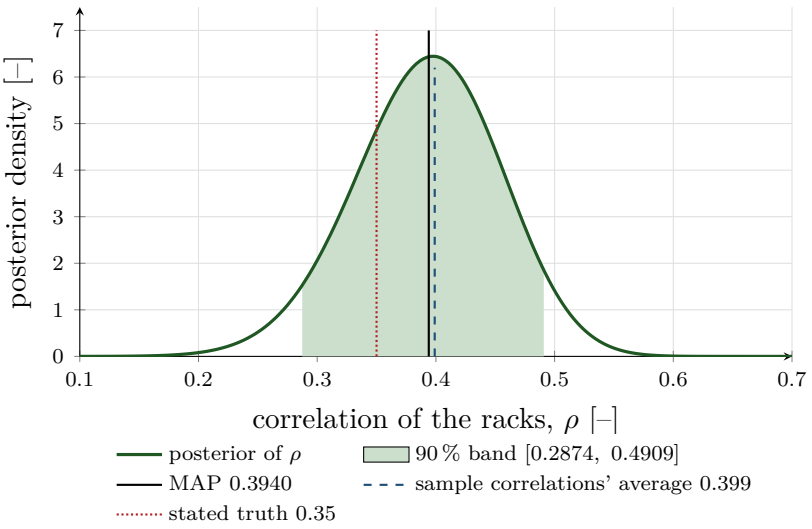


Figure 8.4: The posterior of the correlation of hall B's racks, from the Laplace normal on $\alpha = \operatorname{atanh}\rho$, $\mathcal{N}(0.4165,\ 0.0734^2)$, carried to ρ by the change of variables with the factor $1/(1-\rho^2)$. Shaded: the 90 % band of equation (8.19). Lines: the MAP value, the average of the ten sample correlations, and the stated truth of the record.

the inversion route of Chapter 5 applied to an entry, with $s_\alpha = \sqrt{\Sigma_{\theta,\alpha\alpha}}$. The density of the correlation follows from the normal on α by the change of variables (4.55), with $\mathrm{d}\alpha/\mathrm{d}\rho = 1/(1-\rho^2)$.

Building blocks. The MAP entry $\hat{\alpha}$ and its spread s_α (Table 8.2); the band (8.19).

The correlation. With $\hat{\alpha} = 0.4165$ and $s_\alpha = 0.0734$, the correlation of the racks is

$$\rho = 0.3940, \qquad \rho \in [0.2874,\ 0.4909] \text{ with } 90\,\% \text{ probability.}$$

Its posterior density is drawn in Figure 8.4: the horizontal axis is the correlation, the vertical axis its density; the shaded area is the 90 % band; the solid line the MAP value, the dashed line the sample correlations' average, and the dotted line the stated truth, a check marker for Section 8.4.2.

The racks are correlated, and not weakly: the band excludes zero by a wide margin and stops well short of one. The density leans slightly to the left, the trace of the change of variables, so the band reaches 0.107 below the MAP and 0.097 above it. The sixty days fix the correlation to about a quarter of its value either way. That uncertainty is what the conditional answers of Section 8.3.4 vary.

8.3.3 The coincidence factor at the design population

Planner's question. What is the coincidence factor at the design population? The question asks for the number the design sheet fixes at 0.8, read from the fitted record instead.

General form. A derived quantity of the state is estimated by its posterior distribution, equation (2.9), located at its value at the MAP state.

Table 8.4: The coincidence factor (8.20) of hall B at its design population, at the MAP state of the rack record with the correlation set to three values; the mean of the coincident peak is 330.11 kW in every row.

Correlation	sd of S^* [kW]	99th-percentile peak [kW]	Sum of the racks' 99th percentiles [kW]	Coincidence factor [–]
$\rho = 0$, independent racks	15.57	366.33	410.42	0.8926
$\rho = 0.3940$, the MAP	24.83	387.87	410.42	0.9450
$\rho = 1$, racks peak together	34.53	410.42	410.42	1.0000
The record, largest daily sum over sum of maxima	–	–	–	0.9681
The design sheet's diversity factor	–	–	–	0.8

Specialized form. The question is equation (2.9) with $f(\boldsymbol{\theta})$ the coincidence factor at the 99th percentile: the 99th percentile of the coincident peak over the sum of the racks' own 99th percentiles, both at the design population. Given $\boldsymbol{\theta}$ every quantity is normal, and with $\Phi(2.326) = 0.99$,

$$\mathrm{CF}(\boldsymbol{\theta}) = \frac{m(\boldsymbol{\theta}) + 2.326\, s_S(\boldsymbol{\theta})}{f \sum_k (\mu_k + 2.326\, s_k)}, \tag{8.20}$$

with m and s_S of equations (8.8) and (8.9). The numerator depends on the correlation through s_S and the denominator does not; at $\rho = 1$, $s_S = f \sum_k s_k$ by equation (8.11), and the factor is exactly one. The factor is read at the MAP state, and at the two ends of the correlation, which bracket it for every value of ρ because it increases with ρ.

Building blocks. The MAP state (Table 8.2); the moments (8.8)–(8.11); the factor (8.20).

The factor. Table 8.4 gives it. Each row is a value of the correlation, with the other entries at the MAP state; the columns give the standard deviation of the coincident peak, its 99th percentile, the sum of the racks' 99th percentiles and their ratio. Two rows below them give the record's own factor, its largest daily sum over its sum of maxima, and the sheet's.

At the MAP state the coincidence factor is 0.945, and it lies between 0.893 for independent racks and 1 for racks that peak together. The record's own factor, 0.968, is of the same kind though not the same quantity, a largest day over largest readings rather than a ratio of percentiles. The customary 0.8 is below every one of them, below even the factor of five independent racks: a rule that would hold only for loads that move against each other, more diverse than independent racks. Applied to racks driven by one workload, it is optimistic. Its result says so in its own numbers: the sheet's diversified peak of 331.62 kW is within 2 kW of the coincident peak's mean, 330.11 kW, so the 0.8 turned the sum of the racks' largest days into the hall's average day.

8.3.4 What the answer is made of

Planner's question. How much does the covariance move the answer, and how much the capacity band? The question asks how the probability of Section 8.3.1 changes with the correlation, and which of the two states the uncertainty in that probability comes from.

General form. A conditional probability restricts the posterior to the outcomes with a given value of an entry, equation (3.4); the spread of a derived quantity over the posterior is its

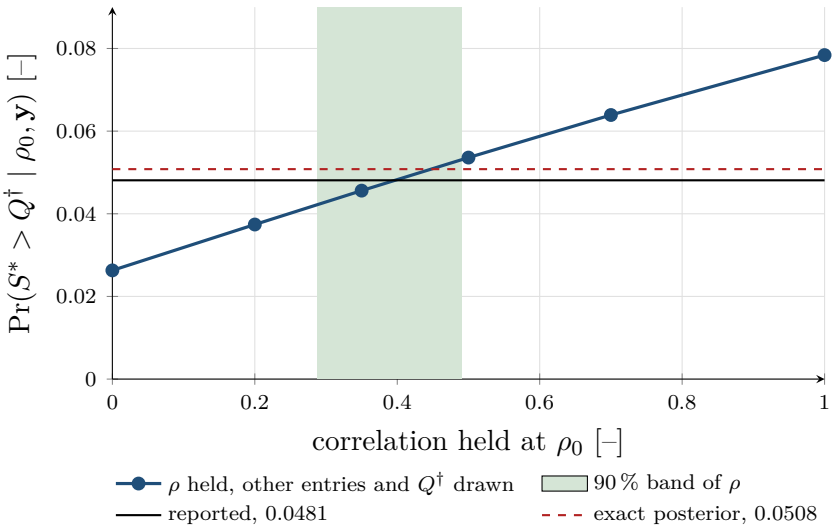


Figure 8.5: The probability that the coincident peak of hall B exceeds its capacity with the correlation of the racks held at ρ_0, equation (8.21), the means, spreads and binding load drawn from their posteriors. Shaded: the 90 % band of ρ, [0.2874, 0.4909]. Lines: the reported and the exact answers, which draw ρ as well.

variance, equation (2.10),

$$\mathrm{Var}(f(\mathbf{z}) \mid \mathbf{y}) = \int \big(f(\mathbf{z}) - \mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}]\big)^2 \, p(\mathbf{z} \mid \mathbf{y}) \, \mathrm{d}\mathbf{z},$$

recomputed with one entry held at a value to see the share of that entry.

Specialized form. The conditional probability is equation (3.4) applied to equation (8.5) with ρ held at a value ρ_0, read by the paired draws with the correlation entry of every $\boldsymbol{\theta}_i$ replaced,

$$\Pr(S^* > Q^\dagger \mid \rho_0, \mathbf{y}) \approx \frac{1}{N} \sum_{i=1}^{N} \pi(\boldsymbol{\theta}_i|_{\rho=\rho_0}, \ \mathbf{z}_i). \tag{8.21}$$

The spread is equation (2.10) with f the exceedance probability $\pi(\boldsymbol{\theta}, \mathbf{z})$ of equation (8.7): its standard deviation over the paired draws, the epistemic spread of the answer. It is computed three times: with every entry drawn; with ρ held at its MAP value in every draw; and with the binding load held at its median, 395.21 kW, in every draw,

$$\mathrm{sd}_i \, \pi(\boldsymbol{\theta}_i, \mathbf{z}_i), \qquad \mathrm{sd}_i \, \pi(\boldsymbol{\theta}_i|_{\rho=\hat{\rho}}, \ \mathbf{z}_i), \qquad \mathrm{sd}_i \, \pi(\boldsymbol{\theta}_i, \ Q^\dagger = 395.21 \,\mathrm{kW}). \tag{8.22}$$

What a held entry takes off the spread is that entry's part in it.

Building blocks. The paired draws (8.17); the exceedance probability (8.7); the conditional draws (8.21); the spreads (8.22).

The level. The conditional probability is drawn against the correlation in Figure 8.5: the horizontal axis is the value at which ρ is held, the vertical axis the probability; the dots are the six values computed, joined by a line; the shaded strip is the correlation's 90 % band; the two horizontal lines are the reported and the exact answers, which average over the correlation.

The probability rises steadily with the correlation, from 0.0263 for independent racks to 0.0784 for racks that peak together, a factor of three. The covariance sets the level of the

Table 8.5: The epistemic spread of the exceedance probability $\pi(\boldsymbol{\theta}, \mathbf{z})$ of hall B, its standard deviation over the 50 000 paired draws, equation (8.22), with one entry held at a time.

Held	sd of π [–]	Remaining of the full [–]
Nothing, every entry drawn	0.1045	1.00
The correlation ρ at its MAP 0.3940	0.1043	1.00
The binding load $Q^{\dagger}$ at its median 395.21 kW	0.0039	0.04

answer: the same racks, with the same means and spreads, against the same capacity, carry one probability or three times it depending on how they move together. Across the correlation's band the line runs from about 0.042 to about 0.053, read between the computed points, and the reported answer sits inside that range, as an average over the band must.

The spread. Table 8.5 gives the spread of the exceedance probability. Each row holds a different entry; the columns give the standard deviation of π over the paired draws and the share of the full spread that remains.

Holding the correlation leaves the spread of the answer where it was, 0.1043 against 0.1045; holding the capacity takes it to 0.0039, four hundredths of itself. The two halves of the question have two different answers. The covariance between the racks is the whole of the level: it moves the probability by a factor of three, and nothing else in the rack record moves it as much. But the confidence in the probability is the capacity's: its spread over the posterior, wide enough that the 95 % point of π is 0.2668, comes almost entirely from the band of the binding load, which is the band of rack B3's unmeasured airflow of Chapter 5. The rack record, with sixty days behind it, is the well-known half of the comparison.

8.4 The checks (step 5)

The answers of Section 8.3 rest on two Laplace normals, on draws, and on a record generated from a stated truth; each is tested here against the calculation it stands for.

8.4.1 The draws, the reweighting and their gap

The reported probability, 0.0481, is an average of 50 000 values of π whose standard deviation is 0.1045, so its Monte Carlo standard error is $0.1045/\sqrt{50\,000} = 0.0005$, the value printed beside it. The exact posterior, 0.0508, uses the same pairs under the weights (8.16), and its gap from the reported value, +0.0027, is more than five times that standard error. The two use the same draws, so the gap is not sampling noise between independent sets. It is the difference between the Laplace normals and the exact posteriors, and the effective sample size, 30 798 of 50 000, against 49 638 for the hall's draws alone, places most of it in the rack record: a posterior on spreads and a correlation, sixty days wide, is not quite normal even on the logarithmic and hyperbolic scales. The exact value has a Monte Carlo error of its own, larger than the draws' for the smaller effective sample, which this chapter does not compute.

The gap matters only if it moves a decision, and here it can. The policy of Chapter 5 asks that a requirement hold with probability at least 0.95. Read as the requirement that the coincident peak stay within the capacity, the reported route gives $1 - 0.0481 = 0.9519$, just above the level, and the exact posterior $1 - 0.0508 = 0.9492$, just below it. The two routes are stated as they stand, and neither is chosen over the other by a rule. Against the design sheet's

Table 8.6: The checks of the coincident-peak answers of hall B. The truth of the capacity is that of Chapter 5's record; the truth of the exceedance probability is π of equation (8.7) at the true rack state and the true binding load.

Quantity and route	Value [– or kW]	Truth [– or kW]	Reading
$\operatorname{atanh}\rho$, MAP (sd 0.0734)	0.4165	0.3654	0.70 of its sd above the truth
Every entry of $\boldsymbol{\theta}$, MAP	Table 8.2	Table 8.2	each within two of its sds of the truth
Binding load, joint draws, median	395.21	378.93	the truth inside the band [345.89, 450.43]
$\Pr(S^* > Q^\dagger)$, paired draws (reported)	0.0481	0.0200	the truth inside the 90 % band of π, [0.0000, 0.2668]
$\Pr(S^* > Q^\dagger)$, exact posterior	0.0508	0.0200	as above
Gap, exact minus reported	+0.0027	–	over five Monte Carlo standard errors
Monte Carlo standard error, reported	0.0005	–	$\mathrm{sd}(\pi)/\sqrt{N}$, $N = 50\,000$
Effective sample size, both states	30 798	–	of 50 000; the hall alone 49 638
Plug-in	0.0038	0.0200	below the truth
Sheet's capacity, $\hat{\boldsymbol{\theta}}$	0.0006	0.0200	below the truth

claim, a pass with 78 kW to spare and an exceedance indistinguishable from zero, both say the same thing.

8.4.2 The answers against the stated truth

The rack record is generated from a stated truth: means (60, 55, 66, 60, 52) kW, spreads (6, 5, 7, 6, 5) kW, correlation 0.35 for every pair, meter error 0.5 kW. Table 8.6 collects the checks: each row is a quantity, the columns its value by the route named, the stated truth, and what the comparison shows.

The correlation's MAP lies 0.70 of its spread above the truth, and every entry of the rack state lies within two of its spreads of its truth; the largest departures are the mean of rack B2, below its truth, and the log-spread of rack B5, above it. The true binding load, 378.93 kW, lies inside the reported band. At the true rack state and the true binding load the exceedance probability is 0.0200, lower than the reported 0.0481 and well inside the 90 % band of π over the posterior. The comparison is the one the answer promises: the truth is one value of π, the answer is the average of π over what the evidence leaves unknown, and with the capacity known to a band some 100 kW wide that average lies above the median value of π and above the truth here. The plug-in and the sheet's route both fall below the truth, the first because it ignores the capacity's band and the second because it also takes the capacity too high.

8.5 The report (step 6)

At its design population the coincident peak of hall B exceeds the hall's capacity on a given day with probability 0.05: 0.0481 by the paired draws and 0.0508 by the exact posterior. That is not the zero a 0.8 diversity factor promises, and not the 0.004 of a plug-in that sets the capacity at one value. The racks move together, with correlation 0.394; the coincidence factor at the design population is 0.945, not 0.8; and the confidence in the answer is the capacity's.

Table 8.7: The report for hall B at its design population of 340 kW, at the drawing's supply of 14.0°C with all three units running. Reported values by paired draws of the rack state and the hall's state.

Question	Answer	Band or routes beside
The racks never all peak at once; what is the probability the coincident peak exceeds the plant's capacity?	0.0481, Monte Carlo se 0.0005 (exact: 0.0508)	plug-in 0.0038; sheet's 410 kW as exact 0.0006; $\rho = 0$ 0.0263; $\rho = 1$ 0.0784
How correlated are the racks?	$\rho = 0.394$	0.287 to 0.491, central 90 %
What is the coincidence factor at the design population?	0.945 at the MAP state	0.893 ($\rho = 0$) to 1.000 ($\rho = 1$); record 0.968; sheet 0.8
How much does the covariance move the answer, and how much the capacity band?	the covariance sets the level: 0.026 to 0.078 as ρ goes from 0 to 1	the capacity sets the spread: sd of π 0.1045; 0.1043 with ρ known; 0.0039 with $Q^{\dagger}$ known

8.5.1 The answers and their bands

Table 8.7 gives the answers in the order of Table 8.1. Each row is a question; the second column gives the reported answer with the exact posterior beside it where it is computed, and the third its band or the routes beside it.

8.5.2 The design-review entry

The design review's entry replaces the sheet's pass. At the design population, on a given day, the coincident peak of the five racks exceeds the load at which hall B's first inlet meets 27°C with probability 0.0481 by the paired draws and 0.0508 by the exact posterior; read as the requirement, 0.9519 and 0.9492, on either side of the 0.95 the policy asks, and the entry records both without choosing. The entry records what the number rests on: a correlation of 0.394 between the racks, which the sixty-day record fixes well and which sets the level of the answer, and the band of the hall's binding load, which sets almost all of the answer's uncertainty and rests on the survey's airflow number for rack B3. The entry closes with the choice the review has: accept a daily exceedance probability of about 0.05 at the design population, or buy the probe at rack B3 before committing to it.

Common mistake

Applying a diversity factor from habit. The factor is a statement about the covariance of the loads; for racks that serve one workload it is near one, and 0.8 turns the sum of the racks' worst days into the hall's average day.

8.6 The next measurement (step 7)

The report leaves an answer whose level is fixed by the rack record and whose spread is almost all the capacity's. Two measurements are open to the review: more of the record, and a reading at the rack that sets the capacity. Each is judged by how much it narrows the spread of the exceedance probability, the quantity on which the decision sits.

Table 8.8: The two candidate measurements for the coincident-peak answer of hall B, ranked by the spread of the exceedance probability they leave. The spreads of π after are approximations from scaled draws of the current posterior, not refits after the measurement.

Candidate	Spread of its entry after	sd of π after [–]
None, today	–	0.1045
Sixty more days of rack records (one constructed continuation, refitted)	atanh ρ: 0.0734 to 0.0524 [–]	0.1036
Inlet probe at rack B3, 0.1 K (preposterior of Chapter 5)	$Q^{\dagger}$: 34.96 to 8.08 [kW]	0.0072

8.6.1 The two candidates

Sixty more days of rack records narrow the correlation. Unlike the reading rows of Chapter 5, the rows of the rack record carry the state in their covariance, so the information they add depends on their values; the chapter estimates it from one constructed continuation of the record, sixty days generated from the same truth and fitted with the first sixty. The posterior spread of atanh ρ falls from 0.0734 to 0.0524, about the factor $1/\sqrt{2}$ of a doubled record. The draws of the rack state are then shrunk about the MAP state, entry by entry, by the ratio of the spreads after and before, and the exceedance probability recomputed with the hall's draws unchanged. The probe at rack B3 is the measurement Chapter 5 recommended, an inlet probe of 0.1 K standard deviation, which takes the standard deviation of the binding load from 34.96 to 8.08 kW by its preposterior, Table 5.9; the draws of the binding load are shrunk about their median by the ratio 8.08/34.96, and the exceedance probability recomputed with the rack state's draws unchanged. Both are approximations from scaled draws, not refits of the posterior after the measurement: they keep each posterior's shape and narrow it.

Table 8.8 gives the result. Each row is a candidate, the columns the spread it leaves on its own entry and the spread of the exceedance probability after it.

The ranking follows the split of Section 8.3.4. Sixty more days of records, on the one constructed continuation, narrow the correlation by more than a quarter and leave the spread of the answer at about 0.1036 by the scaled draws, against 0.1045 today: they refine the half of the answer that is already known. The probe at rack B3 takes the spread to about 0.0072 by the scaled draws, a fourteenth of today's, because it narrows the half that is not.

8.6.2 The recommendation

The measurement to buy before committing hall B to its design population is the probe at rack B3, the same measurement that Chapter 5 recommended for the headroom. It is the measurement again because the question is the same capacity seen from the load side: whatever is asked of hall B near its design load, the band of the answer is the band of rack B3's airflow. The rack record does not need more days for this decision; it would need them to tell one workload from two (Section 8.6.4). Chapter 18 ranks the tests and surveys a planner can commission on the site plant in the same way, by the width each takes off the quantity that decides the build, with refits in place of scaled draws.

8.6.3 What the analysis showed

The design sheet passed the hall with 78 kW to spare, and the analysis puts the probability that the coincident peak exceeds the capacity on a given day at the design population at 0.05, 0.0481 by the draws and 0.0508 by the exact posterior. Three things on the sheet made the difference. Its diversity factor assumed loads more diverse than independent racks would be,

when the record shows racks that move together with correlation 0.394, a coincidence factor of 0.945; the 0.8 placed the design peak at the hall's average day. Its capacity was one number above the median of the capacity's band. And it compared two numbers where the answer is an average over two uncertain states: the plug-in, which takes both states at their best values, is thirteen times too small, because the answer is made in the draws where the capacity is low.

The surprise is the split. The drill question asserts that the covariance between racks is the whole answer. It is the whole of the level: held at zero or at one, the correlation moves the probability by a factor of three, and no other entry of the rack record matters as much. But it is almost none of the confidence: with the correlation known exactly the spread of the exceedance probability is unchanged, and with the capacity known it falls to four hundredths of itself. On this hall the coincident peak is the well-measured side of the comparison and the capacity the poorly measured one, and the measurement that decides the build is a probe at a rack's inlet, not a longer load record.

8.6.4 Limits of the model

A rack's own load and its own inlet. The racks enter the hall only through their sum, and an inlet rises with the hall's heat, not its own rack's. A rack whose inlet draws its own exhaust would have its inlet rise with its own load, so that the busiest rack would bind sooner than the hall's sum suggests; the model has no row that would carry it.

One workload, normal, day by day. One shared factor gives every pair the same correlation; racks serving two workloads would need two factors and a covariance with more entries. The loads are normal at the peak hour, and a workload with a heavier upper tail would raise the probability above the values here. The days are independent, and the probability is for one day: the probability that some day of a season exceeds the capacity is larger, and not by a simple power, because every day of the season shares the same uncertain capacity.

The design population by scaling. Every rack's mean and spread are scaled by 340/300 and the correlation kept. Load added as a different workload, or in new racks with their own airflow, would change both the covariance and the capacity. When the hall reaches its design population is the question of Chapter 6; this chapter asks only what happens once it does.

A pinned supply. The capacity is the binding load at the drawing's design-day supply of 14.0°C with all units running, as in Chapter 5, and with the other heat of an ordinary day; on the design day the envelope's gain adds to the other heat and, by equation (5.19), lowers the binding load by that gain (Chapter 7). A chilled-water plant that cannot hold its supply at the coincident peak would bind first; the central plant's capacity enters in Part III.

Two independent fits. The rack record and the hall's readings are fitted separately, and their posteriors multiply. A shared cause, a busy workload that also changed the airflow, would correlate them, and the factorization (8.6) would not hold.

The constructed record. The sixty days of rack readings are generated from a stated truth with the stated meter error, not taken from a hall, and the comparison with that truth in Section 8.4.2 is a check of the method on its own model, not of the model against a real hall.

The next part. Part III brings in the site's chilled-water plant, and Chapter 9 asks how much redundancy it needs.

8.7 What to retain

Definition 8.1 (Coincident peak and its exceedance)**.** The coincident peak of a group of loads is their sum, $S = \mathbf{1}^{\mathsf{T}}\mathbf{L}$; given the loads' means $\boldsymbol{\mu}$ and covariance $\mathbf{C}$, its mean is $\mathbf{1}^{\mathsf{T}}\boldsymbol{\mu}$ and its variance $\mathbf{1}^{\mathsf{T}}\mathbf{C}\,\mathbf{1}$, whose off-diagonal terms are the diversity; under one shared factor, $\mathbf{1}^{\mathsf{T}}\mathbf{C}\,\mathbf{1} = \sum_k s_k^2 + \rho[(\sum_k s_k)^2 - \sum_k s_k^2]$. The probability that it exceeds an uncertain capacity is the exceedance probability given both states, the load's tail at the capacity, averaged over the posterior of both, and it is reported by paired draws with the exact posterior beside it.

Four conclusions carry beyond hall B. A diversity factor is a statement about a covariance and is tested against a load record; for loads that serve one workload it is near one. The probability that a load exceeds an uncertain capacity is an average over the capacity's band, and a plug-in at the capacity's best value can understate it many times over, because the average is made in the band's low tail. The level and the confidence of such a probability can come from different places: holding each entry in turn shows which entry sets which. And when both sides of a comparison are uncertain, the measurement worth buying narrows the side that is less well known, which may not be the side the question is asked about.

Table 8.9 lists the terms of the chapter with their expressions, units and the subsection that defines each.

Table 8.9: The terms of load diversity and coincident peaks and where each is defined. Units for hall B.

Term	Expression	Units	Meaning	Defined in
One-factor model	$L_{kd} = \mu_k + s_k(\sqrt{\rho}F_d + \sqrt{1-\rho}\,e_{kd})$	kW	a shared workload and each rack's own	8.1.2
Rack state	$(\mu_{1..5}, \ln s_{1..5}, \operatorname{atanh}\rho)$	kW, –	the eleven entries of the record	8.1.2
Coincident peak	$S^* = f\,\mathbf{1}^{\mathsf{T}}\mathbf{L}$	kW	the hall's load at the design population	8.1.2
Rows and posterior				
Exceedance given the states	$\pi = 1 - \Phi\big((Q^\dagger - m)/s_S\big)$	–	the day's variation, aleatory	8.2.1
Variance of a sum	$s_S^2 = f^2\,\mathbf{1}^{\mathsf{T}}\mathbf{C}\,\mathbf{1}$	kW^2	diagonal plus diversity	8.2.1
One-factor covariance	$C_{kk} = s_k^2$, $C_{kj} = \rho s_k s_j$	kW^2	every pair correlated ρ	8.2.1
Day's row	$\frac{1}{2}\ln\det\mathbf{C}_y + \frac{1}{2}(\mathbf{y}_d - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{C}_y^{-1}(\mathbf{y}_d - \boldsymbol{\mu})$	–	a row whose covariance is fitted	8.2.2
Curvature as information	$\boldsymbol{\Sigma}_\theta = \mathbf{H}(\hat{\boldsymbol{\theta}})^{-1}$	mixed	full Hessian, not $\mathbf{J}^{\mathsf{T}}\mathbf{J}$	8.2.2
Paired reweighting	$w_i \propto w_{\theta,i}\,w_{\mathbf{z},i}$	–	exact posterior of both states	8.2.5
Questions and answers				
Paired draws	$\hat{P} = N^{-1}\sum_i \pi(\boldsymbol{\theta}_i, \mathbf{z}_i)$	–	the reported exceedance	8.3.1
Plug-in	$\pi(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$	–	ignores both bands	8.3.1
Correlation band	tanh of the band of $\operatorname{atanh}\rho$	–	inversion on an entry	8.3.2
Coincidence factor	$(m+2.326\,s_S)/\big(f\sum_k(\mu_k+2.326\,s_k)\big)$	–	diversity at the 99th percentile	8.3.3
Conditional answer	π averaged with ρ held	–	what the covariance sets	8.3.4
Spread of a probability	$\mathrm{sd}_i\,\pi$, with one entry held	–	what each state sets	8.3.4

8.8 Exercises

Exercise 8.1. *Formulation.* The design sheet takes the capacity as exactly 410 kW. Write its question as a probability over the rack state alone and state its answer at the MAP state of this chapter. ($\Pr(S^* > 410\,\text{kW} \mid \boldsymbol{\theta}) = \pi(\boldsymbol{\theta},\ Q^\dagger = 410\,\text{kW})$ of equation (8.7), with no average over the hall's state; at $\hat{\boldsymbol{\theta}}$ the coincident peak has mean 330.11 kW and standard deviation 24.83 kW, so the answer is $1 - \Phi\big((410 - 330.11)/24.83\big) = 1 - \Phi(3.22)$, about 0.0006.)

Exercise 8.2. *Derivation.* Derive equation (8.11) from the model (8.1), and check the standard deviation of the coincident peak at the MAP state. (By equation (8.10), the diagonal sums to $\sum_k s_k^2$ and the off-diagonal to $\rho \sum_{k \neq j} s_k s_j = \rho[(\sum_k s_k)^2 - \sum_k s_k^2]$. With the MAP spreads 6.37, 4.88, 7.29, 6.24 and 5.70 kW, $\sum_k s_k = 30.48$ kW and $\sum_k s_k^2 = 188.96\,\text{kW}^2$, so $\mathbf{1}^\mathsf{T}\mathbf{C}\,\mathbf{1} = 188.96 + 0.394 \times (929.03 - 188.96) = 480.55\,\text{kW}^2$; its root is 21.92 kW, and times $f = 1.1333$ it is 24.84 kW, the printed 24.83 kW from the unrounded spreads.)

Exercise 8.3. *Derivation.* Show that the Monte Carlo standard error of the reported probability follows from the spread of π, and say why the gap to the exact posterior is not explained by it. (By equation (8.17) the reported value is a mean of $N = 50\,000$ values of π with standard deviation 0.1045, so its standard error is $0.1045/\sqrt{50\,000} = 0.00047$. The exact value uses the same draws with other weights, so their difference, +0.0027, is the departure of the posterior from its Laplace normal, not the noise of independent samples; it is more than five standard errors, against the printed standard error of 0.0005.)

Exercise 8.4. *Application.* Compute the plug-in probability and say why it is thirteen times smaller than the reported one. (At $\hat{\boldsymbol{\theta}}$ and $Q^\dagger(\hat{\mathbf{z}}) = 396.46$ kW, equation (8.7) gives $1 - \Phi\big((396.46 - 330.11)/24.83\big) = 1 - \Phi(2.67)$. The plug-in reads π at the center of the capacity's band, where it is near its median over the draws, 0.0045; the reported value averages π over the band, and the low-capacity draws, with π up to 0.2668 at the 95 % point, raise the average to 0.0481. The plug-in is 0.0038.)

Exercise 8.5. *Application.* A planner proposes to treat hall B's racks as independent, the most diverse case the one-factor model allows. Compute the coincidence factor at the design population that this implies at the MAP means and spreads. (At $\rho = 0$ the standard deviation of the coincident peak is 15.57 kW, so its 99th percentile is $330.11 + 2.326 \times 15.57 = 366.33$ kW; the sum of the racks' 99th percentiles is 410.42 kW at any correlation; the factor is $366.33/410.42 = 0.8926$.)

Part III

Sizing and Redundancy

Chapter 9

Redundancy Level

9.1 N+1 or N+2 for the site's chillers

The drawing for the site's chilled-water plant is on the table. Its design load is 1 000 kW of heat to the chilled-water loop on the design day, a wet bulb of 22°C, and it meets that load with $N = 3$ chillers of 400 kW nameplate each: 1 200 kW in all, the drawing's 20 % margin. The design sheet then turns to redundancy. With one of the three out, 800 kW remains, short of the design load, so the sheet adds a fourth unit for N+1: 1 600 kW of nameplate, 1 200 kW with one unit out. It asks whether to add a fifth for N+2, 2 000 kW of nameplate, 1 200 kW with two out.

The sheet answers with a textbook availability. It takes the handbook failure rate of a chiller, 1.5 failures a year, and a repair time of 36 h, so that a unit is down a fraction $q = 0.00616$ of the time; it takes the nameplate as exact, so that three units carry the load and two do not; and it counts the ways the units can be down. The plant with three units is available 0.98162 of the time, with four 0.99977, with five 0.999998. Each added unit buys about two more nines, and the sheet has no rule for when the nines stop paying. The planner's question is the drill question, verbatim,

N+1 or N+2?

It is answered with the method of Chapter 4 on the whole site, which enters the book here: the two halls on one chilled-water plant, fitted on a year of its operating record, with the chillers' outage record beside it. What is new is the kind of question. A unit that is down is not a parameter of the plant but one of its states, and the plant's ability to carry the load is a property of every combination of units up and down, each with its probability and its capacity.

Jargon: redundancy level, N, N+1, N+2, wet bulb, design day, as-built capacity, economizer, failure rate, repair rate, mean time to repair, unavailability, service availability, state enumeration, k-out-of-n, expected shortfall, design-condition hours, shed value, annualized cost, capital recovery factor, plug-in

The *redundancy level* of a plant counts the units it carries beyond those the load needs: *N* is the number the design load needs at nameplate, *N+1* one more and *N+2* two more. The *wet bulb* temperature measures the outdoor air's ability to take heat from a cooling tower; a chiller's capacity falls as it rises, and the *design day* is the hot day the drawing is sized for, here a wet bulb of 22°C. The *as-built capacity* of a unit is what it delivers in place, as distinct from its nameplate. An *economizer* cools the loop through a heat exchanger with the cooling-tower water when the wet bulb is low enough, without the chillers' compressors. The *failure rate* λ of a unit is its number of failures per unit of running time, the *repair rate* μ the reciprocal of its *mean time to repair* (MTTR). The *unavailability* q of a unit is the long-run fraction of time it is down. The *service availability* A of a plant is the probability that its units up carry the load, and its unavailability is $1 - A$. A *state enumeration* lists every combination of units up and down with its probability; for identical units it reduces to the *k-out-of-n* count of the textbook, in which the plant works when at least k of its n units do. The *expected shortfall* is the load the plant fails to carry, averaged over its states. The *design-condition hours* are the hours a year at which the design load is met at the design day's conditions; the *shed value* is the cost of a kilowatt-hour of IT load not served. The *annualized cost* of a unit is its installed price times the *capital recovery factor*, the yearly payment that repays the price with interest over the unit's life. A *plug-in* value is computed with every uncertain entry set to one best value.

Key idea

Redundancy is an enumeration with two readers. Every combination of units up and down is a state of the plant, with a probability set by the units' failure and repair rates and a capacity set by their as-built ratings. The availability reader sums the probabilities of the states that carry the load; the capacity reader reads the headroom in each state. Both are derived quantities of one posterior, and their bands can come from different entries of it. The extra unit is bought when the expected shed cost it saves exceeds what it costs a year, and that comparison carries both bands.

9.1.1 Redundancy read as an enumeration

A plant of n units has 2^n states of its units, each unit up or down. Write a state as the vector $\mathbf{s} = (s_1, \dots, s_n)$ with $s_i = 1$ when unit i is up and 0 when it is down. If each unit is down a fraction q of the time, independently of the others, the probability of a state is the product of its units' probabilities, and the capacity it offers is the sum of the capacities of its units that are up, together with anything that does not fail,

$$P(\mathbf{s} \mid q) = \prod_{i=1}^{n} q^{1-s_i} (1-q)^{s_i}, \qquad \mathrm{Cap}(\mathbf{s}) = \sum_{i=1}^{n} s_i \,\mathrm{Cap}_i + Q_e, \tag{9.1}$$

with Cap_i the as-built capacity of unit i and Q_e the duty of a path that is always available, in kW. The plant carries a load L^* in state $\mathbf{s}$ when $\mathrm{Cap}(\mathbf{s}) \geq L^*$. The two readers of the enumeration follow,

$$A = \sum_{\mathbf{s}} P(\mathbf{s} \mid q)\, \mathbf{1}[\mathrm{Cap}(\mathbf{s}) \geq L^*], \qquad E = \sum_{\mathbf{s}} P(\mathbf{s} \mid q)\, \max(L^* - \mathrm{Cap}(\mathbf{s}),\ 0),$$
$$H(\mathbf{s}) = \mathrm{Cap}(\mathbf{s}) - L^*, \tag{9.2}$$

the service availability A, dimensionless; the expected shortfall E, in kW; and the headroom $H(\mathbf{s})$ of each state, in kW, negative in a state that does not carry the load. The availability counts states, and the headroom measures them. A plant can have a large headroom with all units up and a poor availability, if one unit out drops it below the load; or a small headroom and an excellent availability, if one unit out still leaves a little above the load. Capacity margin and availability are two readings of one table and are never equated.

When the units are identical in capacity the states with the same number k of units down share a capacity and a probability, and the sum over states collapses to a sum over k with the binomial coefficient $\binom{n}{k}$ counting the states in each group: the k-out-of-n formula of the textbook. As-built units are not identical, so the enumeration is kept state by state, and the textbook count is the special case. The derivation of the enumeration from the units' up and down processes, and of its reading as minimal cut sets, is that of *Reliability Engineering by Derivation*, Chapters 13 to 15; this chapter uses its result.

Principle 9.1. *A redundancy level is judged on the enumeration of its unit states. Each state carries a probability from the units' rates and a capacity from their as-built ratings; the availability sums the probabilities of the states that carry the load, and the headroom reads the capacity of each. Both are functions of the posterior of the plant, each with its own band.*

9.1.2 The site and its chiller plant (step 0)

From this chapter on, the book's plant is the whole site: the two halls of Part II, hall A and hall B, served by one chilled-water plant. The plant has three existing chillers C1, C2 and C3 of 400 kW nameplate at the design day, one pump with a 12 kW motor, one shared supply pipe that carries the loop's flow to both halls, and one power feed of 1 600 kVA at power factor 0.95. A waterside economizer is drawn but not built: its conductance is a handbook guess.

The physical relations. In a month of operation at site IT load Q_{IT} and wet bulb T_{wb}, both known inputs, the halls put the heat Q_L into the loop, the loop's water rises by ΔT_w through the halls and returns at $T_s + \Delta T_w$, the economizer would take the duty Q_e, and the chillers take the rest. The relations are

$$\begin{aligned}
Q_L &= Q_{\mathrm{IT}} + Q_o, & V &= Q_L/(\rho c_p\,\Delta T_w),\\
\mathrm{Cap}_i &= N_i\,\mathrm{e}^{\kappa_i}\,[1 - d\,(T_{\mathrm{wb}} - 22)], & Q_e &= G_e^{\mathrm{hb}}\,\mathrm{e}^{g_e}\,\max(T_s + \Delta T_w - T_{\mathrm{wb}},\ 0),\\
\mathrm{COP} &= \mathrm{e}^{c}[1 - c_T\,(T_{\mathrm{wb}} - 18)], & P_{\mathrm{ch}} &= (Q_L - Q_e)/\mathrm{COP},\\
\Delta P &= (V/k_p)^2, \quad k_p = \mathrm{e}^{l_k}, & P_{\mathrm{pump}} &= (\Delta P + \Delta P_{\mathrm{h}})\,V/(1000\,\eta), \quad \eta = \mathrm{e}^{l_e},\\
P_{\mathrm{feed}} &= (1 + a)\,Q_{\mathrm{IT}} + P_{\mathrm{ch}} + P_{\mathrm{pump}},
\end{aligned} \tag{9.3}$$

with Q_o the halls' other heat, V the loop flow in L/s, ρc_p the water's heat per liter and kelvin, Cap_i the as-built capacity of chiller i with nameplate N_i, Q_e the economizer's duty with handbook conductance G_e^{hb}, COP the chiller plant's efficiency, P_{ch} the chiller power, ΔP the pressure drop of the shared pipe in kPa, P_{pump} the pump power with ΔP_{h} the halls' own pressure drop, and P_{feed} the power through the feed, all in kW. The pinned quantities are $\Delta T_w = 10\,\mathrm{K}$, $T_s = 14°\mathrm{C}$, $Q_o = 22\,\mathrm{kW}$, $G_e^{\mathrm{hb}} = 25\,\mathrm{kW/K}$, $c_T = 0.02\,\mathrm{K}^{-1}$, $\Delta P_{\mathrm{h}} = 100\,\mathrm{kPa}$ and the auxiliary fraction $a = 0.06$. The free quantities form the site's state of eight entries,

$$\mathbf{z} = (\kappa_{C1},\ \kappa_{C2},\ \kappa_{C3},\ d,\ g_e,\ c,\ l_k,\ l_e)^{\mathsf{T}}, \tag{9.4}$$

the as-built log-ratios of the three chillers, their capacity derate per kelvin of wet bulb above the 22°C rating, the economizer's log-ratio to the handbook, the log of the chillers' efficiency

at 18°C wet bulb, and the logs of the pipe's flow coefficient and the pump's efficiency. Each element has a limit: the chillers with the economizer must carry the loop's heat, the pipe's drop stay within 150 kPa, the pump within its 12 kW motor, the feed within $1\,600 \times 0.95 = 1\,520$ kW. Chapter 13 sets those limits against each other; this chapter asks about the first alone.

The design day. At the design day, $T_{\mathrm{wb}} = 22°\mathrm{C}$, the derate factor is one and the economizer sees a $24 - 22 = 2$ K approach, so that by equation (9.3)

$$\mathrm{Cap}_i = N_i \, \mathrm{e}^{\kappa_i}, \qquad Q_e = 50 \, \mathrm{e}^{g_e} \, \mathrm{kW}, \tag{9.5}$$

and the derate, the efficiency, the pipe and the pump drop out of the chillers' capacity. They stay in the state, because the operating record identifies the as-built ratios only through all of them together.

The outage state. Each chiller fails at the rate λ and is repaired at the rate μ, the same for every unit. Up and down alternate, the up spells with mean $1/\lambda$ and the down spells with mean $1/\mu$, and the long-run fraction of time a unit is down is the mean down spell over the mean cycle,

$$q = \frac{1/\mu}{1/\lambda + 1/\mu} = \frac{\lambda}{\lambda + \mu}, \tag{9.6}$$

with λ and μ in the same unit of inverse time; the steady-state unavailability of *Reliability Engineering by Derivation*. The textbook's $q = 0.00616$ is the handbook rate times the handbook repair time, the same fraction to first order in q.

Modeling choices at the graph. The chillers are taken at the design day, in steady state, with the design load pinned: this is what makes the enumeration of equation (9.1) a statement about one condition, and it fails as a statement about the year, whose other days carry less load at a lower wet bulb (Chapter 23 takes the year). The units fail and are repaired independently, with one rate pair for all: this makes the probability of a state the product of equation (9.1), and it fails under a common cause, such as a shared power supply or a shared control fault, that takes several units down together. The economizer, the pump, the shared pipe and the feed are taken as always available: this keeps the enumeration to the chillers, the element the redundancy question is about, and it fails where a single shared element can take the whole plant down (Chapter 15 enumerates the shared pipe as a state). The economizer is part of every drawing and adds its duty in every state, with the handbook's guess for its conductance.

The graph. The site is drawn in Figure 9.1. The halls on the left put their heat into the loop, which the pump drives through the shared supply pipe to the chiller plant; the three chillers, highlighted, share the derate and the efficiency, and the drawn economizer sits below them with its guess. The dashed arrows leave the graph for the operating record, and the outage record on the right gives the failure and repair rates of a chiller. The row of boxes at the bottom sketches the enumeration of the N+1 drawing, its sixteen states grouped by the number of units down.

9.1.3 The drawings and their enumeration

The question is asked at a scenario node of Section 4.3.8, with three inputs pinned. The wet bulb is pinned at the design day's 22°C and the load at the design load $L^* = 1\,000$ kW to the loop: this makes the answer a statement on the drawing's own terms, and it fails as a statement about any other condition. The drawing is pinned: three alternatives, each with the drawn economizer,

- N, the three existing chillers C1–C3;

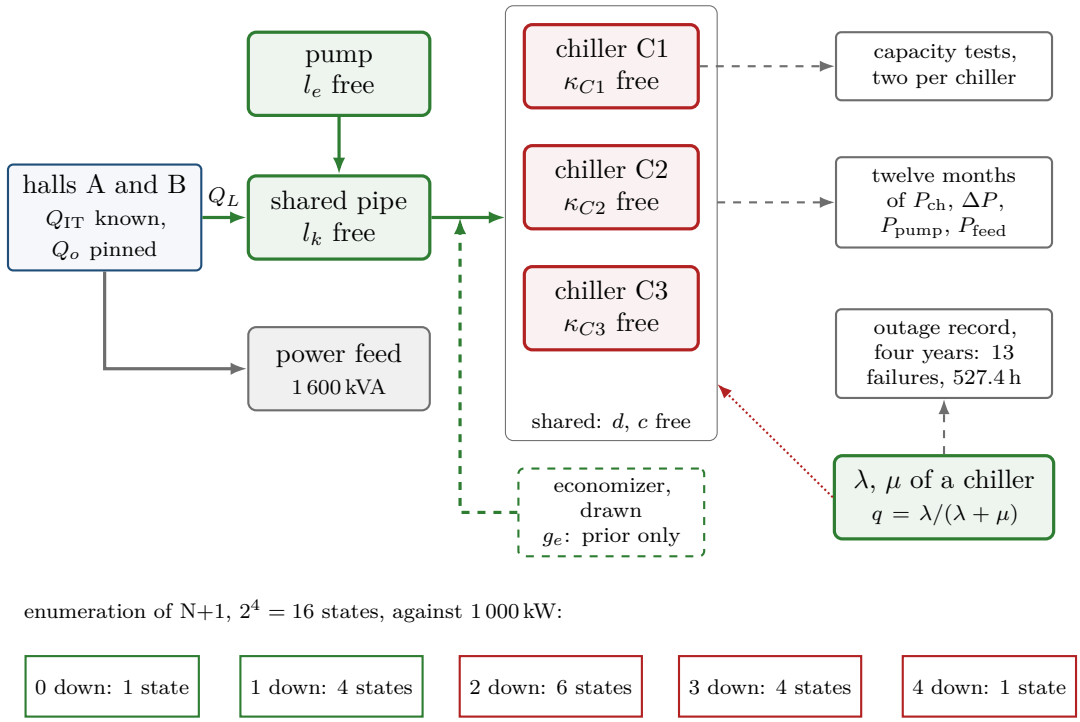


Figure 9.1: The site of Part III. Free (green): the pump's and the pipe's log-coefficients l_e and l_k; the three chillers' as-built log-ratios κ_{C1}, κ_{C2}, κ_{C3} (red border, the units the enumeration counts), with the derate d and the efficiency c they share; and the failure and repair rates λ, μ of a chiller. The economizer (dashed) is drawn but not built, and its log-ratio g_e keeps the handbook prior. Pinned: the halls' other heat, the loop's water rise and the feed's rating. Observed (dashed arrows): two capacity tests per chiller, twelve monthly windows of chiller, pump and feed power and pipe pressure drop, and four years of outages. Bottom: the states of the N+1 drawing grouped by the number of units down; green groups carry the 1 000 kW design load at the posterior means, red groups do not (Table 9.4).

- N+1, the three and one new 400 kW unit;
- N+2, the three and two new 400 kW units.

A new unit is not yet built, so its as-built ratio is not in the record. It is taken to be of the existing units' model, with the mean of their log-ratios and a stated between-unit spread,

$$\kappa_n = \tfrac{1}{3}(\kappa_{C1} + \kappa_{C2} + \kappa_{C3}) + 0.04\,\varepsilon_n, \qquad \varepsilon_n \sim \mathcal{N}(0,\ 1), \tag{9.7}$$

independently for each new unit n, the spread of 0.04 stated as that of the existing units' ratios about their mean; this is a modeling choice, and it fails if a new unit is of another model or is commissioned to another standard.

At the scenario, a drawing of n units has the state probabilities (9.1) with q of equation (9.6), and the capacity of a state is, by equation (9.5),

$$\mathrm{Cap}(\mathbf{s}) = \sum_{i=1}^{n} s_i\, 400\,\mathrm{e}^{\kappa_i} + 50\,\mathrm{e}^{g_e}, \qquad L^* = 1\,000\,\mathrm{kW}, \tag{9.8}$$

with κ_i the as-built log-ratio of unit i, an entry of $\mathbf{z}$ for the existing units and κ_n of equation (9.7) for a new one. The full unknown of the chapter is

$$\boldsymbol{\theta} = (\mathbf{z},\ \lambda,\ \mu,\ \varepsilon_{n1},\ \varepsilon_{n2}), \tag{9.9}$$

the site's state (9.4), the chillers' rates and the two new units' deviations. Given $\boldsymbol{\theta}$, the readers of equation (9.2) are numbers: the unavailability $U(\boldsymbol{\theta}) = 1 - A(\boldsymbol{\theta})$, the expected shortfall $E(\boldsymbol{\theta})$ and the headroom of every state. The unit-state variation is aleatory, already summed inside them; what is not known about $\boldsymbol{\theta}$ is the band.

9.1.4 The planner's questions (step 1)

The planner asks four questions of the three drawings, and step 1 writes each over the unknown (9.9), with the evidence on the right of the bar. Table 9.1 lists them. The first column is the question in the planner's words, the drill question first; the second the general form of its category in the notation of Chapter 2; the third its mathematical form for the site; the fourth its category; the fifth the computation of Section 2.3.6 that answers it; and the last the subsection where it is answered. The drill question is a design question answered by an expected consequence: each step up in redundancy against its cost. The second is a prediction under a threshold summed over the enumeration. The third and fourth are estimation: the capacity each drawing carries, and what the band of the availability is made of. The answers are given in the order three, two, one, four, because the decision reads the first two.

9.1.5 The evidence as rows (step 2)

Step 2 attaches two records to the graph. They share no entry: the operating record sees the site's state $\mathbf{z}$, and the outage record sees the rates λ and μ.

The operating record. The first record is the plant's year and its commissioning. Each chiller was tested twice for capacity, at two different wet bulbs, which together identify its as-built ratio and the derate: C1 read 386.70 kW at 18°C and 363.56 kW at 22°C, C2 388.98 kW at 21°C and 425.84 kW at 15°C, and C3 396.21 kW at 15°C and 361.82 kW at 20°C, each test of standard deviation 8 kW. Twelve monthly windows follow, from a January at 19.83°C wet bulb and 580.64 kW of IT load to a December at 18.54°C and 660.00 kW; each window reads the chiller power, the pipe's pressure drop, the pump power and the feed power, of standard deviations 3 kW, 1.5 kPa, 0.2 kW and 5 kW. The economizer is not built, so no window sees

Table 9.1: The questions asked of the site's chiller plant at the design day. $\mathbf{y}$ is the operating record, the capacity tests, the outage record and the priors; U_a, E_a and H_a are the unavailability, the expected shortfall and the headroom of drawing a, equation (9.2); c_u is the annualized cost of one unit.

Planner's question	General form	Mathematical form	Category	Computation	Answered in
N+1 or N+2?	$\arg\min_a$ $\mathbb{E}[L(a, \boldsymbol{\theta}) \mid \mathbf{y}]$	annual shed cost saved, 200 h × R\$ 1 500/kWh × $(E_a - E_{a+1})$, against c_u, for N to N+1 and N+1 to N+2	design	expected consequence	9.3.3
What is each drawing's service availability, and how sure?	$\Pr(f \geq f^* \mid \mathbf{y})$ over an enumeration	$\mathbb{E}[\sum_{\mathbf{s}} P(\mathbf{s})$ $\mathbf{1}(\mathrm{Cap}(\mathbf{s}) \geq$ 1 000 kW) $\mid \mathbf{y}]$, with the 90 % band of U_a	prediction under a threshold	posterior	9.3.2
How much capacity does each carry, all up and in each state?	moments	H_a with all units up, median and band; $\mathrm{Cap}(\mathbf{s})$ by the number of units down	estimation	posterior	9.3.1
What is the availability band made of?	conditional variance	$\mathrm{sd}(U_{\mathrm{N}+1})$ with q known and with the capacity known	estimation	posterior	9.3.4

it, and its log-ratio g_e keeps its prior. Eight prior rows complete the record: each chiller's log-ratio about zero with standard deviation 0.08, the derate about $0.015\,\mathrm{K}^{-1}$ within 0.005, the economizer's log-ratio about zero within 0.35, and the efficiency, the pipe and the pump about their handbook values, with the spreads listed in Table 9.2. The record has 54 readings and 8 priors, 62 rows on 8 entries.

The outage record. The second record is four years of the three chillers' outages: 12 unit-years of exposure, 13 failures, and 527.4 h of repair in all, a mean of 40.6 h a repair. The failures are counted rows on λ, a Poisson count in the exposure, with the handbook rate of 1.5 failures a year as the mean of a gamma prior of shape 2. The repair times are rows on μ, each an exponential time, with a weak gamma prior of shape 1 whose mean repair time is the handbook's 36 h.

Modeling choices at the rows. Each reading of the operating record is taken as Gaussian, unbiased and independent given the state: this makes the record's objective a sum of squares, and it fails if a power meter carries a bias the rows do not model. The failures are a Poisson process with one rate over the four years and the three units: this makes the count and the exposure the whole of the record's evidence on λ, and it fails if the units age, if one unit is worse than the others, or if failures cluster. The repair times are exponential with one rate: this makes their sum and their number the whole of the evidence on μ, and it fails if repairs have a typical duration with a long tail, such as a part on order. The new units are given the existing units' rates: this is the only knowledge of them until they run.

9.2 The posterior of the state (step 3)

All four questions are read from one posterior, that of the unknown (9.9) under the two records of step 2. This section derives what the answers need from it and computes each part: the one

fit of the site, the posteriors of the rates, and the draws that carry them together.

9.2.1 What the answer needs from the state

The planner's answers are expectations and quantiles of the readers U_a, E_a and H_a over the posterior. The service availability of drawing a is the probability that the plant carries the load at a moment of the design day, and by the law of total probability of equation (3.2) it is the availability given the unknown, averaged over its posterior,

$$\bar{U}_a = 1 - \bar{A}_a = \int U_a(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{y})\, \mathrm{d}\boldsymbol{\theta}, \qquad U_a(\boldsymbol{\theta}) = \sum_{\mathbf{s}} P(\mathbf{s} \mid q)\, \mathbf{1}\big[\mathrm{Cap}(\mathbf{s}) < L^*\big]. \tag{9.10}$$

By equations (9.1) and (9.8), the probability of a state needs q alone, and so, by equation (9.6), λ and μ; the capacity of a state needs the three log-ratios, g_e and the new units' deviations. The two records share no entry, and the new units' ratios depend on $\mathbf{z}$ only through equation (9.7), so the posterior factors,

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = p(\mathbf{z} \mid \mathbf{y}_{\mathrm{op}})\; p(\lambda \mid \mathbf{y}_{\mathrm{out}})\; p(\mu \mid \mathbf{y}_{\mathrm{out}})\; \phi(\varepsilon_{n1})\, \phi(\varepsilon_{n2}), \tag{9.11}$$

with $\mathbf{y}_{\mathrm{op}}$ the operating record, $\mathbf{y}_{\mathrm{out}}$ the outage record and ϕ the standard normal density. The failures and the repairs are separate rows of the outage record on separate entries, so their two posteriors factor too. The remaining three subsections compute the three factors: the site's state by the one fit, the rates by their conjugate posteriors, and the new units by their stated spread.

9.2.2 The one fit of the site

The posterior of $\mathbf{z}$ is the fit of the operating record, built as in Chapter 4: the 62 rows of step 2, each a reading's departure from equation (9.3) over its standard deviation or a prior's departure over its spread, form the residual vector; the MAP state minimizes half its squared length by the normal equations of equation (4.29); and the Laplace covariance $\boldsymbol{\Sigma}$ of equation (4.38) is the inverse of $\mathbf{J}^{\mathsf{T}}\mathbf{J}$ at the minimum. The iteration from the priors' means takes the objective from 608.6492 to 29.7038 in three steps, with the gradient's norm below 10^{-4} by the fourth. Twice the objective, 59.408 on 54 degrees of freedom, has an upper tail probability of 0.285: the readings sit within their noise. Parts III to VII read this one fit, and its derivation is not repeated in later chapters.

Table 9.2 gives the fit. Each row is an entry of equation (9.4); the columns give its prior mean and spread, its MAP value and posterior spread $\sqrt{\Sigma_{ii}}$, the stated truth from which the record was generated, a check marker for Section 9.4.2, and the ratio of the posterior spread to the prior's. The three as-built log-ratios are known to about a fifth of their prior spread, and all three lie below zero: at the MAP state the chillers deliver 365.2, 384.5 and 356.3 kW at the design day against their 400 kW nameplate. The pipe and the efficiency are known best, read by every month's pressure drop and chiller power. The economizer's log-ratio stays at its prior with ratio 1.000, since no reading sees it, so every drawing with the economizer carries the handbook's guess and its spread of 0.35.

9.2.3 The outage record's posteriors

The failure count is Poisson in the exposure and the prior is gamma, so the posterior of λ is the gamma of equation (2.39), the prior's shape plus the count and its rate plus the exposure. With shape 2 and mean 1.5 failures a year, the prior's rate is $2/1.5 = 1.33\,\mathrm{yr}$, and

$$\lambda \mid \mathbf{y}_{\mathrm{out}} \sim \mathrm{Gamma}\big(2{+}13,\ 1.33\,\mathrm{yr}{+}12\,\mathrm{yr}\big), \qquad \mu \mid \mathbf{y}_{\mathrm{out}} \sim \mathrm{Gamma}\big(1{+}13,\ 36\,\mathrm{h}{+}527.4\,\mathrm{h}\big). \tag{9.12}$$

Table 9.2: The one fit of the site: the eight entries of the state on the operating record, with the prior, the MAP value and posterior spread, and the stated truth of the record. The as-built ratios, the economizer and the efficiency are log-ratios; the derate is per kelvin.

Entry [unit]	Prior mean [entry's]	Prior sd [entry's]	MAP [entry's]	$\sqrt{\Sigma_{ii}}$ [entry's]	Truth [entry's]	Ratio [–]
κ_{C1} [–]	0.0000	0.0800	−0.0909	0.0158	−0.1054	0.197
κ_{C2} [–]	0.0000	0.0800	−0.0394	0.0175	−0.0619	0.218
κ_{C3} [–]	0.0000	0.0800	−0.1157	0.0189	−0.1278	0.236
d [K^{-1}]	0.0150	0.0050	0.0152	0.0029	0.0180	0.577
g_e [–]	0.0000	0.3500	0.0000	0.3500	−0.2231	1.000
c [–]	1.7047	0.1500	1.6176	0.0061	1.6094	0.041
l_k [–]	0.6931	0.3000	0.8083	0.0046	0.8087	0.015
l_e [–]	−0.3567	0.1000	−0.2891	0.0190	−0.3285	0.190

Table 9.3: The posteriors of a chiller's failure rate, mean time to repair and unavailability, from the four-year outage record, by 20 000 draws of the gamma posteriors (9.12).

Quantity [unit]	Mean [unit]	Median [unit]	5 % point [unit]	95 % point [unit]
Failure rate λ [1/yr]	1.12418	1.09712	0.69289	1.63629
Mean time to repair $1/\mu$ [h]	43.37695	41.23764	27.25707	66.93155
Unavailability q [–]	0.00554	0.00514	0.00277	0.00959

The second follows the same way: each repair time t_j is exponential with rate μ, so the 13 repairs contribute the likelihood $\mu^{13}\,\mathrm{e}^{-\mu\sum_j t_j}$, the kernel of a gamma in μ, and the prior of shape 1 and mean repair time 36 h adds one to the shape and 36 h to the rate. Both posteriors are exact; no Laplace normal stands in for them. Their mean failure rate is $15/13.33 = 1.125$ failures a year, and their mean repair time, the mean of $1/\mu$, is $563.4/13 = 43.3$ h.

The posteriors are drawn, and q is computed in each draw by equation (9.6). Table 9.3 gives the result of 20 000 draws: each row is a quantity, the columns its mean, median and 90 % band.

The record moves both rates from the handbook's. The failure rate falls from 1.5 to a mean of 1.124 a year, since 13 failures in 12 unit-years is fewer than the handbook expects; the repair time rises from 36 to a mean of 43.4 h, since the record's repairs averaged 40.6 h. The two partly cancel in q, whose mean 0.00554 lies below the textbook's 0.00616. Its band is wide: 0.00277 to 0.00959, a factor of about three and a half, because thirteen failures fix a rate to about a quarter of itself.

9.2.4 The new units and the draws

The answers are computed on 20 000 draws of the unknown, each made of a draw $\mathbf{z}_i$ of the site's state from the Laplace normal $\mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$, a draw (λ_i, μ_i) from the exact posteriors (9.12), and two draws $\varepsilon_{n1,i}$ and $\varepsilon_{n2,i}$ for the new units' deviations, by the factorization (9.11). In each draw the 2^n states of each drawing are enumerated by equations (9.1) and (9.8), and the readers U_a, E_a and the headroom with all units up are recorded. The reported answers are the plain averages and quantiles over the draws, the state route. The exact posterior is the same draws under the weights of equation (5.16), with the objective of the site's fit: only the site's state needs reweighting, since the rates are drawn from their exact posteriors and the new units from their stated normal. The draws are made with the seed 9; their weights have an effective sample size of 19 913 of the 20 000, so the Laplace normal of the fit loses almost nothing against the exact posterior.

9.3 The answers (step 4)

Every answer is read from the draws of Section 9.2.4. The reported value is the average or quantile over the draws, the state route; the exact posterior stands beside it, and the plug-in enumeration at the posterior means stands beside both as the single-value route.

9.3.1 Capacity in each state

Planner's question. How much capacity does each drawing carry, all up and in each state? The question asks for the capacity reader of the enumeration: the headroom with every unit up, with its band, and the capacity left as units go down.

General form. An estimation question asks for the moments of a derived quantity under the posterior, equation (2.10),

$$\mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}] = \int f(\mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

with the band from the quantiles of its distribution, equation (2.11).

Specialized form. The question is equation (2.10) with the unknown $\boldsymbol{\theta}$ of equation (9.9) in place of $\mathbf{z}$ and, for drawing a of n_a units,

$$f(\boldsymbol{\theta}) = H_a(\boldsymbol{\theta}) = \sum_{i=1}^{n_a} 400\, \mathrm{e}^{\kappa_i} + 50\, \mathrm{e}^{g_e} - L^*, \tag{9.13}$$

the headroom with all units up, by equation (9.8); its median and 90 % band are read from the draws. The capacity by the number of units down is read from the enumeration at the posterior means, q at its mean, the existing units at the MAP state and the new units at the mean existing ratio: for each number k of units down, the summed probability of the $\binom{n_a}{k}$ states, the range of their capacities and headrooms, and how many carry the load.

Building blocks. The MAP state and the draws of the site (Section 9.2.2); the mean of q (Table 9.3); the scenario capacities (9.8); the enumeration (9.1).

The enumeration. Table 9.4 gives the enumeration at the posterior means, $q = 0.00554$, with the existing units at 365.2, 384.5 and 356.3 kW, each new unit at 368.5 kW, and the economizer at 50.00 kW. Each row is a group of states with the same number of units down. The columns carry both readers: on the availability side the group's summed probability and how many of its states carry 1 000 kW; on the capacity side the range of the group's capacities and of its headrooms over the design load. The same capacities are drawn in Figure 9.2, one short bar per group and drawing, with the design load as a horizontal line.

Every drawing loses about 370 kW per unit down, the as-built capacity of one chiller, and the design load falls in the same place in all three: between the group whose capacity is near 1 150 kW, which carries it with 140 to 170 kW to spare, and the group near 790 kW, which falls 200 to 230 kW short. In N that boundary is between all up and one down; in N+1 between one and two down; in N+2 between two and three down. No group straddles the load, so in each drawing a state's carrying the load is decided by how many units are down, not by which. That is the fact on which the availability reader rests in Section 9.3.2, and it holds because the gap between the groups, about 340 kW, is wide against the spread of the as-built units, about 30 kW across a group.

The capacity reader with all units up, read over the draws, gives the headroom of N a median of 157.7 kW with a 90 % band of 124.6 to 200.8 kW; of N+1, 527.1 kW, band 480.9 to

Table 9.4: The enumeration of the three drawings at the design day and the posterior means, the availability reader and the capacity reader side by side. $q = 0.00554$; existing units at the MAP state, new units at the mean existing ratio; economizer 50.00 kW in every state. Headroom is capacity less the 1 000 kW design load.

Drawing	Units down [–]	States [–]	Probability [–]	Carrying 1 000 kW [–]	Capacity [kW]	Headroom [kW]
N	0	1	9.835×10^{-1}	1	1 156.1	156.1
	1	3	1.643×10^{-2}	0	771.5 to 799.8	−228.5 to −200.2
	2	3	9.145×10^{-5}	0	406.3 to 434.5	−593.7 to −565.5
	3	1	1.697×10^{-7}	0	50.0	−950.0
N+1	0	1	9.780×10^{-1}	1	1 524.6	524.6
	1	4	2.178×10^{-2}	4	1 140.0 to 1 168.3	140.0 to 168.3
	2	6	1.819×10^{-4}	0	771.5 to 803.0	−228.5 to −197.0
	3	4	6.751×10^{-7}	0	406.3 to 434.5	−593.7 to −565.5
	4	1	9.396×10^{-10}	0	50.0	−950.0
N+2	0	1	9.726×10^{-1}	1	1 893.1	893.1
	1	5	2.707×10^{-2}	5	1 508.5 to 1 536.8	508.5 to 536.8
	2	10	3.015×10^{-4}	10	1 140.0 to 1 171.5	140.0 to 171.5
	3	10	1.678×10^{-6}	0	771.5 to 803.0	−228.5 to −197.0
	4	5	4.672×10^{-9}	0	406.3 to 434.5	−593.7 to −565.5
	5	1	5.202×10^{-12}	0	50.0	−950.0

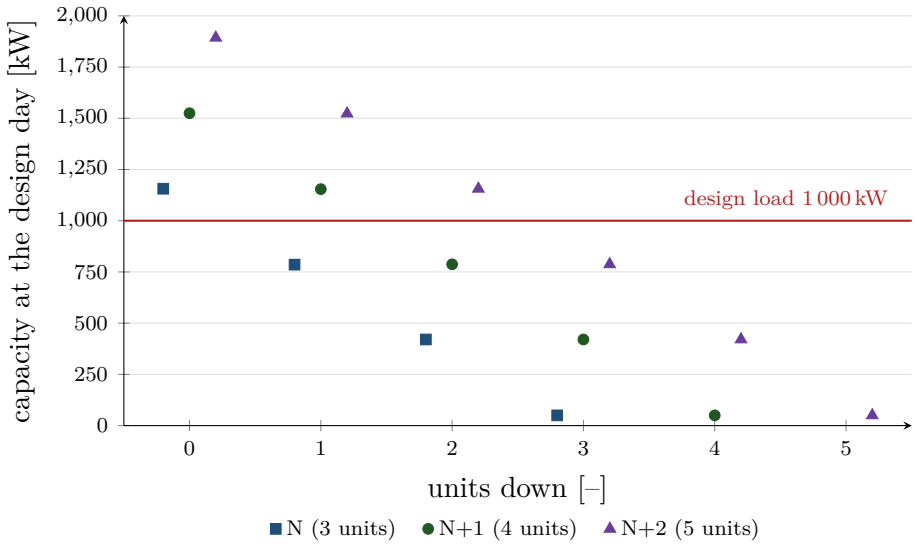


Figure 9.2: Capacity by the number of units down for the three drawings, at the design day and the posterior means (Table 9.4). The markers sit at the middle of each group's range, which is at most 32 kW wide (Table 9.4) and so lies inside the marker, from the state that keeps the weakest units to the one that keeps the strongest; the groups of one drawing are offset slightly to the left (N) or right (N+2). The red line is the 1 000 kW design load.

580.2 kW; and of N+2, 896.1 kW, band 838.7 to 958.8 kW. In none of the 20 000 draws does the all-up capacity of any drawing fall below the load. The bands widen with each new unit, because each new unit adds its own stated spread to the existing units' uncertainty. How the existing units' as-built capacity stands against the drawing's margin is the question of Chapter 16; here it is one column of the enumeration.

9.3.2 The service availability of each drawing

Planner's question. What is each drawing's service availability, and how sure? The question asks for the availability reader: the probability that the plant carries the design load, with its band.

General form. A prediction under a threshold is the posterior probability of an event, equation (2.12),

$$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) \geq f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

asked here of an event that is random even when the state is known: the plant's capacity at a moment depends on which units are down.

Specialized form. The question is equation (2.12) with f the available capacity $\mathrm{Cap}(\mathbf{s})$ of equation (9.8), $f^* = L^* = 1\,000\,\mathrm{kW}$, and the integral taken over both the unit states and the unknown, which is the service unavailability $\bar{U}_a$ of equation (9.10). It is reported by the draws,

$$\bar{U}_a \approx \frac{1}{N}\sum_{i=1}^{N} U_a(\boldsymbol{\theta}_i), \qquad \bar{E}_a \approx \frac{1}{N}\sum_{i=1}^{N} E_a(\boldsymbol{\theta}_i), \tag{9.14}$$

with $N = 20\,000$, and its band is the 90 % band of the conditional unavailability $U_a(\boldsymbol{\theta}_i)$ over the draws: how far the unavailability itself is known. The exact posterior is the same sums under the weights; the plug-in route enumerates once at the posterior means. Since no group straddles the load (Section 9.3.1), the conditional unavailability is the probability that too many units are down, the binomial tail

$$U_a(\boldsymbol{\theta}) = \sum_{k=r_a}^{n_a} \binom{n_a}{k} q^k (1-q)^{n_a-k}, \qquad r_{\mathrm{N}} = 1,\ r_{\mathrm{N+1}} = 2,\ r_{\mathrm{N+2}} = 3, \tag{9.15}$$

with r_a the fewest units down that leave less than the load, and to leading order $U_a \approx \binom{n_a}{r_a} q^{r_a}$: $3q$, $6q^2$ and $10q^3$.

Building blocks. The draws and the weights (Section 9.2.4); the enumeration (9.1) in each draw; the readers (9.2).

The two readers. Table 9.5 sets the answer beside the capacity reader. Each row is a drawing. The availability columns give the mean unavailability by the draws with the exact posterior's beside it, its 90 % band, and the plug-in at the posterior means; the capacity columns give the expected shortfall, the exact posterior's beside it, and the headroom with all units up. The bands of the unavailability are drawn in Figure 9.3 on a logarithmic axis, one row per drawing, with the textbook's value of the opening scene and the stated truth, a check marker for Section 9.4.2, marked on each.

By the draws, the unavailability of N is 1.650×10^{-2}, of N+1 2.104×10^{-4}, and of N+2 2.582×10^{-6}: service availabilities of 0.98350, 0.99979 and 0.9999974. Each unit added divides the unavailability by about eighty, as the leading terms $3q$, $6q^2$ and $10q^3$ of equation (9.15)

Table 9.5: The availability reader and the capacity reader of the three drawings at the design day, by 20 000 draws of the posterior, the exact posterior beside each; the plug-in enumerates once at the posterior means (Table 9.4).

	Availability reader			Capacity reader		
Drawing	Unavailability (exact) [–]	90 % band [–]	Plug-in [–]	Expected shortfall (exact) [kW]	Headroom all up [kW]	90 % band [kW]
N	1.650×10^{-2} (1.651×10^{-2})	8.285×10^{-3} to 2.850×10^{-2}	1.652×10^{-2}	3.4923 (3.4985)	157.7	124.6 to 200.8
N+1	2.104×10^{-4} (2.104×10^{-4})	4.584×10^{-5} to 5.449×10^{-4}	1.826×10^{-4}	0.0444 (0.0445)	527.1	480.9 to 580.2
N+2	2.582×10^{-6} (2.583×10^{-6})	2.115×10^{-7} to 8.697×10^{-6}	1.683×10^{-6}	0.0005 (0.0005)	896.1	838.7 to 958.8

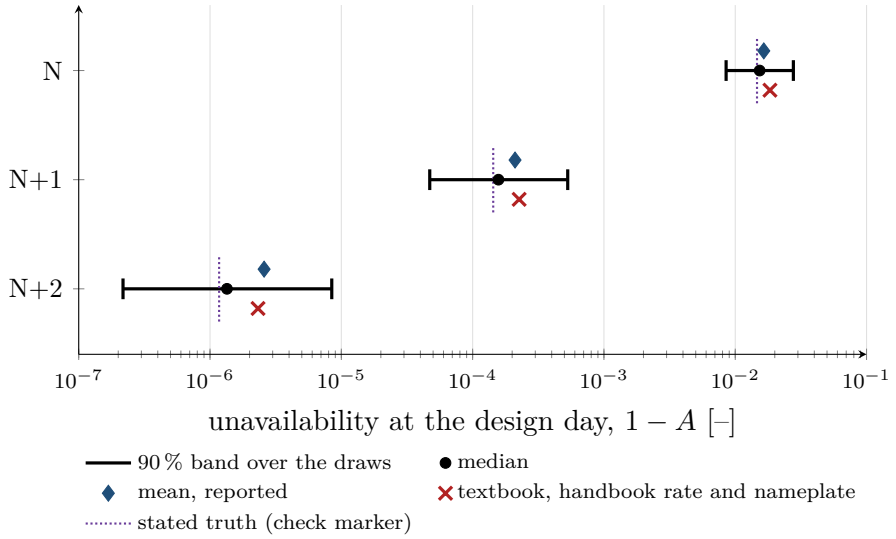


Figure 9.3: The unavailability of the three drawings at the design day, $\bar{U}_a$ of equation (9.10), on a logarithmic axis. Bars: the 90 % band of the conditional unavailability over the 20 000 draws, with its median (dot); diamonds, above each bar: the mean over the draws, the reported answer. Below each bar: the textbook's value of the opening scene (red cross). Dotted: the unavailability at the stated truth of the record, a check marker for Section 9.4.2.

require: each is the last times a number of the order of $1/q$. The exact posterior moves the first by one unit in the fourth figure, 1.651×10^{-2}, leaves the second at 2.104×10^{-4}, and moves the third to 2.583×10^{-6}.

The bands are wide. The unavailability of N+1 runs from 4.584×10^{-5} to 5.449×10^{-4}, a factor of about twelve, and that of N+2 from 2.115×10^{-7} to 8.697×10^{-6}, a factor of about forty. The band of q is a factor of three and a half, and the unavailability of a drawing that needs r_a units down goes as q^{r_a}, so each step of redundancy raises the band's width to one more power. The more redundant the drawing, the less its availability is known in proportion, though the less it matters in kilowatts.

The plug-in is not the answer, and the gap grows with the redundancy. For N it is 1.652×10^{-2}, the reported value to three figures, since $3q$ is linear in q and the mean of a linear function is the function of the mean. For N+1 it is 1.826×10^{-4} against the reported 2.104×10^{-4}, and for N+2 1.683×10^{-6} against 2.582×10^{-6}. The mean of q^2 exceeds the square of the mean of q by the variance of q, and the mean of q^3 exceeds the cube by more: the plug-in drops the outage record's uncertainty from an answer that is convex in it, and understates the unavailability of N+2 by about a third.

The textbook's values, 1.838×10^{-2}, 2.261×10^{-4} and 2.321×10^{-6}, lie inside the three bands and above the reported means of N and N+1. They come near the reported answers by two errors that partly cancel: a handbook q above the record's mean, and a single value of q where the answer averages a convex function of it. Taking the nameplate as exact costs the textbook nothing here, since by Section 9.3.1 the as-built shortfall moves every group's capacity without moving any group across the load.

The capacity reader tells another story with the same states. The expected shortfall of N is 3.4923 kW, mostly the one-down states' 200 to 230 kW shortfall times their probability; that of N+1 is 0.0444 kW and of N+2 0.0005 kW. The headroom with all units up, by contrast, grows by about 370 kW a unit. Read together, the two columns of Table 9.5 say what N+1 buys: not more headroom on an ordinary hour, which N already has, but a plant whose second-weakest state still carries the load.

9.3.3 N+1 or N+2

Planner's question. N+1 or N+2? The question asks which redundancy level to build, each extra unit set against what it saves.

General form. A design question chooses the action of least expected loss, equation (2.23),

$$\mathcal{R}(a) = \mathbb{E}\big[L(a, \mathbf{z}) \mid \mathbf{y}\big], \qquad a^* = \arg\min_a \mathcal{R}(a),$$

with the loss of an action its cost plus the consequence of what happens under it.

Specialized form. The question is equation (2.23) with the actions the three drawings and the loss, per year, the annualized cost of the units added plus the cost of the load shed in the design-condition hours,

$$\mathcal{R}(a) = m_a\, c_u + h_d\, v\, \mathbb{E}\big[E_a(\boldsymbol{\theta}) \mid \mathbf{y}\big], \qquad c_u = \text{R\$}\,1\,600\,000 \times 0.131\,\text{yr}^{-1} = \text{R\$}\,209\,600\,\text{yr}^{-1}, \tag{9.16}$$

with m_a the number of units added, $h_d = 200$ h a year the design-condition hours, $v = \text{R\$}\,1\,500$ per kWh the shed value, and c_u the installed price of one 400 kW unit times the capital recovery factor of 0.131 a year, fifteen years at ten percent; all are stated inputs, not estimates. One step from drawing a to drawing $a+1$ pays when its saving exceeds the cost of its unit,

$$S_{a\to a+1}(\boldsymbol{\theta}) = h_d\, v\, \big[E_a(\boldsymbol{\theta}) - E_{a+1}(\boldsymbol{\theta})\big], \qquad \text{step pays if } \mathbb{E}[S_{a\to a+1} \mid \mathbf{y}] > c_u, \tag{9.17}$$

and the probability $\Pr(S_{a\to a+1} > c_u \mid \mathbf{y})$ says how sure that is. The saving is read in each draw, with the same states for both drawings, so that its band carries the correlation between them.

Table 9.6: The two steps of redundancy against the annualized cost of one 400 kW unit, R$ 209 600 a year, at 200 design-condition hours a year and R$ 1 500 per kWh shed. Draws of the posterior; the exact posterior's mean saving beside.

Step	Gain in availability [–]	90 % band [–]	Shortfall removed [kW]	Saving (exact) [R$/yr]	90 % band [R$/yr]	$\Pr(S > c_u)$ [–]
N to N+1	1.629×10^{-2}	8.239×10^{-3} to 2.795×10^{-2}	3.4479	1 034 370 (1 036 215)	501 712 to 1 810 363	1.0000
N+1 to N+2	2.078×10^{-4}	4.563×10^{-5} to 5.362×10^{-4}	0.0438	13 149 (13 175)	2 787 to 34 234	0.0000

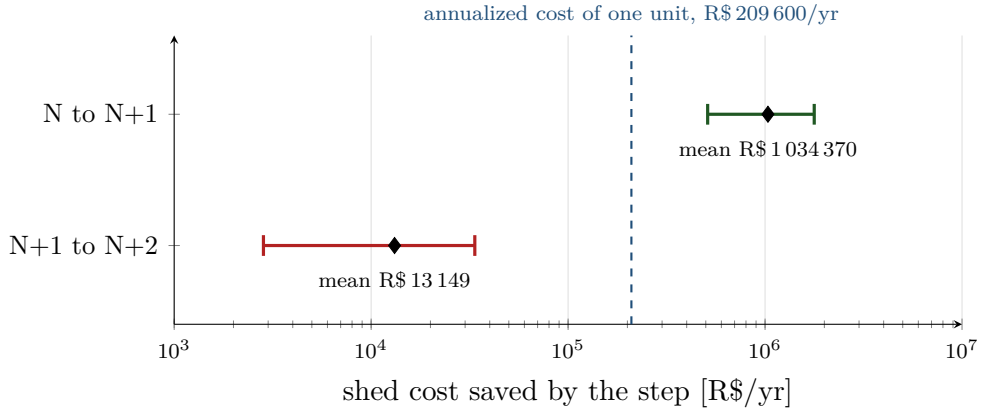


Figure 9.4: The shed cost saved per year by each step of redundancy, $S_{a \to a+1}$ of equation (9.17): the 90 % band over the draws (bar) and the mean (diamond), on a logarithmic axis, against the annualized cost of the unit the step adds (dashed). The first band lies wholly to the right of the cost and the second wholly to the left.

Building blocks. The expected shortfalls in each draw (Section 9.3.2); the stated inputs h_d, v, the unit price and the capital recovery factor.

The increments. Table 9.6 gives the two steps. Each row is a step; the columns give the gain in availability and its band, the fall in the expected shortfall, the saving per year with its band and the exact posterior's mean, the annualized cost of the unit, and the probability that the saving exceeds it. The saving's band is set against the cost in Figure 9.4, on a logarithmic axis in reais a year.

The first step pays and the second does not, and neither is close. N to N+1 removes 3.4479 kW of expected shortfall, which at 200 h a year and R$ 1 500 per kWh is R$ 1 034 370 a year, almost five times the unit's R$ 209 600; even the lower end of its band, R$ 501 712, is more than twice the cost, and the saving exceeds the cost in every draw. N+1 to N+2 removes 0.0438 kW, worth R$ 13 149 a year; even the upper end of its band, R$ 34 234, is a sixth of the cost, and the saving falls short in every draw. The exact posterior moves the two means to R$ 1 036 215 and R$ 13 175. The answer is N+1.

The gain in availability tells the same story in other units: N to N+1 gains 1.629×10^{-2}, N+1 to N+2 gains 2.078×10^{-4}. The textbook's way of reading it, two more nines for each unit, is true and beside the point. A nine is worth what it saves, and the first unit saves about eighty times what the second does, because it is bought against the states that are likely.

9.3.4 What the availability band is made of

Planner's question. What is the availability band made of? The question asks which part of the unknown, the outage rates or the as-built capacities, stands behind the band of the unavailability.

General form. The variance of a derived quantity under the posterior, equation (2.10),

$$\operatorname{Var}(f(\mathbf{z}) \mid \mathbf{y}) = \int \left(f(\mathbf{z}) - \mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}]\right)^2 p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

read here against the variance that remains when part of the state is known.

Specialized form. The question is equation (2.10) with $f = U_{\mathrm{N+1}}$ of equation (9.10), split by the law of total variance of equation (3.32) between q and the capacity entries $\boldsymbol{\kappa} = (\mathbf{z}, \varepsilon_{n1})$,

$$\operatorname{Var}(U \mid \mathbf{y}) = \mathbb{E}\big[\operatorname{Var}(U \mid q, \mathbf{y}) \mid \mathbf{y}\big] + \operatorname{Var}\big(\mathbb{E}[U \mid q, \mathbf{y}] \mid \mathbf{y}\big), \tag{9.18}$$

with the first term the variance that remains when q is known and the second the share of q. It is read on the draws by holding one part: q at its posterior mean in every draw, or the site's state at its MAP value in every draw, and taking the standard deviation of U over the draws that remain. By equation (9.15), if no group of states straddles the load in any draw, U is a function of q alone, the first term is zero, and holding the capacity changes nothing.

Building blocks. The draws (Section 9.2.4); the conditional unavailability (9.15); the mean of q and the MAP state.

The split. The standard deviation of the unavailability of N+1 over the posterior is 1.840×10^{-4}. With q known at its mean it is zero; with the capacity known at the MAP state it is 1.840×10^{-4}, unchanged. The band of the availability is entirely the outage record's. The reason is the one Section 9.3.1 found: with one unit out, N+1 keeps 1 140.0 to 1 168.3 kW, and with two out 771.5 to 803.0 kW, so across the capacity's band the design load never moves from the gap between those groups, and the as-built ratings decide no state.

The headroom's band is the other way round. The headroom with all units up is a function of the capacity entries alone, by equation (9.13); it does not depend on q at all, and its band of 480.9 to 580.2 kW for N+1 is entirely the capacity's: the as-built ratios, the new unit's spread and the economizer's guess, whose shares are the question of Chapter 17. The two readers of one enumeration have their bands from disjoint parts of the posterior. A measurement that narrows the as-built capacity narrows the headroom and leaves the availability's band where it is; more outage years narrow the availability and leave the headroom. This is why availability and margin must never be equated: they are not two scales of one quantity but two quantities, known from different evidence.

9.4 The checks (step 5)

The answers of Section 9.3 rest on draws, on the Laplace normal of the site's fit and on a record generated from a stated truth; each is tested here, and the answer is tested against the one input it depends on most, the design load.

Table 9.7: The checks of the redundancy answers. The reported values are the means over 20 000 draws; beside them, the exact posterior or the route named; the truth is the stated truth of the records, with the new units at the mean true ratio.

Quantity	Reported [– or kW]	Beside [– or kW]	Truth [– or kW]	Reading
q, posterior mean	0.00554	–	0.00491	inside [0.00277, 0.00959]
$\bar{U}$, N	1.650×10^{-2}	exact 1.651×10^{-2}	1.465×10^{-2}	inside the band
$\bar{U}$, N+1	2.104×10^{-4}	exact 2.104×10^{-4}	1.435×10^{-4}	inside the band
$\bar{U}$, N+2	2.582×10^{-6}	exact 2.583×10^{-6}	1.173×10^{-6}	inside the band
$\bar{U}$, N+1, plug-in	1.826×10^{-4}	–	1.435×10^{-4}	drops the convexity in q
$\bar{U}$, N+1, textbook	2.261×10^{-4}	–	1.435×10^{-4}	handbook q, nameplate
$\bar{E}$, N	3.4923	exact 3.4985	3.4640	close
$\bar{E}$, N+1	0.0444	exact 0.0445	0.0339	above the truth
$\bar{E}$, N+2	0.0005	exact 0.0005	0.0003	above the truth
Unavailability of N+1 against the load, by the draws				
700 kW	1.040×10^{-6}	–	–	three or more down
800 kW	1.381×10^{-4}	–	–	most two-down states
900 kW	2.102×10^{-4}	–	–	every two-down state
1 000 kW, design load	2.104×10^{-4}	–	–	the same plateau
1 100 kW	5.177×10^{-4}	–	–	one-down states in low draws
Load carried with the largest as-built unit out, MAP				
N, N+1, N+2 [kW]	771.5, 1 140.0, 1 508.5			

9.4.1 The draws against the exact posterior

The reported values average the 20 000 draws with equal weight; the exact posterior reweights the same draws by equation (5.16), the site's state only, with an effective sample size of 19 913. The gaps, exact minus reported, lie far inside the Monte Carlo standard error of the reported mean. In the unavailability the gap is 2.173×10^{-6} against a standard error of 4.561×10^{-5} for N, 6.926×10^{-8} against 1.301×10^{-6} for N+1, and 1.390×10^{-9} against 2.883×10^{-8} for N+2, each about a twentieth of the standard error. In the expected shortfall the exact posterior adds 0.0062 kW to the 3.4923 kW of N and 0.0001 kW to that of N+1; in the savings it adds R$ 1 845 to R$ 1 034 370 and R$ 26 to R$ 13 149 a year. The unavailability barely moves because, by Section 9.3.4, it does not depend on the site's state at all, and the reweighting acts only on that state; the shortfall of N moves the most, because it measures how far the one-down states fall short, and that is a capacity.

9.4.2 The answers against the stated truth

The operating record and the outage record are generated from a stated truth: the entries in the last column of Table 9.2, and a chiller failure rate of 1.2 a year with repairs of 36 h, a q of 0.00491. Table 9.7 collects the checks: each row is a quantity, the columns its reported value, the value beside it, the truth, and the reading. At the truth, with the new units at the mean true ratio, the unavailability of N is 1.465×10^{-2}, of N+1 1.435×10^{-4}, and of N+2 1.173×10^{-6}, each inside its band (Figure 9.3) and below the reported mean, since the true q of 0.00491 lies below the posterior mean of 0.00554 and inside its band. The true shortfall of N is 3.4640 kW, of N+1 0.0339 kW, of N+2 0.0003 kW. The decision at the truth is the reported one: the true shortfall removed by N+2 is smaller still than the reported.

9.4.3 The design load against the drawings' steps

The availability of N+1 was found to depend on the outage record alone, and the reason was a gap: the design load sits between the plant's capacity with two units out and with one. The lower part of Table 9.7 tests that gap by moving the load. At 700 kW only the states with three or more units down fail, and the unavailability is 1.040×10^{-6}. At 800 kW most of the two-down states, whose capacities run from 771.5 to 803.0 kW, fail, and it is 1.381×10^{-4}. At 900 and 1 000 kW every two-down state fails and every one-down state carries, and the unavailability is flat, 2.102 and 2.104×10^{-4}: the design load sits on a plateau. At 1 100 kW it rises to 5.177×10^{-4}, below the 1 140.0 kW the plant carries at the MAP state with its largest unit out, because in the low draws of the capacity the weakest one-down states no longer carry 1 100 kW. The plateau of each drawing ends near the load it carries with its largest unit out: 771.5 kW for N, 1 140.0 kW for N+1, 1 508.5 kW for N+2. A design load near one of those values would make the availability depend on the as-built capacity, and the split of Section 9.3.4 would change; at 1 000 kW it is 140 kW clear of the nearest.

9.5 The report (step 6)

Buy the fourth unit, not the fifth. At the design day the site's chiller plant with N+1 and the drawn economizer carries the 1 000 kW design load with service availability 0.99979, an unavailability of 2.104×10^{-4} with a 90 % band of 4.584×10^{-5} to 5.449×10^{-4}, a band that is the outage record's. The step from N to N+1 saves about five times its annualized cost; the step to N+2 saves a sixteenth of it.

9.5.1 The answers and their bands

Table 9.8 gives the answers in the order of Table 9.1. Each row is a question; the second column gives the reported answer with the exact posterior beside it, and the third its band or the routes beside it.

9.5.2 The design-review entry

The design review's entry replaces the sheet's nines with a decision. The site's chiller plant is built to N+1: four 400 kW units with the drawn economizer. On the design day it carries the 1 000 kW design load with availability 0.99979 by the draws and by the exact posterior, against the textbook's 0.99977; the unavailability's 90 % band, 4.584×10^{-5} to 5.449×10^{-4}, is set by the four-year outage record and by nothing in the as-built ratings. With all units up the plant has a headroom of 527.1 kW, band 480.9 to 580.2 kW, and with any one unit out it still carries 1 140.0 to 1 168.3 kW. The fifth unit is not bought: at 200 design-condition hours a year and R$ 1 500 per kWh shed, it would save R$ 13 149 a year, with a band whose upper end, R$ 34 234, is a sixth of its annualized cost of R$ 209 600. The entry records its cost basis as stated inputs: the design-day hours stand in for the year, and Chapter 23 replaces them with the year's load and wet bulb.

Common mistake

Reading redundancy as nines. Each unit adds about two nines whether or not they are worth buying; the decision sets the shortfall each unit removes, in the hours and at the value the site carries, against what the unit costs.

Table 9.8: The report for the site's chiller plant at the design day, 1 000 kW to the loop with the drawn economizer. Reported values by 20 000 draws of the posterior; the exact posterior beside each. Cost basis: 200 design-condition hours a year, R$ 1 500 per kWh shed, R$ 209 600 a year per unit.

Question	Answer	Band or routes beside
N+1 or N+2?	N+1: the step N to N+1 saves R$ 1 034 370/yr (exact: R$ 1 036 215), N+1 to N+2 R$ 13 149/yr (exact: R$ 13 175), against R$ 209 600/yr	R$ 501 712 to R$ 1 810 363 and R$ 2 787 to R$ 34 234; $\Pr(S > c_u)$ 1.0000 and 0.0000
What is each drawing's service availability, and how sure?	unavailability N 1.650×10^{-2}, N+1 2.104×10^{-4}, N+2 2.582×10^{-6} (exact: 1.651×10^{-2}, 2.104×10^{-4}, 2.583×10^{-6})	90 % bands 8.285×10^{-3} to 2.850×10^{-2}; 4.584×10^{-5} to 5.449×10^{-4}; 2.115×10^{-7} to 8.697×10^{-6}; textbook 1.838×10^{-2}, 2.261×10^{-4}, 2.321×10^{-6}
How much capacity does each carry, all up and in each state?	headroom all up: N 157.7 kW, N+1 527.1 kW, N+2 896.1 kW (medians); N+1 with one unit out 1 140.0 to 1 168.3 kW, with two out 771.5 to 803.0 kW	124.6 to 200.8, 480.9 to 580.2, 838.7 to 958.8 kW, central 90 %
What is the availability band made of?	the outage record: sd of U_{N+1} 1.840×10^{-4}, 0 with q known	1.840×10^{-4} with the capacity known; the headroom's band is the capacity's

9.6 The next measurement (step 7)

Step 7 asks which test or survey of the existing plant would narrow the quantity that decides the build. Here that quantity is the saving of each step against its cost, and the answer is unusual: none.

9.6.1 A decision no measurement moves

A measurement narrows a band; it changes a decision only if the band straddles the point where the decision turns. The saving of the first step runs from R$ 501 712 to R$ 1 810 363 a year, wholly above the cost of R$ 209 600; the saving of the second runs from R$ 2 787 to R$ 34 234, wholly below it. Neither band touches the cost. More years of the outage record would narrow q and with it both savings, by Section 9.3.4, but they would narrow bands that already lie on one side; capacity tests of the existing units would narrow the headroom and leave the savings where they are. No measurement of the existing plant moves this decision, and the review should buy none for it.

9.6.2 What would move it

What would move the decision is not a measurement of the plant but a change in the stated inputs. The saving of the second step is proportional to the design-condition hours h_d and to the shed value v, by equation (9.17), so N+2 pays when their product grows by the ratio of the cost to the saving: at the mean, $209\,600/13\,149 = 15.9$, and even at the upper end of the saving's band, $209\,600/34\,234 = 6.1$. A site that ran at design conditions nearly sixteen times as many hours, or valued a kilowatt-hour of shed load nearly sixteen times as high, would buy the fifth unit. Neither number is estimated here; both are stated inputs, and the planner who doubts them should restate them, not measure the chillers. Chapter 18 ranks the tests and

surveys a planner can commission on the site plant by the width each takes off the quantity that decides the build, and it is there that a decision near its turning point meets them.

9.6.3 What the analysis showed

The design sheet compared the drawings by their nines and had no rule for when to stop. The analysis answers the question it asked: N+1. The step to N+1 saves almost five times its cost in shed load and the step to N+2 a sixteenth of its cost, and the bands of both savings lie clear of the cost, so no measurement of the plant would change the answer.

Three findings carry the answer. First, the textbook's availability is close to the posterior's for N+1, 2.261 against 2.104×10^{-4}, by two errors that partly cancel: the handbook rate is above the record's, and a single value of q understates an answer that is convex in q, so that the plug-in at the record's mean, 1.826×10^{-4}, is further from the answer than the textbook. Second, the availability is known far less well than its nines suggest: its band spans a factor of twelve for N+1 and forty for N+2, the outage record's factor of three and a half in q raised to the power of the units the drawing must lose. Third, and beyond this site, the availability and the headroom of one enumeration take their bands from disjoint evidence. The availability band is entirely the outage record's, since the design load sits on a plateau between the plant's capacity with one unit out and with two; the headroom band is entirely the capacity's. A margin that narrows does not move the availability, and an outage record that lengthens does not move the margin, which is why availability and capacity margin are reported as two columns and never equated.

9.6.4 Limits of the model

The design day for the year. The enumeration is at the design day and the design load, and the cost of a shortfall is that day's shortfall times 200 stated hours a year. The year's other hours carry less load at a lower wet bulb, where the economizer does more and the chillers can carry more; Chapter 23 replaces the design-day basis with the year's expected unserved energy.

Independent units with one rate. Every chiller fails and is repaired independently, with one failure rate and one repair rate. A common cause that takes two units down together, a shared control power supply or a shared condenser-water fault, would raise the two-down probability of N+1 above $6q^2$ and could change the answer; a unit worse than the others, or rates that change with age, would change the record's evidence on q. The new units are given the existing units' rates.

Shared elements taken as available. The pump, the shared pipe and the feed are taken as always available, and the economizer as always delivering its duty. A shared element that fails takes every state down at once, and its unavailability adds to the plant's whatever the redundancy of the chillers; Chapter 15 enumerates the shared pipe as a state.

Steady-state unavailability. The fraction q is the long-run fraction of time down. Planned maintenance, which takes a unit out on purpose and leaves the plant one unit short for its duration, is not in it; Chapter 22 asks whether the plant can be maintained without leaving the load uncovered.

The new units' capacity. A new unit is the existing units' model with a stated spread of 0.04 in its log-ratio. A new unit of another model, or commissioned to a tighter standard, would move the headroom columns, though not, at this design load, the availability.

One unit size and the present load. The drawings add 400 kW units to the existing three at the design load. Chapter 10 compares unit sizes at the same total, and Chapter 11 lets the load grow toward the year at which the redundancy is spent.

The constructed record. The operating record, the capacity tests and the outage record are generated from a stated truth with the stated reading errors, not taken from a site, and the comparison with that truth in Section 9.4.2 is a check of the method on its own model, not of the model against a real plant.

9.7 What to retain

Definition 9.1 (Redundancy by enumeration)**.** For a plant of n units, each down a fraction $q = \lambda/(\lambda + \mu)$ of the time independently, the state enumeration gives each of the 2^n states $\mathbf{s}$ the probability $\prod_i q^{1-s_i}(1-q)^{s_i}$ and the capacity $\sum_i s_i \,\mathrm{Cap}_i$ plus whatever does not fail. Its two readers are the service availability, the summed probability of the states that carry the load, and the capacity, the headroom of each state; the expected shortfall joins them. Over the posterior of the rates and the as-built capacities each reader has its own mean and band, reported by draws with the exact posterior beside them. A redundancy step is bought when the expected shed cost it removes exceeds the annualized cost of its unit.

Four conclusions carry beyond this site. A redundancy level is decided by cost, not by nines: each unit adds about the same number of nines, and the shortfall it removes falls by the same factor. The availability of a redundant plant is a power of q, so the uncertainty of the outage record is raised to that power: the more redundant the plant, the less its availability is known in proportion, and a plug-in at the mean rate understates it. When the design load lies on a plateau between the capacity with r units out and with $r-1$, the availability is a function of the outage rates alone and the headroom of the capacities alone; a load near the end of a plateau couples them. And a decision whose bands lie on either side of the cost needs no measurement; only a change in the stated inputs can move it.

Table 9.9 lists the terms of the chapter with their expressions, units and the subsection that defines each.

Table 9.9: The terms of redundancy level and where each is defined. Units for the site's chiller plant.

Term	Expression	Units	Meaning	Defined in
State probability	$P(\mathbf{s} \mid q) = \prod_i q^{1-s_i}(1-q)^{s_i}$	–	units down independently	9.1.1
State capacity	$\mathrm{Cap}(\mathbf{s}) = \sum_i s_i \,\mathrm{Cap}_i + Q_e$	kW	units up plus the economizer	9.1.1
Availability, shortfall	$A = \sum P\,\mathbf{1}[\mathrm{Cap} \geq L^*]$, $E = \sum P \max(L^* - \mathrm{Cap}, 0)$	–, kW	the two readers	9.1.1
Site state	$(\kappa_{C1..C3}, d, g_e, c, l_k, l_e)$	–, 1/K	the eight entries	9.1.2
Unit unavailability	$q = \lambda/(\lambda+\mu)$	–	long-run fraction down	9.1.2
New unit	$\kappa_n = \bar{\kappa} + 0.04\,\varepsilon_n$	–	existing model, stated spread	9.1.3
Rows and posterior				
Factored posterior	$p(\mathbf{z})\,p(\lambda)\,p(\mu)\,\phi(\varepsilon_{n1})\,\phi(\varepsilon_{n2})$	–	two records, no shared entry	9.2.1
Rate posteriors	Gamma(15, 13.33 yr), Gamma(14, 563.4 h)	–	failures and repairs	9.2.3
Questions and answers				
Headroom all up	$H_a = \sum_i 400\,\mathrm{e}^{\kappa_i} + 50\,\mathrm{e}^{g_e} - L^*$	kW	the capacity reader	9.3.1
Binomial tail	$U_a = \sum_{k \geq r_a} \binom{n_a}{k} q^k (1-q)^{n_a-k}$	–	no group straddles the load	9.3.2
Plug-in	U_a at the posterior means	–	drops the convexity in q	9.3.2
Annual risk	$\mathcal{R}(a) = m_a c_u + h_d v\,\mathbb{E}[E_a]$	R$/yr	cost plus shed load	9.3.3
Step saving	$S_{a\to a+1} = h_d v(E_a - E_{a+1})$	R$/yr	against c_u	9.3.3
Band split	$\mathrm{Var}(U) = \mathbb{E}[\mathrm{Var}(U \mid q)] + \mathrm{Var}(\mathbb{E}[U \mid q])$	–	outage record or capacity	9.3.4

9.8 Exercises

Exercise 9.1. *Formulation.* The planner asks whether N+1 would still carry the load if the design load were revised down to 800 kW. Write the question as a probability over the enumeration and state its answer from this chapter's draws. ($\bar{U}_{\mathrm{N+1}}(800) = \mathbb{E}[\sum_{\mathbf{s}} P(\mathbf{s} \mid q)\,\mathbf{1}(\mathrm{Cap}(\mathbf{s}) < 800\,\mathrm{kW}) \mid \mathbf{y}]$, equation (9.10) with $L^* = 800$ kW. The two-down states of N+1 run from 771.5 to 803.0 kW, so most of them now fail in a draw and a few carry, and the unavailability by the draws is 1.381×10^{-4}.)

Exercise 9.2. *Derivation.* Derive the unavailability $q = \lambda/(\lambda+\mu)$ of a unit from its mean up and down spells, and check it against the posterior means. (A unit's cycle is an up spell of mean $1/\lambda$ and a down spell of mean $1/\mu$; the long-run fraction down is $(1/\mu)/(1/\lambda + 1/\mu) = \lambda/(\lambda+\mu)$, which is λ/μ to first order when repairs are short. With the posterior means 1.12418 failures a year and 43.37695 h of repair, $\lambda/\mu = 1.12418 \times 43.37695/8\,760 = 0.00557$, and the mean of q over the draws, which averages the ratio rather than taking the ratio of the averages, is 0.00554.)

Exercise 9.3. *Derivation.* Show that the unavailability of N+1 is $6q^2$ to leading order, and use it to check the ends of its band. (By Section 9.3.1 the plant fails exactly when two or more of its four units are down, so by equation (9.15) $U = \binom{4}{2}q^2(1-q)^2 + \binom{4}{3}q^3(1-q) + q^4 = 6q^2 + O(q^3)$. Since U depends on q alone and rises with it, its quantiles are those of q carried through: at the 5 % point of q, $6 \times 0.00277^2 = 4.60 \times 10^{-5}$, and at the 95 % point, $6 \times 0.00959^2 = 5.52 \times 10^{-4}$, which the higher terms lower by a little over one percent; the band by the draws is 4.584×10^{-5} to 5.449×10^{-4}.)

Exercise 9.4. *Application.* Recompute the textbook's unavailability of N+1 and say why it lies close to the reported answer. (With $q = 0.00616$ and nameplate units, three of four carry 1 200 kW and two carry 800 kW, so the plant fails with two or more down: $6q^2 = 6 \times 0.00616^2 = 2.28 \times 10^{-4}$ to leading order. The handbook q is above the record's mean of 0.00554, which raises the textbook value, and the textbook squares one value of q where the answer averages q^2 over the record's band, which lowers it; the two partly cancel against the reported 2.104×10^{-4}. The textbook's exact value is 2.261×10^{-4}.)

Exercise 9.5. *Application.* A colleague argues that N+2 is justified at a higher shed value. Find the shed value at which the step from N+1 to N+2 pays at the mean saving, and at the upper end of the saving's band. (By equation (9.17) the saving is proportional to v, so the step pays when v exceeds R$ 1 500 times the ratio of the cost to the saving: $209\,600/13\,149 = 15.9$ at the mean, a shed value of about R$ 23 900 per kWh, and $209\,600/34\,234 = 6.1$ at the 95 % point, about R$ 9 200 per kWh. At the stated R$ 1 500 per kWh the mean saving is R$ 13 149 a year.)

Chapter 10

Unit Sizing

10.1 Two quotes for one chiller plant

The site needs a new chiller plant of 1 200 kW nameplate, the drawing for its design load of 1 000 kW to the chilled-water loop at the design day, a wet bulb of 22°C. Two quotes are on the table, each with the waterside economizer the drawing adds. The first is three units of 400 kW for R$4.8 million, whose units run at part load with a coefficient of performance (COP) factor of 0.97 against the larger units. The second is two units of 600 kW for R$4.2 million, at a factor of 1.00. The design sheet sets them side by side with its rule of thumb: smaller units are more available, because one unit out of three leaves 800 kW of nameplate where one unit out of two leaves 600 kW; but they are less efficient and dearer. The sheet stops there, with an argument on each side and no number that weighs them.

The planner's question is the drill question, verbatim,

Two 600 kW units or three 400 kW?

It is answered with the method of Chapter 4 on the site plant that Chapter 9 introduces: its one fit, its outage record and its enumeration of unit states. What is new is that two drawings that do not exist yet are compared on three readings at once, availability, efficiency and cost, each with the band the evidence leaves, and that the comparison asks whether the difference between them is real.

Jargon: unit sizing, nameplate, as-built ratio, design day, unavailability, loss-of-load probability, enumeration of unit states, steady-state unavailability, mean time to repair, part-load COP factor, waterside economizer, expected shed load (expected unserved energy), capital recovery factor, paired draws, availability crossover

Unit sizing is the choice of how a plant's total capacity is split among its units. The *nameplate* of a unit is its rated capacity at the design day; its *as-built ratio* κ is the logarithm of its delivered capacity over the nameplate, zero at the rating. The *design day* is the ambient condition the drawing is sized for. The *unavailability* $U(Q)$ of a plant at a load Q is the probability that the units in service cannot carry that load; it is the *loss-of-load probability* at that load, and the availability is $1 - U(Q)$. The *enumeration of unit states* lists every combination of units up and down with its probability and the capacity it leaves. The *steady-state unavailability* q of one unit is the long-run share of time it is down, from its failure rate λ and its repair rate μ; the *mean time to repair* (MTTR) is $1/\mu$. The *part-load COP factor* f scales a unit's coefficient of performance, the cooling it delivers per unit of electricity, for the part load it runs at in a given drawing. A *waterside economizer* cools the loop with outdoor air through a cooling tower and a heat exchanger, without the chillers, whenever the wet bulb is below the return water. The *expected shed load* is the expected cooling energy the plant cannot deliver over a period, in kWh, the expected unserved energy of that period. The *capital recovery factor* turns a price paid once into an equal annual cost over the plant's life at a stated interest rate. *Paired draws* evaluate both drawings on the same draw of the state, so that what they share cancels in their difference. The *availability crossover* is the load at which the ordering of two drawings' unavailabilities reverses.

Key idea

Unit size enters the enumeration twice: in which states carry the load, and in how many units can fail. Three units of 400 kW carry a load below their one-unit-out capacity through any single outage, so they fail only on a double outage, and two units of 600 kW fail on any single one: at such a load the smaller units fail as the square of a unit's unavailability and the larger as its first power. Above the one-unit-out capacity of the smaller units, both drawings fail on a single outage, and three units have more ways to fail than two. The rule of thumb is a statement about one side of a load, and the enumeration says which side the plant is on. The cost comparison adds the annualized price and the energy to the expected shed load, and reads the difference of the two drawings on paired draws of one posterior, so that the band of the difference says whether the difference is real.

10.1.1 Unit size read through the enumeration

A drawing with n units of capacities $\mathrm{Cap}_1, \ldots, \mathrm{Cap}_n$ has 2^n states of its units, each unit up or down. Chapter 9 derives the enumeration, after the reliability book (*Reliability Engineering by Derivation*, Chapters 13 to 15): with every unit down independently with the same steady-state unavailability q, a state $\mathbf{s} = (s_1, \ldots, s_n)$ with $s_i = 1$ for a unit up has the probability and leaves the capacity

$$P(\mathbf{s}) = \prod_{i=1}^{n} q^{\,1-s_i}\,(1-q)^{s_i}, \qquad \mathrm{Cap}(\mathbf{s}) = \sum_{i=1}^{n} s_i\,\mathrm{Cap}_i + Q_e, \qquad q = \frac{\lambda}{\lambda+\mu}, \tag{10.1}$$

with Q_e the economizer's duty, which does not fail with the chillers. The unavailability at a load Q is the summed probability of the states that cannot carry it,

$$U(Q) = \sum_{\mathbf{s}} P(\mathbf{s})\,\mathbf{1}\big[\mathrm{Cap}(\mathbf{s}) < Q\big]. \tag{10.2}$$

Unit size acts through the indicator. For n units of equal size and a small q, two cases cover the loads a planner meets. If the load is above the capacity left by any one unit out, every single outage fails it, and the unavailability is that of at least one unit down; if the load is below the capacity left by any one unit out and above that left by any two, only the double outages fail it. Expanding the products of equation (10.1) to the first power of q that survives,

$$U(Q) = 1 - (1-q)^n \approx n\,q \quad \text{(single outages fail)}, \qquad U(Q) \approx \binom{n}{2} q^2 \quad \text{(single outages carry)}. \tag{10.3}$$

The two cases differ by a factor of about q, a few parts in a thousand for a chiller on handbook rates. At a load the smaller units carry with one unit out and the larger do not, three units of 400 kW fail as $3q^2$ and two of 600 kW as $2q$; at a load neither carries with one unit out, the three fail as $3q$ and the two as $2q$, and the ordering reverses. The load at which it reverses is the capacity of the smaller units with their largest unit out, the availability crossover, and it is a derived quantity of the state: its band is the band of the as-built units.

> **Principle 10.1.** *Smaller units are more available only at loads that one unit out can still carry. Above that load every drawing fails on a single outage, and the drawing with more units has more ways to fail.*

10.1.2 The site and the two drawings (step 0)

The plant is the site of Chapter 9: two halls on one chilled-water loop, three existing chillers C1–C3 of 400 kW nameplate, one pump, the shared supply pipe, the power feed and the drawn economizer. This chapter replaces the chillers by each of two new drawings and asks the enumeration of each the same questions. Both drawings are new: every unit is drawn, none is built, so the comparison is like for like.

The physical relations. A chiller's as-built capacity at the wet bulb T_{wb} is its nameplate N_i times its as-built ratio, derated by d per kelvin of wet bulb above the rating condition of 22°C; a new unit's ratio is that of the existing units' model, their mean, with a spread of its own between units,

$$\mathrm{Cap}_i = N_i\,\mathrm{e}^{\kappa_i}\,\big(1 - d\,(T_{\mathrm{wb}} - 22)\big), \qquad \kappa_i = \bar{\kappa} + 0.04\,\varepsilon_i, \qquad \bar{\kappa} = \tfrac{1}{3}\,(\kappa_{C1} + \kappa_{C2} + \kappa_{C3}), \tag{10.4}$$

with $\varepsilon_i \sim \mathcal{N}(0,1)$ independent between units and $N_i = 400$ or 600 kW. The economizer passes heat from the return water at $T_{\mathrm{wr}} = 24$°C (the 14°C supply plus the 10 K design rise) to the outdoor air through its handbook conductance 25 kW/K, scaled by its log-ratio g_e,

$$Q_e = 25\,\mathrm{e}^{g_e}\,\max(T_{\mathrm{wr}} - T_{\mathrm{wb}},\ 0), \qquad \mathrm{COP} = \mathrm{e}^{c}\,\big(1 - 0.02\,(T_{\mathrm{wb}} - 18)\big), \tag{10.5}$$

in kW, and the chillers carry what the economizer leaves, at the plant's coefficient of performance, which falls by 2 % per kelvin of wet bulb. The relations are the site's, as Chapter 9 states them; nothing in them is re-chosen here.

The two drawings. Drawing *3×400* has $n = 3$ units of $N_i = 400$ kW, drawing *2×600* has $n = 2$ of $N_i = 600$ kW; both total 1 200 kW of nameplate. The design sheet states their prices and COP factors, and their nameplate with one unit out, 800 and 600 kW.

The scenario nodes. Two kinds of scenario are asked of each drawing, each a node of Section 4.3.8 that pins its inputs at values not yet seen. The availability scenario pins the wet bulb at the design day's 22°C and the loop heat Q^* at four loads: 682 kW, today's site IT load of 660 kW plus 22 kW of other heat; 800 and 900 kW; and the design load of 1 000 kW. The annual scenario pins the twelve months of the record year: each month's loop heat $Q_{L,m}$, its IT load plus the 22 kW, and its wet bulb $T_{\mathrm{wb},m}$, held for 730 h, a twelfth of the year. From these, three derived quantities of each drawing a are read. The chiller energy of the year is the chiller duty over the COP, divided by the drawing's COP factor f_a; the expected shed load is the expected shortfall of each month's states, summed over the months; and the annual cost adds the price c_a annualized by the capital recovery factor $\rho = 0.131\,\mathrm{yr}^{-1}$ (15 years at 10 %), the energy at $p_E = \mathrm{R\$}\,0.75/\mathrm{kWh}$ and the shed load at $p_S = \mathrm{R\$}\,1\,500/\mathrm{kWh}$:

$$E_a = \sum_{m=1}^{12} 730\,\frac{Q_{L,m} - Q_{e,m}}{\mathrm{COP}_m\, f_a}, \qquad S_a = \sum_{m=1}^{12} 730 \sum_{\mathbf{s}} P(\mathbf{s})\,\max(Q_{L,m} - \mathrm{Cap}_m(\mathbf{s}),\ 0),$$
$$L(a,\mathbf{z}) = \rho\, c_a + p_E\, E_a + p_S\, S_a, \qquad D = L(3{\times}400,\mathbf{z}) - L(2{\times}600,\mathbf{z}), \tag{10.6}$$

in kWh/yr and R$/yr, with $\mathrm{Cap}_m(\mathbf{s})$ the capacity of equation (10.1) at month m's wet bulb and economizer duty. A negative difference D means the three smaller units cost less over the year. The headroom of each drawing at the design day is its capacity with all units up, with its largest unit out, and with its two largest units out, less the design load,

$$H_{\mathrm{all}} = \sum_i \mathrm{Cap}_i + Q_e - 1\,000\,\mathrm{kW}, \qquad H_{-1} = \sum_i \mathrm{Cap}_i - \max_i \mathrm{Cap}_i + Q_e - 1\,000\,\mathrm{kW},$$
$$H_{-2} = \sum_i \mathrm{Cap}_i - \mathrm{Cap}_{(1)} - \mathrm{Cap}_{(2)} + Q_e - 1\,000\,\mathrm{kW}, \tag{10.7}$$

with $\mathrm{Cap}_{(1)} \geq \mathrm{Cap}_{(2)}$ the two largest units of the draw; the last two are maxima over the units, read jointly. For drawing 2×600, H_{-2} leaves the economizer alone.

Modeling choices at the graph. Every unit of either size fails and is repaired at the same rates as the existing 400 kW chillers: this lets the outage record of the existing units speak for units not yet built, and it fails if a 600 kW unit fails more or less often, or takes longer to repair, than a 400 kW one; the site has no record of the larger size (Section 10.6.4). The units fail independently given q: this gives the products of equation (10.1), and it fails under a shared cause such as a common electrical supply, the question of Chapter 15. The units share one q, which is uncertain: given the outage record all units move together with it, a common cause of the kind of Section 3.4.4 acting on what is known rather than on what happens. A new unit's ratio is the existing units' mean with a spread of 0.04, the spread of the existing units' as-built ratios, the same for both sizes: this makes the two drawings equally good units of one model, and it fails if the larger units are built to another tolerance. The economizer is on the drawing for both, with its handbook conductance, and it does not fail. The annual cost is read on the record year's months at today's loads: this compares the drawings on the plant as it runs now, and the load path is the question of Chapter 11.

The graph. The computation is drawn in Figure 10.1. The three boxes at the top are what the site already knows: the fit of the existing chillers and the plant efficiency, the economizer's handbook guess and the outage record. The two drawings in the middle take their units' ratios from the fit through equation (10.4); the enumeration below them takes the units' capacities, the economizer and q; the two scenario nodes on the left pin the loads and wet bulbs; and the four derived quantities at the bottom are the answers' raw material.

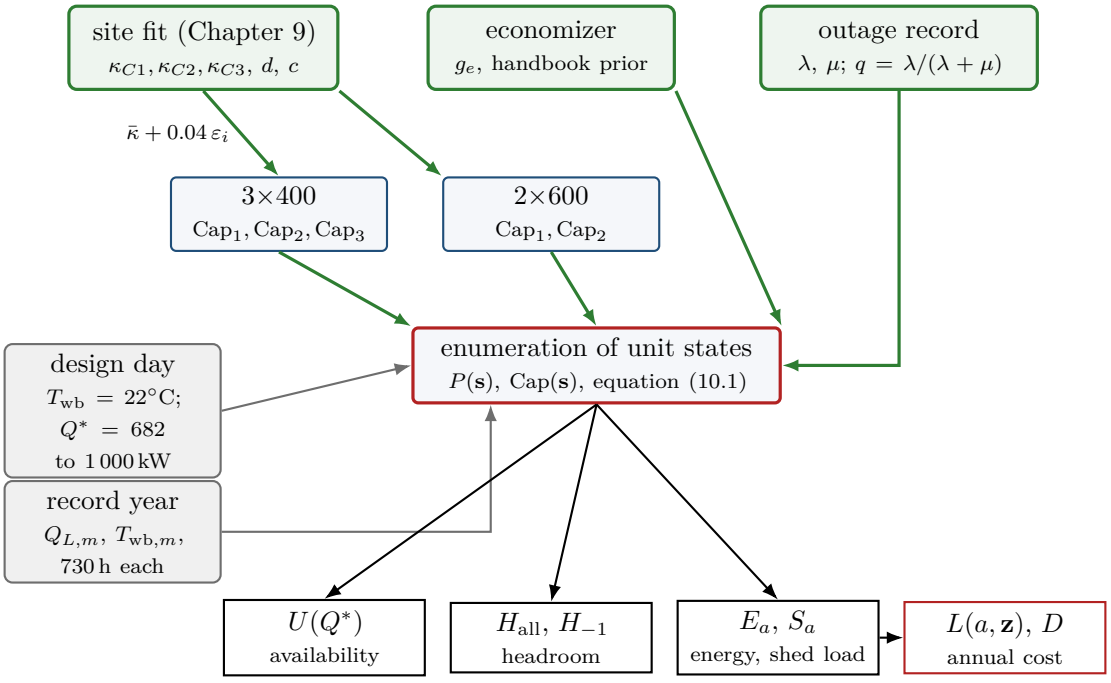


Figure 10.1: The comparison of the two drawings on the site. Green: the entries the site already knows, the fit of the existing chillers' as-built ratios $\kappa_{C1}, \kappa_{C2}, \kappa_{C3}$, their derate d and the plant efficiency c; the economizer's log-ratio g_e, which keeps its handbook prior; and the chillers' failure and repair rates from the outage record. Blue: the two drawings, whose new units take the existing units' mean ratio with a spread of 0.04, equation (10.4). Red: the enumeration of unit states, equation (10.1). Gray: the two scenario nodes, the design day at four loads and the twelve months of the record year. Bottom: the derived quantities, equations (10.2), (10.7) and (10.6); the annual cost adds the annualized price to the energy and the shed load.

Table 10.1: The questions asked of the two drawings. $\mathbf{y}$ is the site's operating record and capacity tests, the outage record and the priors of Chapter 9; U is the unavailability (10.2), H the headroom (10.7), and L and D the annual cost and its difference (10.6).

Planner's question	General form	Mathematical form	Category	Computation	Answered in
Two 600 kW units or three 400 kW?	$\arg\min_a$ $\mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$	$\mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$ for both drawings, and $\Pr(D < 0 \mid \mathbf{y})$	design	expected consequence	10.3.4
Which is more available, and at what load?	$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y})$ over an enumeration	$\mathbb{E}[U(Q^*) \mid \mathbf{y}]$ for $Q^* = 682$ to 1 000 kW; $\Pr(U_{3\times400} < U_{2\times600} \mid \mathbf{y})$	prediction under a threshold	posterior	10.3.1
What headroom does each leave?	$\Pr(f(\mathbf{z}) \leq f^* \mid \mathbf{y})$, quantiles	H_{all} and H_{-1} at 1 000 kW: median and 90 % band	estimation	posterior	10.3.2
How much does efficiency cost?	$\mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}]$ and quantiles of a derived quantity	$p_E\,(E_{3\times400} - E_{2\times600})$: mean and 90 % band	estimation	posterior	10.3.3
Is the difference real?	$\Pr(f(\mathbf{z}) > f^* \mid \mathbf{y})$	$\Pr(D < 0 \mid \mathbf{y})$, with the 90 % band of D	prediction under a threshold	posterior	10.3.5

10.1.3 The planner's questions (step 1)

The planner asks five questions of the two drawings, and step 1 writes each as a probability, a quantile or an expectation over the state of the site, with the unknown on the left of the bar and the evidence on the right. Table 10.1 lists them. The first column is the question in the planner's words, the drill question first; the second the general form of its category in the notation of Chapter 2, for a quantity $f(\mathbf{z})$ derived from the state and a level f^* of it; the third its mathematical form for the two drawings; the fourth its category; the fifth the computation of Section 2.3.6 that answers it; and the last the subsection where it is answered. The first is a design question read as an expected consequence: the drawing that costs less over the year. The second and the last are predictions under a threshold: which drawing is more available, at which load, and whether the cost difference has a sign the evidence settles. The third and the fourth are estimation.

10.1.4 The evidence as rows (step 2)

This chapter adds no reading. The evidence is the site's, attached in Chapter 9 and read here through its posterior. Its rows are of three kinds. The operating record gives two capacity tests per existing chiller at two wet bulbs and twelve monthly windows of chiller power, pipe pressure drop, pump power and feed power, eight entries of the site's state fitted on them. The outage record gives the chillers' count over four years, 13 failures in 12 unit-years with 527.4 hours of repair, as a count row on the failure rate and a duration row on the repair rate, each with a gamma prior. The priors complete the rows; among them the economizer's log-ratio g_e, which no window of the record sees, because the economizer is drawn and not built.

Which rows reach which answer. The availability at a load depends on the units' capacities, through which states carry it, and on q, through each state's probability. The capacities rest on the capacity tests, which fix the existing chillers' ratios and the derate, and through equation (10.4) the new units' mean ratio; and on the economizer's prior. The probabilities rest on the outage record alone. The energy rests on the plant efficiency c, which the monthly chiller

power fixes, and on the economizer, which nothing in the record fixes. So before any number, the record says where each band will come from: the shed load's from the outage record, the energy's from the economizer's handbook guess, and the headroom's from the capacity tests, the new units' spread and the economizer.

Modeling choices at the rows. The outage record of the existing 400 kW units is taken as the record of every unit of either drawing; the two drawings share it, and so share the same draw of q in every comparison. The record year's months are the annual scenario's months: their IT loads, from 580.6 to 660.0 kW, and their wet bulbs, from 10.2 to 19.8°C, are inputs of the record, not readings of the drawings. The prices, the COP factors, the energy price, the value of shed load and the capital recovery factor are stated planning inputs, printed where they are used; the chapter does not estimate them.

10.2 The posterior of the state (step 3)

All five questions of Table 10.1 are read from one posterior, the site's, which Chapter 9 computes. This section states what each answer takes from it and how the two drawings are evaluated on it; it does not refit.

10.2.1 The site's posterior

The state of the site has two parts. The first is the eight entries of the plant, $(\kappa_{C1}, \kappa_{C2}, \kappa_{C3}, d, g_e, c, l_k, l_e)$, with the three as-built ratios, the derate, the economizer's log-ratio, the plant efficiency and the pipe's and the pump's entries, fitted on the operating record with the normal equations of Chapter 4; its Laplace normal is $\mathcal{N}(\hat{\mathbf{z}}, \mathbf{\Sigma})$. Two facts of that fit carry into this chapter. The as-built ratios are known to about a fifth of their prior spread and the derate to about three fifths, and the plant efficiency c to a standard deviation of 0.0061; the economizer's g_e keeps its handbook prior, standard deviation 0.35, a ratio of 1.000 to its prior, because no row sees it. The second part is the chillers' failure rate λ and repair rate μ, with conjugate gamma posteriors from the outage record; Chapter 9 gives $q = \lambda/(\lambda + \mu)$ a posterior mean of 0.00554 and a 90 % band of 0.00277 to 0.00959, a factor of about three and a half between its ends. The pipe's and the pump's entries do not enter any answer of this chapter.

10.2.2 The new units and the paired draws

Every answer is a derived quantity of the state through the enumeration, which is a sum over states and has no useful gradient where a state's capacity crosses the load. So the answers are read from draws, and the reported value is the state route: the Laplace draws pushed through the exact enumeration. Draw i takes

$$\mathbf{z}_i \sim \mathcal{N}(\hat{\mathbf{z}}, \mathbf{\Sigma}), \quad \lambda_i,\ \mu_i \sim \text{their gamma posteriors}, \quad q_i = \frac{\lambda_i}{\lambda_i + \mu_i}, \quad \kappa_{ij} = \bar{\kappa}(\mathbf{z}_i) + 0.04\,\varepsilon_{ij}, \tag{10.8}$$

with $j = 1, 2, 3$ the new units' ratios of that draw. Drawing 3×400 uses all three, drawing 2×600 the first two. Both drawings are evaluated on the same draw: the same plant entries, the same q_i, the same economizer and the same new-unit ratios where they overlap. The draws are paired, so that what the drawings share, the economizer and the plant efficiency above all, moves both alike and cancels in their difference.

Each draw is pushed through the exact enumeration (10.1) of each drawing: at the design day for the four loads of the availability scenario, and month by month for the energy and the shed load of equation (10.6). A probability that compares the drawings, such as $\Pr(U_{3\times400}(Q^*) <$

$U_{2\times600}(Q^*) \mid \mathbf{y})$, is the share of draws in which the event holds. The draws are Chapter 9's, $N = 20\,000$ of them from seed 9.

10.2.3 The exact posterior by reweighted draws

The Laplace normal approximates the posterior of the plant entries, and the exact posterior is computed beside it as in Section 5.2.3: each draw $\mathbf{z}_i$ takes the weight of equation (5.16), the exact posterior over the normal that produced it, with the site's objective in place of hall B's. The rates need no weight: their gamma posteriors are exact and are drawn directly. The weights are nearly even: their effective sample size (ESS) is 19 913 of the 20 000 draws. Every reported answer has its exact counterpart, the same draws under the weights, and step 5 sets the two side by side.

10.3 The answers (step 4)

Every answer is read from the paired draws of Section 10.2.2, the reported value first and the exact posterior beside it. The availability comes first, because the cost comparison rests on it.

10.3.1 Availability against load

Planner's question. Which is more available, and at what load? The question asks for each drawing's unavailability at the design day, at loads from today's to the design load, and for the probability that the smaller units are the more available.

General form. A prediction under a threshold is a tail of a derived quantity under the posterior, equation (2.12),

$$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) \geq f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

here over an enumeration: the event is the plant's states falling short of the load, whose probability given the state is itself a sum.

Specialized form. The unavailability of drawing a at the load Q^* is equation (2.12) with the event $\mathrm{Cap}(\mathbf{s}) < Q^*$ over the units' states and over the state of the site, the inner sum the enumeration (10.2) and the outer integral the posterior,

$$\begin{aligned} \mathbb{E}\big[U_a(Q^*) \mid \mathbf{y}\big] &\approx \frac{1}{N}\sum_{i=1}^{N}\sum_{\mathbf{s}} P(\mathbf{s};\ q_i)\, \mathbf{1}\big[\mathrm{Cap}(\mathbf{s};\ \mathbf{z}_i) < Q^*\big], \\ \Pr\big(U_{3\times400} < U_{2\times600} \mid \mathbf{y}\big) &\approx \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[U_{3\times400,i} < U_{2\times600,i}\big], \end{aligned} \tag{10.9}$$

with $N = 20\,000$ paired draws. The first is the posterior mean of the loss-of-load probability; the second is equation (2.12) with $f = U_{2\times600} - U_{3\times400}$ and $f^* = 0$, the share of the posterior in which the smaller units are the more available. The exact posterior is the same sums under the weights.

Building blocks. The paired draws (10.8); the capacities (10.4) and the economizer (10.5) at 22°C; the enumeration (10.1); the weights of Section 10.2.3.

Table 10.2: The unavailability of the two drawings at the design day against the load, from 20 000 paired draws and from the same draws reweighted to the exact posterior, and the share of draws in which drawing 3×400 is the more available.

Load Q^* [kW]	U, 3×400 [–]	exact [–]	U, 2×600 [–]	exact [–]	$\Pr(U_{3\times400} < U_{2\times600})$ [–]
682	1.060×10^{-4}	1.060×10^{-4}	1.090×10^{-2}	1.091×10^{-2}	0.9967
800	1.056×10^{-2}	1.063×10^{-2}	1.104×10^{-2}	1.104×10^{-2}	0.3684
900	1.648×10^{-2}	1.648×10^{-2}	1.104×10^{-2}	1.104×10^{-2}	0.0015
1 000	1.650×10^{-2}	1.651×10^{-2}	1.104×10^{-2}	1.104×10^{-2}	0.0000

Table 10.3: The 90 % bands of the two drawings' unavailabilities at the design day, the 5 % and 95 % points over the 20 000 paired draws.

Load Q^* [kW]	3×400, 5 % [–]	3×400, 95 % [–]	2×600, 5 % [–]	2×600, 95 % [–]
682	2.296×10^{-5}	2.743×10^{-4}	5.282×10^{-3}	1.899×10^{-2}
800	4.717×10^{-5}	2.487×10^{-2}	5.531×10^{-3}	1.909×10^{-2}
900	8.229×10^{-3}	2.848×10^{-2}	5.531×10^{-3}	1.909×10^{-2}
1 000	8.285×10^{-3}	2.850×10^{-2}	5.531×10^{-3}	1.909×10^{-2}

The unavailability at four loads. Table 10.2 gives, for each load, the two drawings' unavailabilities by the draws and by the exact posterior, and the probability that the smaller units are the more available. Table 10.3 gives, for the same loads, the 5 % and 95 % points of each drawing's unavailability over the draws. The same unavailabilities are plotted against the load in Figure 10.2, on a logarithmic vertical axis: the red line is drawing 3×400, the blue line drawing 2×600, the bars beside each load their 90 % bands, the open markers the stated truth of the record, a check marker for Section 10.4.3; the shaded strip is the 90 % band of the capacity drawing 3×400 keeps with its largest unit out, and the dashed line its median, the availability crossover.

At today's load of 682 kW the rule of thumb holds, and by far: the smaller units' unavailability is 1.060×10^{-4} against 1.090×10^{-2}, about a hundred times lower, and they are the more available in 0.9967 of the draws. This is the second case of equation (10.3) against the first. Three units of 400 kW carry 682 kW with one unit out, so only a double outage fails them, about $3q^2$; two units of 600 kW do not, so any single outage fails them, about $2q$. The draws in which the smaller units lose, a share of 0.0033, can only be draws whose as-built units are weak enough that single outages of the 400 kW units no longer carry 682 kW, since otherwise $3q^2$ is below $2q$.

At the design load the rule of thumb fails. At 1 000 kW neither drawing carries the load with one unit out, both are in the first case of equation (10.3), and the unavailabilities are 1.650×10^{-2} for three units against 1.104×10^{-2} for two, a ratio of $1.650/1.104 = 1.49$, near the $3q/2q$ that the approximation predicts. The two units are the more available in every draw: $\Pr(U_{3\times400} < U_{2\times600}) = 0.0000$. The surprise of the comparison is this reversal. *The smaller units win on availability only at loads that one unit out can still carry*; the larger units' unavailability barely moves between today's load and the design load, from 1.090 to 1.104×10^{-2}, because neither of their states with one unit out carries any of these loads, while the smaller units' rises by a factor of about a hundred and fifty.

The crossover lies near 800 kW. There the smaller units' unavailability, 1.056×10^{-2}, is still just below the larger units' 1.104×10^{-2} on average, but they are the more available in only 0.3684 of the draws: 800 kW lies inside the band of the capacity that three units keep with their

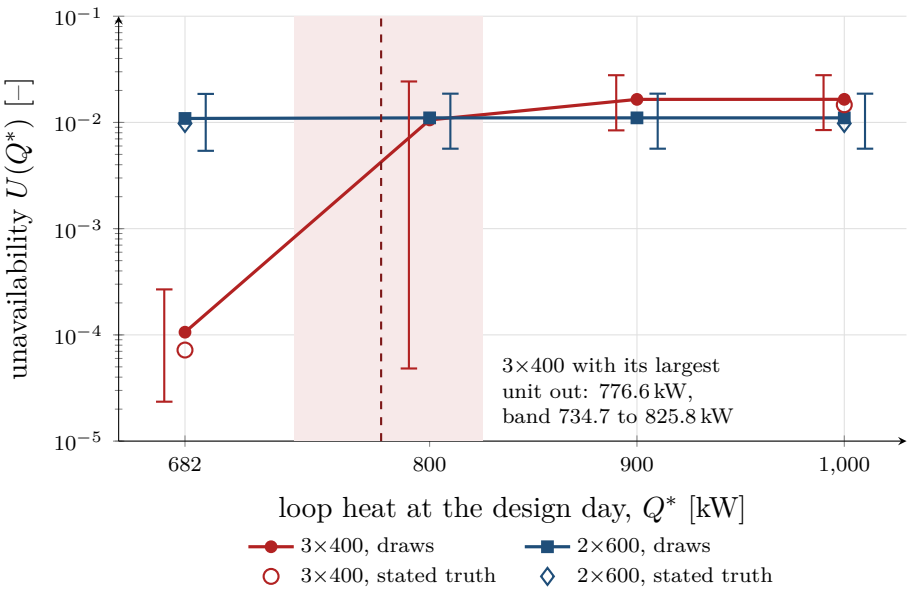


Figure 10.2: The unavailability of the two drawings at the design day against the loop heat, equation (10.9), by the paired draws (Table 10.2). Bars: the 90 % bands of Table 10.3, 3×400 drawn 10 kW left of each load and 2×600 10 kW right of it, for legibility. Open markers: the stated truth, a check marker. Shaded: the 90 % band of the capacity drawing 3×400 keeps at the design day with its largest unit out, the design load plus the headroom H_{-1} of Table 10.4; dashed: its median, the availability crossover.

largest unit out, whose median is 776.6 kW (Section 10.3.2), so in some draws one unit out still carries it and in most it does not. Its 90 % band in Table 10.3 shows the split: it runs from 4.717×10^{-5}, the double-outage level of today's load, to 2.487×10^{-2}, the single-outage level of the design load, and holds the larger units' whole band, $[5.531 \times 10^{-3},\ 1.909 \times 10^{-2}]$, inside it. By 900 kW the smaller units are the more available in 0.0015 of the draws, and their band, $[8.229 \times 10^{-3},\ 2.848 \times 10^{-2}]$, has moved up with them.

10.3.2 Headroom all up and with the largest unit out

Planner's question. What headroom does each leave? The question asks for each drawing's capacity over the design load at the design day, with all units running and with its largest unit out.

General form. An estimation question about a derived quantity asks for its posterior distribution, equation (2.9),

$$\Pr(f(\mathbf{z}) \le f^* \mid \mathbf{y}) = \int \mathbf{1}[f(\mathbf{z}) \le f^*]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

summarized by its median and its central 90 % interval, equation (2.11).

Specialized form. The question is equation (2.9) with $f = H_{\text{all}}$, $f = H_{-1}$ and $f = H_{-2}$ of equation (10.7), at $T_{\text{wb}} = 22°\text{C}$, for each drawing. The headroom with the largest unit out, or the two largest, is a minimum over the units of the capacity left, so it is read jointly from the paired draws, each taking the largest units of its own draw; no single unit's gradient carries it.

Table 10.4: The headroom of the two drawings over the 1 000 kW design load at the design day, equation (10.7), from the paired draws: medians and 90 % bands.

State	3×400, median [kW]	5 % [kW]	95 % [kW]	2×600, median [kW]	5 % [kW]	95 % [kW]
All up, H_{all}	158.3	104.0	218.4	158.4	96.5	225.8
Largest out, H_{-1}	−223.4	−265.3	−174.2	−407.7	−448.6	−359.6
Two largest out, H_{-2}	−592.8	−623.2	−550.9	−949.8	−972.1	−911.6

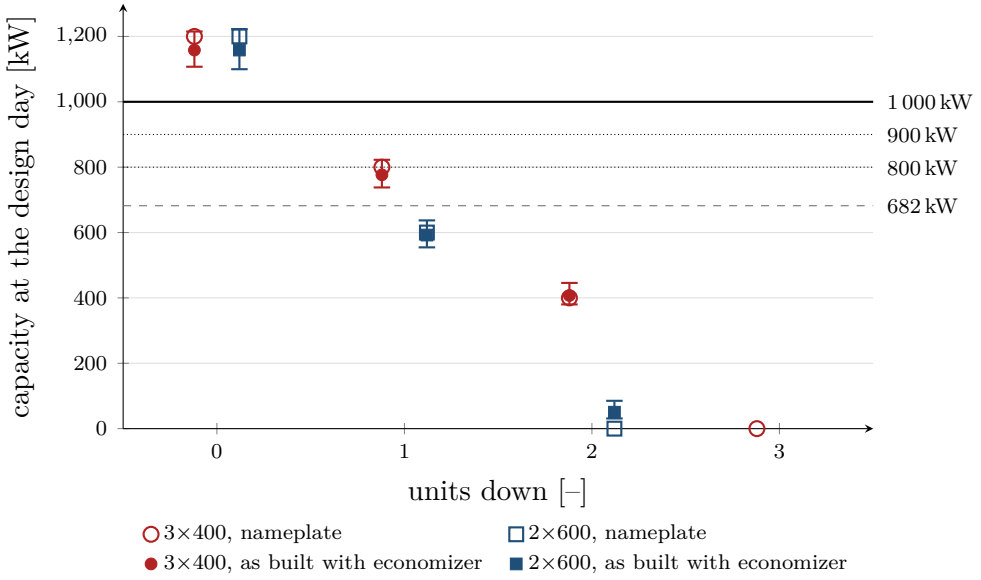


Figure 10.3: The capacity of the two drawings at the design day by the number of units down. Open markers: the nameplate of the units up, the design sheet's reading. Filled markers and bars: the as-built capacity with the drawn economizer, median and 90 % band from the paired draws, all units up, with the largest unit out and with the two largest out; the design load plus the headrooms of Table 10.4. With three units down the smaller units keep only the economizer, whose as-built value is not printed and is not drawn. Lines: the four loads of the availability scenario.

Building blocks. The paired draws (10.8); the capacities (10.4) and the economizer (10.5) at the design day.

The headroom. Table 10.4 gives the median and the 90 % band of the three headrooms for both drawings, one row per state. The same capacities are set against the four loads in Figure 10.3: the horizontal axis is the number of units down, the vertical axis the capacity at the design day; open markers are the design sheet's nameplate, filled markers with bars the as-built capacity with the drawn economizer, its median and its 90 % band, for the states the headroom reads; the horizontal lines are the four loads of the availability scenario.

All up, the two drawings leave the same headroom, 158.3 and 158.4 kW at the median: the same nameplate, the same mean as-built ratio and the same economizer. The as-built capacity, 1 158.3 kW with the economizer, sits below the 1 200 kW of nameplate, because the existing units' ratios, which the new units inherit, are below their rating. The larger units' band is the wider, [96.5, 225.8] against [104.0, 218.4] kW: the between-unit spread of 0.04 averages over three units in one drawing and over two in the other.

With the largest unit out neither drawing carries the design load: the headroom is −223.4 kW

for three units and −407.7 kW for two, both bands wholly below zero. Both are drawings of N units, with no redundancy at the design load, which is the question of Chapter 9 and not of this one. The figure shows why the availability reverses: the smaller units' one-out capacity, 776.6 kW, lies between today's load and the design load, and the larger units' one-out capacity, 592.3 kW, lies below all four loads. With the two largest units out the smaller units still keep one unit and the economizer, $1\,000 - 592.8 = 407.2$ kW at the median, band [376.8, 449.1] kW; the larger units keep the economizer alone, 50.2 kW, band [27.9, 88.4] kW, the economizer's handbook band.

10.3.3 The cost of efficiency

Planner's question. How much does efficiency cost? The question asks what the smaller units' lower part-load COP factor costs in energy over a year, with its band.

General form. The moments of a derived quantity under the posterior, equation (2.10),

$$\mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}] = \int f(\mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

with its central 90 % interval beside it.

Specialized form. The question is equation (2.10) with $f = p_E\,(E_{3\times400} - E_{2\times600})$ of equation (10.6). The two drawings carry the same duty at the same COP in every month and differ only in the factor, so on a paired draw

$$p_E\,(E_{3\times400} - E_{2\times600}) = p_E\,E_{2\times600}\Big(\frac{1.00}{0.97} - 1\Big) = 0.0309\,p_E\,E_{2\times600}, \tag{10.10}$$

a fixed share of the larger units' energy: the band of the extra cost is the band of the energy itself, scaled.

Building blocks. The paired draws (10.8); the economizer and the COP (10.5) at each month's wet bulb; the energy of equation (10.6).

The energy. The draws give the smaller units a chiller energy of 692 MWh/yr, 90 % band [431, 881], and the larger 672 MWh/yr, band [418, 854]. The extra energy cost of three 400 kW units is

$$p_E\,(E_{3\times400} - E_{2\times600}) = \text{R\$ } 15\,577/\text{yr}, \qquad 90\,\%\text{ band } [9\,688,\ 19\,814]\ \text{R\$/yr},$$

and the exact posterior gives the same mean, R$15,577/yr. The band is wide, more than a factor of two between its ends, and it is not the COP's: the record fixes the plant efficiency c to 0.0061. It is the economizer's. By equation (10.5) the economizer takes $25\,\mathrm{e}^{g_e}$ kilowatts per kelvin of the gap between the return water and the wet bulb, a gap of up to 14 K in the coolest months, so the chillers' duty over the year moves with e^{g_e}, and g_e is the handbook guess with its prior standard deviation of 0.35. The efficiency penalty is real in sign in every draw, since the factor is stated; its size is as uncertain as the economizer.

10.3.4 The annual cost

Planner's question. Two 600 kW units or three 400 kW? The question asks which drawing costs less over a year, counting its price, its energy and the load it sheds.

Table 10.5: The annual cost of the two drawings over the record year's months at today's loads, equation (10.6): means over the paired draws, with 90 % bands. Capital at the stated prices, R$4.8 and 4.2 million, times $0.131\,\mathrm{yr}^{-1}$; energy at R$0.75/kWh; shed load at R$1,500/kWh.

Part [unit]	3×400	2×600	3×400 minus 2×600
Capital [R$/yr]	628 800	550 200	78 600
Energy [R$/yr]	519 249	503 672	15 577, band [9 688, 19 814]
Expected shed load [kWh/yr]	50.3, band [4.8, 151.4]	173.7, band [25.8, 527.9]	−123.4
Expected shed load [R$/yr]	75 449	260 508	−185 059
Total [R$/yr]	1 223 498, band [966 050, 1 475 965]	1 314 380, band [947 040, 1 950 497]	−90 881, band [−475 245, 64 297]
Total, exact posterior [R$/yr]	–	–	−91 469

General form. A design decision read as an expected consequence chooses the action of least expected loss, equation (2.23),

$$\mathcal{R}(a) = \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}] = \int L(a, \mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \qquad a^* = \arg\min_a \mathcal{R}(a).$$

Specialized form. The question is equation (2.23) with the two actions 3×400 and 2×600 and the loss $L(a, \mathbf{z})$ of equation (10.6): the certain cost $\rho\, c_a$, the energy $p_E\, E_a$, and the expected shed load S_a, whose inner sum over the units' states is the enumeration and plays the part of $\sum_e C_e \Pr(e \mid \mathbf{z}, a)$, valued at p_S. The expected loss is the mean over the paired draws, and the difference

$$\mathbb{E}[D \mid \mathbf{y}] = \mathcal{R}(3{\times}400) - \mathcal{R}(2{\times}600) \approx \frac{1}{N} \sum_{i=1}^{N} D_i \tag{10.11}$$

is the mean of the paired differences; its band is read from their quantiles.

Building blocks. The paired draws (10.8); the enumeration (10.1) at each month's capacities and economizer; the loss (10.6); the stated prices, energy price, value of shed load and capital recovery factor.

The annual comparison. Table 10.5 gives the three parts of each drawing's annual cost and their total, with the expected shed load in kWh beside its value, and the difference of the totals in the last column. Each row is a part; the bands are the 90 % bands of the paired draws.

The expected annual costs are R$1,223,498 for three units and R$1,314,380 for two, and the mean of the paired differences is

$$\mathbb{E}[D \mid \mathbf{y}] = -\mathrm{R\$}\,90\,881/\mathrm{yr}, \qquad 90\,\%\ \text{band}\ [-475\,245,\ 64\,297]\ \mathrm{R\$/yr}, \tag{10.12}$$

with the exact posterior at −R$ 91 469/yr. The expected loss leans to the smaller units, and the parts say why. The larger units are cheaper to buy, by 628 800 − 550 200 = 78 600 reais a year, and cheaper to run, by R$15,577/yr; the smaller units shed less, 50.3 against 173.7 kWh/yr, worth R$75,449 against R$260,508 a year at the value of shed load. The expected shed load decides it, and it is the term the outage record carries: at today's loads the larger units shed on any single outage in the warmer months, when one 600 kW unit and the economizer fall short of the loop heat, and the smaller units only on a double outage, as equation (10.3) has it. The totals' own bands are wide, because each carries the energy's band; in the difference the energy moves both drawings alike and cancels, and what is left is chiefly the shed load's.

10.3.5 Whether the difference is real

Planner's question. Is the difference real? The question asks how sure the planner can be that the smaller units are the cheaper, not on average but in the plant that will be built.

General form. A prediction under a threshold, equation (2.12),

$$\Pr(f(\mathbf{z}) > f^* \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) > f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}.$$

Specialized form. The question is equation (2.12) with $f = -D$ of equation (10.6) and $f^* = 0$: the posterior probability that drawing 3×400 costs less over the year,

$$\Pr(D < 0 \mid \mathbf{y}) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[D_i < 0], \qquad \mathrm{se} = \sqrt{\hat{P}\,(1 - \hat{P})/N}, \tag{10.13}$$

with $\hat{P}$ the share and se its Monte Carlo standard error; the exact posterior sums the weights of the same draws.

Building blocks. The paired differences D_i of Section 10.3.4; the weights of Section 10.2.3.

The answer. By equation (10.13),

$$\Pr(3{\times}400 \text{ cheaper} \mid \mathbf{y}) = 0.585,$$

from the paired draws (0.5846) and from the exact posterior (0.5851), with a Monte Carlo standard error of $\sqrt{0.5846 \times 0.4154/20\,000} = 0.0035$. The draws also give nine percentiles of the difference: the 1 %, 5 %, 10 %, 25 %, 50 %, 75 %, 90 %, 95 % and 99 % points are −1 074 755, −475 245, −298 378, −118 602, −20 714, 30 299, 55 021, 64 297 and 76 701 reais a year. They are plotted in Figure 10.4 as the posterior cumulative distribution of the difference: the horizontal axis is the difference D in thousands of reais a year, negative where the smaller units are cheaper; the vertical axis the posterior probability that D lies below that value; the filled markers the nine percentiles and the share 0.5846 at zero, joined by straight lines with nothing smoothed between them; the dotted line that share; the solid line at zero, the dashed line the mean, and the bar the 90 % band.

The band says the difference is not yet real. The smaller units are the cheaper in 58 % of the posterior, and the larger in 42 %: the 90 % band of the difference runs from R$475,245 a year in the smaller units' favor to R$64,297 a year in the larger units', and contains zero well inside it. The mean favors the smaller units by R$90,881 a year, but the mean sits far from the middle of the posterior: the median is −R$20,714 a year, under a quarter of the mean, and the band reaches 384 000 reais a year below the mean and only 155 000 above it, a long tail on the side of the smaller units that runs to −R$1,074,755 at the 1 % point. The other side is short: the larger units can win by no more than what they save in capital and energy, R$78,600 a year plus an energy saving whose 95 % point is R$19,814, and the 99 % point of the difference is R$76,701. That tail is the posterior of q. Where the chillers fail often and repair slowly, the larger units' single-outage shed load is large and the smaller units win by hundreds of thousands of reais; where they fail seldom, the shed load of either drawing is small, and the larger units' lower price and better efficiency win by a few tens of thousands. The outage record of 13 failures in 12 unit-years does not yet say which of these plants the site has. The answer to the drill question, at this record, is a probability of 0.585, and it is stated as the answer: the evidence leans to three 400 kW units, and it does not decide between them and two of 600 kW.

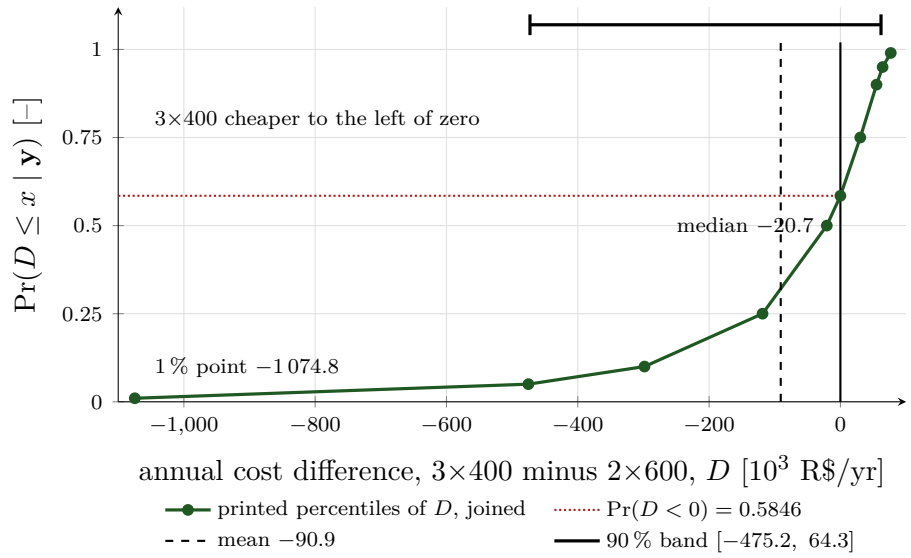


Figure 10.4: The posterior of the annual cost difference D of equation (10.6), in thousands of reais a year, as its cumulative distribution. Markers: the nine percentiles the draws give (1 % to 99 %) and the share 0.5846 of draws below zero; the lines between them are straight and carry no information of their own. Dotted: the curve's height at zero, $\Pr(D < 0) = 0.5846$, the probability that the smaller units are the cheaper. Solid line: zero. Dashed: the mean of the paired differences, −90.9, well left of the median −20.7. Bar: the 90 % band of the draws.

10.4 The checks (step 5)

The answers of Section 10.3 rest on the Laplace normal of the site's entries, on a finite set of draws, and on a stated truth that generated the record. Each is tested here.

10.4.1 The draws against the exact posterior

The reported answers are means and shares over the Laplace draws; the exact posterior reweights the same draws. In the unavailabilities of Table 10.2 the gaps, exact minus reported, are at most 7×10^{-5}, at 800 kW for the smaller units (1.063×10^{-2} against 1.056×10^{-2}), and 10^{-5} or zero elsewhere. The extra energy cost has the same mean, R$15,577/yr, by both. The mean annual difference moves by $-91\,469 - (-90\,881) = -588$ reais a year, under a percent of the mean, and the probability that the smaller units are cheaper by $0.5851 - 0.5846 = 0.0005$, smaller than its Monte Carlo standard error of 0.0035. The weights' effective sample size, 19 913 of 20 000, says the same: the reweighting barely moves any draw. The posterior of the site's entries is close enough to the Laplace normal that the comparison reads the same under either, as in Chapter 9.

10.4.2 The approximations of the enumeration

The two cases of equation (10.3) are tested on the exact enumeration. At the design load both drawings are in the first case, and the ratio of their unavailabilities, 1.49, is the 3/2 of $3q$ over $2q$ to within the small curvature $1 - (1 - q)^n$ adds. At today's load the larger units are in the first case, with 1.090×10^{-2}, close to twice the posterior mean of q; the smaller units are in the second, about a hundredth of that. Between the two cases, at 800 kW, the smaller units' unavailability lies between $3q^2$ and $3q$, as it must when one unit out carries the load in some draws and not in others. No approximation enters the reported numbers; the enumeration is exact in every draw, and the approximations serve only to read them.

Table 10.6: The checks of the comparison of the two drawings. The draws are 20 000 paired draws of the site's posterior; the exact posterior reweights them; the truth is the stated truth of the record. The gaps are taken from the printed values.

Route	$U_{3\times400}$, 682 kW [–]	$U_{2\times600}$, 682 kW [–]	$U_{3\times400}$, 1 000 kW [–]	$U_{2\times600}$, 1 000 kW [–]	$\Pr(D<0)$ [–]
Paired draws, reported	1.060×10^{-4}	1.090×10^{-2}	1.650×10^{-2}	1.104×10^{-2}	0.5846
Exact posterior	1.060×10^{-4}	1.091×10^{-2}	1.651×10^{-2}	1.104×10^{-2}	0.5851
Gap, exact minus reported	0	1×10^{-5}	1×10^{-5}	0	0.0005
Monte Carlo standard error	–	–	–	–	0.0035
90 % band, draws	$[2.296\times10^{-5}$, $2.743\times10^{-4}]$	$[5.282\times10^{-3}$, $1.899\times10^{-2}]$	$[8.285\times10^{-3}$, $2.850\times10^{-2}]$	$[5.531\times10^{-3}$, $1.909\times10^{-2}]$	–
Stated truth	7.201×10^{-5}	9.791×10^{-3}	1.465×10^{-2}	9.791×10^{-3}	–

10.4.3 The answers against the stated truth

The site's record is generated from a stated truth, and the last check compares the availabilities with it, the truth's as-built ratios and derate with the new units at the truth's mean ratio and the truth's outage rates. Table 10.6 collects the checks: each row is a route or the truth, the columns the two drawings' unavailabilities at today's load and at the design load, and the cost comparison where it applies. At today's load the truth gives the smaller units 7.201×10^{-5} and the larger 9.791×10^{-3}, both inside the reported 90 % bands; at the design load 1.465×10^{-2} and 9.791×10^{-3}, again inside the bands, and the ordering the answers report. Every reported value lies above the truth's, by about an eighth at the design load: the outage record's repairs averaged 40.6 hours against the truth's 36, so the posterior of q sits above the truth's, and every unavailability with it. The reversal itself is in the truth: the smaller units are the more available at 682 kW and the less available at 1 000 kW.

10.5 The report (step 6)

The rule of thumb is true at today's load and false at the design load. At 682 kW three 400 kW units are about a hundred times more available than two of 600 kW; at the design load of 1 000 kW the two are the more available, in every draw; and the crossover lies near 800 kW, the capacity the smaller units keep with one unit out. On cost, three 400 kW units are cheaper in 58 % of the posterior: a lean, not a decision.

10.5.1 The answers and their bands

Table 10.7 gives the answers in the order of Table 10.1. Each row is a question; the second column gives the reported answer with the exact posterior beside it, the third its band or the readings that qualify it.

10.5.2 The design-review entry

The design review's entry records the comparison as the enumeration reads it, not as the rule of thumb states it. At today's loop heat of 682 kW, three units of 400 kW are about a hundred times more available than two of 600 kW, 1.060×10^{-4} against 1.090×10^{-2}, because one of them can be out without shedding. At the design load of 1 000 kW the two are the more available, 1.104×10^{-2} against 1.650×10^{-2}, because neither drawing carries that load with a unit out and three units have more ways to fail. The ordering reverses near 800 kW. On the year at

Table 10.7: The report for the two drawings, at the design day for availability and headroom and over the record year's months at today's loads for energy and cost. Reported values by the paired draws of the site's posterior; the exact posterior beside each where it differs.

Question	Answer	Band
Two 600 kW units or three 400 kW?	expected cost R$1,223,498/yr for 3×400 against R$1,314,380/yr for 2×600; difference −R$90,881/yr (exact: −R$91,469/yr)	difference −R$475,245 to +R$64,297 a year, central 90 %; a lean to 3×400
Which is more available, and at what load?	3×400 at 682 kW, 1.060×10^{-4} against 1.090×10^{-2}, in 0.9967 of the draws; 2×600 at 1 000 kW, 1.104×10^{-2} against 1.650×10^{-2}, in every draw	crossover near 800 kW, where 3×400 is the more available in 0.3684 of the draws; exact values within 7×10^{-5}
What headroom does each leave?	all up, 158.3 and 158.4 kW; largest unit out, −223.4 and −407.7 kW	all up [104.0, 218.4] and [96.5, 225.8] kW; unit out [−265.3, −174.2] and [−448.6, −359.6] kW
How much does efficiency cost?	R$15,577/yr more energy for 3×400 (exact: the same)	R$9,688 to R$19,814 a year; the economizer's guess
Is the difference real?	Pr(3×400 cheaper) = 0.585 (exact: 0.585)	not at this record: the band of the difference contains zero

today's loads, the smaller units' expected cost is lower by R$90,881, but they are the cheaper in only 58 % of the posterior, and the entry records that probability, 0.585, as the answer: a lean to three 400 kW units that the evidence does not yet make a decision.

Two consequences follow, and the entry records both. The choice turns on where the plant will spend its life, so it is made on the expected load profile: Chapter 11's load path, on which the plant serves loads from today's 682 kW up to the design load, which the path reaches in year 2.81, and the comparison read along that path rather than at one load. And the term that decides the cost is the expected shed load, whose band is the outage record's, so before committing the review buys the outage evidence that narrows q, which Chapter 18 values.

Common mistake

Choosing the smaller units because they are more available. They are, below the load one unit out can carry; above it they are less available, and the design load is above it. Read the unavailability at each load the plant will serve, and read the cost difference with its band.

10.6 The next measurement (step 7)

The report leaves a cost difference whose sign the evidence does not settle. A further measurement is one more row of the evidence, and what decides whether it is worth buying is which entry it narrows.

10.6.1 What decides the comparison

The difference has three parts, and they carry different bands. The capital, 78 600 reais a year in favor of the larger units, is stated and has no band. The energy, R$15,577 a year in their favor, has the economizer's band, but it moves both drawings alike, and only its fraction 0.0309 of equation (10.10) enters the difference. The expected shed load is the term with the band that crosses zero: R$185,059 a year in favor of the smaller units at the mean (260 508 − 75 449), and in its band the outage record's. The threshold is simple to state. The smaller units are the

cheaper when their saving in shed load exceeds what they cost more in capital and energy, at the mean

$$p_S\,(S_{2\times600} - S_{3\times400}) > 78\,600 + 15\,577 = 94\,177 \text{ R\$/yr},$$
$$S_{2\times600} - S_{3\times400} > 94\,177/1\,500 = 62.8 \text{ kWh/yr},$$

and by equation (10.3) the larger units' shed load grows as q while the smaller units' grows as q^2, so the saving grows with q and the threshold is a threshold on q. Its band is the band of q of Section 10.2.1, a factor of three and a half, made of the failure rate λ and the mean time to repair. Two kinds of evidence would narrow it: more outage history, more unit-years of failures and repair hours, which narrows both rates; and for the energy, a test of the economizer's conductance once it is built, or a survey of the tower and the exchanger before it is.

10.6.2 The recommendation

No candidate is computed in this chapter. The measurement that matters is the one that narrows q: the outage record, through its repair hours as much as through its failure count, since the mean time to repair enters q as strongly as the failure rate does. The economizer test narrows the energy, which matters less to the difference than to either total. Chapter 18 computes the preposterior over the outage record and the economizer test and ranks them by the width each takes off the quantity that decides the build, here the probability that the smaller units are the cheaper.

10.6.3 What the analysis showed

The design sheet's rule of thumb is right at today's load and wrong at the design load, and the enumeration says why: unit size decides which states carry a load, and the number of units decides how many ways the plant can fail. Below the capacity that three 400 kW units keep with one unit out, about 777 kW at the median, they fail only on a double outage and are about a hundred times more available than two of 600 kW; above it both drawings fail on any single outage, and three units fail more often than two, by close to the ratio of three to two. The headroom with all units up is the same for both, and with the largest unit out neither carries the design load; the difference is in where the one-out capacity falls among the loads the plant will serve. The efficiency penalty of the smaller units is certain in sign and uncertain in size, R$15,577 a year with a band set by the economizer's handbook guess. The annual cost leans to the smaller units by R$90,881 a year at the mean, through the shed load they avoid at today's loads, and the probability that they are the cheaper is 0.585: the difference is not yet real. The band is the outage record's, and so is the decision.

10.6.4 Limits of the model

One outage record for two sizes. The units of both drawings take the existing 400 kW chillers' failure and repair rates. A 600 kW unit may fail at another rate, and its repairs, on a larger compressor, may take longer; the comparison of shed load, which decides the cost, would move with either. The site has no record of the larger size.

Independent failures given q. The units fail independently given their shared unavailability. A shared cause, such as one electrical supply or one controls panel for all units, would correlate the outages and weaken the smaller units' advantage at today's load, where they rely on a double outage being rare; Chapter 15 takes up shared paths.

The year at today's loads. The annual cost is read on the record year's months at today's loads, between 602.6 and 682.0 kW of loop heat, where the smaller units' advantage is largest. On the load path the plant reaches the design load in under three years, and above the crossover the larger units shed less; the comparison along the path is Chapter 11's, and the year-round expected unserved energy Chapter 23's.

The economizer on the drawing. Both drawings carry the economizer with its handbook conductance, and it never fails. The energy band is the guess's band, and an economizer that underperforms would raise both drawings' energy and the shed load of the larger units in the cooler months, where it lets one 600 kW unit carry the load.

Stated prices and a stated value of shed load. The prices, the COP factors, the energy price, the value of shed load and the capital recovery factor are stated inputs with no band. The value of shed load multiplies the term that decides the comparison; at a lower value the larger units would be favored, and at a higher one the smaller.

The constructed record. The site's operating record and outage record are generated from a stated truth with stated errors and rates, not taken from a plant, and the comparison with that truth in Section 10.4.3 is a check of the method on its own model, not of the model against a real site.

The next chapter. Chapter 11 lets the load grow along its path and asks when the plant must be extended, with its lead time.

10.7 What to retain

Definition 10.1 (Availability of a drawing against load)**.** The unavailability of a drawing of n units at a load Q is the summed probability of the units' states whose capacity, with the economizer, falls short of the load, $U(Q) = \sum_{\mathbf{s}} P(\mathbf{s})\,\mathbf{1}[\mathrm{Cap}(\mathbf{s}) < Q]$, averaged over the posterior of the units' capacities and of their unavailability q. For a small q it is about nq at a load no single outage leaves carried and about $\binom{n}{2}q^2$ at a load every single outage leaves carried. Two drawings are compared on paired draws of one posterior, and the probability that one is the better is the share of draws in which it is.

Four conclusions carry beyond this site. The availability of a drawing is a function of the load, not a property of the drawing: smaller units are more available below their one-unit-out capacity and less available above it, and the crossover is a derived quantity of the as-built units with a band of its own. Capacity and availability are two readings of one enumeration and do not stand in for each other: the two drawings leave the same headroom with all units up and differ by a factor of a hundred in availability at today's load. A comparison of two drawings is read on paired draws, so that what they share cancels and the band of the difference is the band of what differs. And the answer to whether a difference is real is a probability, stated as the answer: here 0.585, a lean that the outage record, not the price sheet, will settle.

Table 10.8 lists the terms of the chapter with their expressions, units and the subsection that defines each.

Table 10.8: The terms of unit sizing and where each is defined. Units for the site.

Term	Expression	Units	Meaning	Defined in
State probability	$P(\mathbf{s}) = \prod_i q^{1-s_i}(1-q)^{s_i}$	–	one combination of units up and down	10.1.1
Unavailability	$U(Q) = \sum_{\mathbf{s}} P(\mathbf{s})\,\mathbf{1}[\mathrm{Cap}(\mathbf{s}) < Q]$	–	loss-of-load probability at a load	10.1.1
Two cases	$U \approx nq$; $U \approx \binom{n}{2}q^2$	–	single outages fail or carry	10.1.1
Unit capacity	$N_i\,\mathrm{e}^{\kappa_i}(1 - d(T_{\mathrm{wb}} - 22))$	kW	as built, derated	10.1.2
New unit's ratio	$\bar{\kappa} + 0.04\,\varepsilon_i$	–	the existing units' model	10.1.2
Economizer duty	$25\,\mathrm{e}^{g_e}\max(T_{\mathrm{wr}} - T_{\mathrm{wb}}, 0)$	kW	handbook guess, drawn	10.1.2
Headroom	H_{all}, H_{-1}, H_{-2}, eq. (10.7)	kW	all up; one or two largest out	10.1.2
Annual cost	$\rho\,c_a + p_E E_a + p_S S_a$	R$/yr	capital, energy, shed load	10.1.2
Posterior and answers				
Paired draws	eq. (10.8), both drawings on one draw	–	what is shared cancels	10.2.2
Comparison probability	$\Pr(U_{3\times400} < U_{2\times600} \mid \mathbf{y})$	–	share of draws	10.3.1
Availability crossover	capacity with the largest unit out	kW	where the ordering reverses	10.3.1
Extra energy cost	$0.0309\,p_E\,E_{2\times600}$	R$/yr	the COP factor's price	10.3.3
Expected loss	$\mathcal{R}(a) = \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$	R$/yr	the design decision	10.3.4
Reality of a difference	$\Pr(D < 0 \mid \mathbf{y})$ with the band of D	–	whether the sign is settled	10.3.5

10.8 Exercises

Exercise 10.1. *Formulation.* Write the planner's question "at 900 kW, are three 400 kW units more available than two of 600 kW?" as a probability over the site's state, name the event inside it, and state its answer. ($\Pr(U_{3\times400}(900) < U_{2\times600}(900) \mid \mathbf{y})$, the general form (2.12) with $f = U_{2\times600} - U_{3\times400}$ and $f^* = 0$; each U is the enumeration (10.2) at 900 kW and 22°C on one paired draw. The answer is the share of draws in which the event holds, 0.0015.)

Exercise 10.2. *Derivation.* From equation (10.1), derive the two cases of equation (10.3) for $n = 3$ and $n = 2$, and check the ratio of the two drawings' unavailabilities at the design load. (With every single outage failing the load, $U = 1 - (1-q)^n = nq - \binom{n}{2}q^2 + \ldots \approx nq$. With every single outage carrying it and every double failing it, $U = \binom{n}{2}q^2(1-q)^{n-2} + q^n \approx \binom{n}{2}q^2$, which for $n = 3$ is $3q^2$. At 1 000 kW both are in the first case, so the ratio is near $3q/2q = 1.5$; the draws give 1.650×10^{-2} and 1.104×10^{-2}, a ratio of 1.49.)

Exercise 10.3. *Derivation.* Show that the extra energy of the smaller units is a fixed share of the larger units' energy on every paired draw, and say why its band is the economizer's. (Both drawings carry the same duty $Q_{L,m} - Q_{e,m}$ at the same COP_m, so $E_{3\times400} = E_{2\times600}\,(1.00/0.97)$ and the difference is $0.0309\,E_{2\times600}$. The energy's band comes from $Q_{e,m} = 25\,\mathrm{e}^{g_e}(24 - T_{\mathrm{wb},m})$, since g_e keeps its handbook prior while c is fixed by the record. The extra cost is R$15,577/yr, band R$9,688 to R$19,814.)

Exercise 10.4. *Application.* Give the expected shed-load saving at which the smaller units break even, at the mean capital and energy difference, and compare it with the mean saving. (They cost $78\,600 + 15\,577 = 94\,177$ reais a year more in capital and energy, which at R$1,500/kWh is 62.8 kWh/yr of shed load. The mean saving is $173.7 - 50.3 = 123.4$ kWh/yr, above the

break-even, which is why the mean difference favors them by R$90,881/yr; but the saving is proportional to q and its band reaches below the break-even, so $\Pr(D < 0 \mid \mathbf{y}) = 0.585$.)

Exercise 10.5. *Application.* A reviewer proposes to compare the drawings at the design load alone, since that is what the plant is drawn for. State what the comparison would conclude on availability and on headroom, and what it would miss. (At 1 000 kW the larger units are the more available in every draw, 1.104×10^{-2} against 1.650×10^{-2}, and neither drawing carries the load with its largest unit out, at -223.4 and -407.7 kW of headroom. It would miss that at today's load the smaller units are about a hundred times more available, and that the shed load at today's loads is what makes them the cheaper in a share 0.585 of the posterior.)

Chapter 11

Expansion Timing

11.1 When the fourth chiller has to be ordered

The site's chilled-water plant has three chillers of 400 kW nameplate at the design day, and the two halls it serves carry 660 kW of IT load today, which with the halls' other heat puts 682.0 kW on the loop. The planning brief says the IT load grows 15 % a year, and a fourth chiller takes 40 weeks from order to commissioning, 0.767 of a year. The design sheet gives two dates for the order, both from nameplate. Its N+1 rule asks that any two units carry the load, 800 kW of nameplate; the load path breaks that rule in month 15, year 1.25, so the sheet says to order by month 5.8. Its other date lets the three units run all up to their 1,200 kW of nameplate; the path reaches that in month 50, year 4.17, and the order falls due in month 40.8.

The two dates are three years apart, neither has a band, and neither says what the planner will accept. The planner does say it: a shortfall on the design day is tolerated with a probability up to 5 %, and not beyond. The question of this chapter is the drill question in the planner's words,

When does the next unit have to be ordered so that it is in before the probability of a shortfall passes 5 %?

It is answered with the posterior of the site that Chapter 9 fits and the enumeration of unit states it builds, carried along the load path as Chapter 6 carried hall B's binding load. What is new is the event read at each node of the path: not a constrained entry against its limit, but a shortfall in some state of the units, weighted by that state's probability, and a date read where the averaged probability first reaches the planner's tolerance.

Jargon: shortfall, loss-of-load probability, lead time, order date, tolerance, unavailability, state enumeration, as-built capacity, all-up capacity, redundancy-risk plateau, crossing month, per-draw crossing month

A *shortfall* on the design day is the event that the chillers that are up cannot carry the heat on the loop; its probability on that day is the design day's *loss-of-load probability*. The *lead time* ℓ is the time from the order of a unit to its commissioning, and the *order date* is the date by which the order must be placed for the unit to be in service by a given month. The *tolerance* α is the shortfall probability the planner accepts, here 0.05. A unit's *unavailability* q is the long-run share of time it is down for repair, $q = \lambda/(\lambda+\mu)$ with λ its failure rate and μ its repair rate. A *state enumeration* lists every combination of units up and down with its probability and its capacity. A chiller's *as-built capacity* is what it delivers as installed, which may differ from its nameplate; the *all-up capacity* is the sum over the units with all of them running. The *redundancy-risk plateau* is the stretch of the path over which the all-up state carries the load and no state with a unit down does, so that the shortfall probability equals the probability that any unit is down. The *crossing month* $m^\dagger$ is the first month in which the shortfall probability reaches the tolerance, and the *per-draw crossing month* is the same month read for one state of the posterior, as if the plant's capacity were known.

Key idea

The order date is a crossing along a scenario chain, less the lead time. At each month of the load path the shortfall probability is the enumeration of unit states, which carries the chance that a unit is down, averaged over the posterior of the as-built capacities and the outage rates. Each state of the posterior crosses the tolerance in its own month, where the load passes its all-up capacity, and those months make the date's band. A state that has crossed there contributes a shortfall probability of one to the average, so the averaged probability reaches 5 % no later than the month by which 5 % of the states have crossed: the reported date is the early end of its own band, by construction.

11.1.1 A shortfall date along a load path

The quantity the chapter owes is a month, the order date m_o, and it is built from the top down. The order date is the crossing month less the lead time,

$$m_o = m^\dagger - \ell, \qquad m^\dagger = \min\{m : P_m \geq \alpha\}, \tag{11.1}$$

in months from today, with $\ell = 12 \times 40 \times 7/365 = 9.2$ months and $\alpha = 0.05$. The crossing month is the first month whose shortfall probability P_m reaches the tolerance. That probability is the posterior expectation of the shortfall probability given the plant,

$$P_m = \mathbb{E}[\pi_m(\mathbf{z}, q) \mid \mathbf{y}] = \int \pi_m(\mathbf{z}, q)\, p(\mathbf{z}, q \mid \mathbf{y})\, \mathrm{d}\mathbf{z}\, \mathrm{d}q, \tag{11.2}$$

with $\mathbf{z}$ the site's state and q the chillers' unavailability. The shortfall probability given the plant, π_m, is the enumeration of the three units' states $s = (s_1, s_2, s_3)$, $s_i = 1$ for a unit up and 0 for a unit down: each state's probability times the indicator that its capacity falls short of month m's loop heat L_m,

$$\begin{aligned} \pi_m(\mathbf{z}, q) &= \sum_{s \in \{0,1\}^3} p(s \mid q)\, \mathbf{1}[C_s(\mathbf{z}) < L_m], \\ p(s \mid q) &= \prod_{i=1}^{3} q^{1-s_i}\,(1-q)^{s_i}, \qquad C_s(\mathbf{z}) = \sum_{i=1}^{3} s_i\, C_i(\mathbf{z}), \end{aligned} \tag{11.3}$$

the units taken as failing independently, each with the same q. The capacity of unit i is its as-built capacity at the design day. On the site plant a chiller's capacity is its nameplate N_i scaled by its as-built log-ratio κ_i and derated by d per kelvin of wet bulb above the rating condition T_{rate}; the chillers are rated at the design day's wet bulb, 22°C, so on that day the derate drops out,

$$C_i(\mathbf{z}) = N_i\,\mathrm{e}^{\kappa_i}\,(1 - d\,(T_{wb} - T_{\text{rate}})) = 400\,\mathrm{e}^{\kappa_i}\,\text{kW at } T_{wb} = T_{\text{rate}} = 22^{\circ}\text{C}, \quad C_{\text{all}}(\mathbf{z}) = \sum_{i=1}^{3} C_i(\mathbf{z}). \tag{11.4}$$

The loop heat is the load path of equation (6.1), read monthly, with the halls' other heat Q_o added,

$$L_m = Q_{\text{IT},0}\,(1+\gamma)^{m/12} + Q_o, \qquad m = 0, 1, \ldots, 72, \tag{11.5}$$

in kW, with $Q_{\text{IT},0} = 660\,\text{kW}$, $\gamma = 0.15\,\text{yr}^{-1}$ and $Q_o = 682.0 - 660 = 22\,\text{kW}$. Nothing in the state changes with the month: as in Chapter 6 the scenario node moves along the path, and the posterior stays where the fit put it.

The regimes of the enumeration. The enumeration has the shape of a staircase in the load. A single unit never carries the loop heat, since $400\,\mathrm{e}^{\kappa_i}$ is far below L_0; so a state with two units down always falls short, and the states that matter are the all-up state and the three states with one unit down. As the load rises it passes first the smallest one-unit-out capacity, $C_{\text{all}} - \max_i C_i$, then the largest, $C_{\text{all}} - \min_i C_i$, then the all-up capacity, and

$$\pi_m(\mathbf{z}, q) = \begin{cases} 3q^2(1-q) + q^3, & L_m \le C_{\text{all}} - \max_i C_i, \\ 1 - (1-q)^3 - k_m\,q\,(1-q)^2, & \text{between, with } k_m \text{ one-out states carrying,} \\ 1 - (1-q)^3, & C_{\text{all}} - \min_i C_i < L_m \le C_{\text{all}}, \\ 1, & L_m > C_{\text{all}}. \end{cases} \tag{11.6}$$

The third line is the redundancy-risk plateau: every state with a unit down falls short and the all-up state carries, so the shortfall probability is the probability that some unit is down, nearly $3q$. The fourth line is the step to one: once the load passes the all-up capacity no state carries it.

The month each state of the plant crosses. For one state $(\mathbf{z}, q)$ of the posterior the shortfall probability crosses the tolerance at the first step of the staircase that reaches α. The plateau reaches 0.05 only if $1 - (1-q)^3 \ge 0.05$, that is $q \ge 1 - 0.95^{1/3} = 0.017$, about three times the mean unavailability the outage record gives, $\bar{q} = 0.00554$ (Chapter 9). Below that, the state crosses at the step to one, the first month in which the load passes its all-up capacity,

$$m^{\dagger}(\mathbf{z}, q) = \min\{m : \pi_m(\mathbf{z}, q) \ge \alpha\} = \min\{m : L_m > C_{\text{all}}(\mathbf{z})\} = \lfloor \tau(C_{\text{all}}(\mathbf{z})) \rfloor + 1,$$
$$\tau(C) = \frac{12\,\ln\big((C - Q_o)/Q_{\text{IT},0}\big)}{\ln(1+\gamma)}, \tag{11.7}$$

with $\tau(C)$ the month, counted continuously, in which the path reaches the loop heat C: equation (6.3) in months and with the other heat added. For every state whose plateau stays below the tolerance the per-draw crossing month is an increasing function of the all-up capacity alone, so its quantiles are the all-up capacity's quantiles mapped, as Chapter 6 mapped the binding load's to years; the outage rates set the height of the plateau and not the month. A state with $q \ge 0.017$, a chiller fleet unreliable enough that its plateau alone passes 5 %, crosses early, in the first months of the path, as soon as the load passes its one-unit-out capacities.

The reported month against the per-draw months. Write $F_m = \Pr(m^\dagger(\mathbf{z}, q) \le m \mid \mathbf{y})$ for the share of the posterior that has crossed by month m. Take first every state's plateau below the tolerance. A state that has crossed then has $L_m > C_{\text{all}}$ and contributes $\pi_m = 1$ to the average of equation (11.2); a state that has not contributes the plateau or less, never below zero. Hence

$$P_m = F_m + (1 - F_m)\,\mathbb{E}[\pi_m \mid m^\dagger > m,\ \mathbf{y}] \ \ge\ F_m, \qquad \text{so} \qquad F_m \ge \alpha \ \Rightarrow\ P_m \ge \alpha \ \Rightarrow\ m^\dagger \le m^\dagger_\alpha, \tag{11.8}$$

with $m^\dagger_\alpha$ the α-quantile of the per-draw crossing month. The reported month can be no later than the 5 % point of the per-draw months. It can be earlier: on the plateau the states that have not crossed contribute about $3q$, and the outage record puts the posterior of q apart from that of the capacities, so the expectation in equation (11.8) is close to the plateau's mean $\bar{p}$ there, and $P_m \ge \alpha$ needs only $F_m \ge (\alpha - \bar{p})/(1 - \bar{p})$, a share below α. A state that crosses on its plateau contributes at least α but not one, so the bound can fail by at most the share of such states; on the site they are 0.0007 of the posterior (Section 11.3.3), far too few to move a month.

> **Principle 11.1.** *On a known load path the shortfall probability of a month is the enumeration of unit states averaged over the posterior of the plant. Each state of the posterior crosses the tolerance where the load passes its all-up capacity, unless its redundancy risk alone exceeds the tolerance; the reported month, the first in which the average reaches the tolerance, is never later than the tolerance's quantile of the per-draw months.*

11.1.2 The chiller plant as it stands (step 0)

The plant is the site of Chapter 9: the two halls on one chilled-water loop, served by three chillers C1–C3 of 400 kW nameplate at the design day, one pump, a shared supply pipe and a power feed. The chillers' as-built capacities are carried by their log-ratios κ_{C1}, κ_{C2}, κ_{C3} and the derate d; the site's state is the eight entries of Chapter 9,

$$\mathbf{z} = (\kappa_{C1},\ \kappa_{C2},\ \kappa_{C3},\ d,\ g_e,\ c,\ l_k,\ l_e)^{\mathsf{T}}, \tag{11.9}$$

with g_e the economizer's log-conductance, c the chiller plant's log-efficiency, l_k the shared pipe's log flow coefficient and l_e the pump's log-efficiency. To these the outage record adds the chillers' failure rate λ and repair rate μ, read through the unavailability $q = \lambda/(\lambda + \mu)$.

Modeling choices at the plant. The plant is the plant as it stands: three chillers, without the waterside economizer, which is drawn and not built. This is what makes the date the date for the plant the site has, and it fails if the economizer is built before the crossing; Section 11.4.3 builds it. The shortfall event is the chillers' alone: the pump, the pipe and the feed are held to carry what the chillers deliver. This keeps the enumeration to the units the order is for, and it fails where another element binds first; the shared pipe is checked in Section 11.4.3 and taken up in Chapter 13. The event is read on the design day of each year, wet bulb 22°C, the condition the planner's tolerance is stated for, not over the hours of the year.

The scenario chain. Each month $m = 0, \dots, 72$ of the path (11.5) is a scenario node at the design day, with the loop heat pinned at L_m: the chain of Chapter 6, here for the site and read monthly rather than yearly. At each node the three units are enumerated by equation (11.3). The month grid is a choice: it dates the crossing to a whole month, the step at which a plan can act, and it rounds every date up to the month's end.

Table 11.1: The questions asked of the site's chiller plant on its load path. $\mathbf{y}$ is the site's operating record, its capacity tests and its outage record; π_m is the enumeration (11.3); $m^\dagger(\mathbf{z}, q)$ is the per-draw crossing month (11.7); ℓ is the lead time.

Planner's question	General form	Mathematical form	Category	Computation	Answered in
When does the next unit have to be ordered so that it is in before the probability of a shortfall passes 5 %?	first crossing of a tail along a scenario chain	$\min\{m : P_m \geq 0.05\} - \ell$	prediction under a threshold	posterior	11.3.2
How does the shortfall probability grow along the path?	$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y})$ at each node	$P_m = \mathbb{E}[\pi_m \mid \mathbf{y}]$, month by month	prediction under a threshold	posterior	11.3.1
How sure is the date?	quantiles of a derived quantity	$m^\dagger(\mathbf{z}, q)$: median and 90 % band	estimation	posterior	11.3.3
What sets the date?	moments of a derived quantity	the all-up capacity's band against the plateau of one unit out	estimation	posterior	11.3.4

11.1.3 The planner's questions (step 1)

The planner asks four questions of the plant on its load path, and step 1 writes each as a probability, a quantile or a moment over the posterior of the site's state and the outage rates. Table 11.1 lists them, in the columns of Chapter 5: the question in the planner's words, the drill question first; its general form in the notation of Chapter 2; its mathematical form for the site; its category; the computation of Section 2.3.6 that answers it; and the subsection where it is answered. The first two are predictions under a threshold, the first read as the month in which the tail along the path first reaches the tolerance. The last two are estimation: the band of the date, and what the date is made of.

11.1.4 The evidence as rows (step 2)

Step 2 adds no row. The evidence is that of Chapter 9, in two records that act on separate entries. The operating record, two capacity tests per chiller at two wet bulbs and twelve monthly windows of chiller, pump and feed power and of the pipe's pressure drop, is a set of rows on the eight entries of $\mathbf{z}$; the outage record, the chillers' failure counts and repair hours over four years, is a set of count rows on λ and μ.

Which row acts on which entry. The all-up capacity at the design day, equation (11.4), depends on the three log-ratios alone, with gradient

$$\nabla_{\mathbf{z}} C_{\text{all}} = (C_{C1},\ C_{C2},\ C_{C3},\ 0,\ 0,\ 0,\ 0,\ 0)^{\mathsf{T}}, \tag{11.10}$$

in kW, since $\partial C_i / \partial \kappa_i = C_i$ and the derate drops out at the rating wet bulb. Of the rows of the operating record only the capacity tests have a nonzero in the columns of the κ_i: a test of unit i at wet bulb T reads $400\,\mathrm{e}^{\kappa_i}(1 - d\,(T - 22))$, with the entries C_i in κ_i's column and $-400\,\mathrm{e}^{\kappa_i}(T - 22)$ in d's; the monthly power and pressure rows see the efficiency, the pipe and the pump, not the capacity. Read as observability, the date is known as well as the capacity tests know the log-ratios, with the derate entering only through the tests taken away from 22°C.

The outage rows have no nonzero in any column of $\mathbf{z}$, and the rows of $\mathbf{z}$ none in those of λ and μ, so the posterior factors,

$$p(\mathbf{z}, \lambda, \mu \mid \mathbf{y}) = p(\mathbf{z} \mid \mathbf{y}_{\text{op}})\; p(\lambda, \mu \mid \mathbf{y}_{\text{out}}), \tag{11.11}$$

with $\mathbf{y}_{\text{op}}$ the operating record and $\mathbf{y}_{\text{out}}$ the outage record.

Modeling choices at the rows. The three chillers share one failure rate and one repair rate, read from their pooled record: this gives one q for every unit, and it fails if one unit is much less reliable than the others. The units fail independently of each other: this makes the state probability a product, and it fails for a cause shared by the three, such as their common supply of power or water, the subject of Chapter 15. The unavailability is the long-run share of time down, applied to the design day as to any other: this assumes no maintenance is planned for that day.

11.2 The posterior of the state (step 3)

All four questions of Table 11.1 are read from one posterior: the site's one fit, with the outage posteriors beside it, as Chapter 9 builds them. No row is added and no fit is repeated; this section cites the posterior and the draws the answers use.

11.2.1 The fit of the site, cited

The MAP state $\hat{\mathbf{z}}$ of the eight entries (11.9) and the Laplace covariance $\boldsymbol{\Sigma} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}$ at $\hat{\mathbf{z}}$ are those of Chapter 9, the Gauss–Newton fit of the operating record in the form of equation (4.38). The economizer is not in the operating record, so g_e keeps its handbook prior; it does not enter this chapter's plant. At $\hat{\mathbf{z}}$ the as-built units fall short of their nameplate, and with one unit out the remaining two carry between 721.5 and 749.8 kW, the smaller with the largest unit, C2, out. Chapter 9 tabulates the entries with their spreads and the stated truth. The failure and repair rates have conjugate gamma posteriors from the outage record, and the unavailability's posterior is that of $\lambda/(\lambda + \mu)$ under them.

11.2.2 The draws of the state and of the outage rates

The answers read the posterior through the joint draws of Chapter 9: states $\mathbf{z}_i$ from the Laplace normal $\mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$, each paired with an unavailability q_i drawn from the outage posterior, independent of it by equation (11.11), and the weights w_i of equation (5.16) that reweight the states to the exact posterior of the operating record. Each draw is pushed through the capacities (11.4) and the enumeration (11.3) at every month of the path, which gives one row $\pi_m(\mathbf{z}_i, q_i)$, $m = 0, \ldots, 72$, per draw. Every answer of step 4 is a function of these rows: their average over the draws, equal weights for the reported value and the weights w_i for the exact one; or, per draw, the first month in which the row reaches the tolerance.

11.3 The answers (step 4)

Every answer is read from the draws of Section 11.2.2. The reported value is the posterior average by the equal-weight draws, the joint reading over the units' states and the plant's capacities; the exact posterior stands beside it with the gap. The month-by-month probability comes first, because the date is read off it.

Table 11.2: The probability of a shortfall on the design day along the load path, equation (11.12): the existing three chillers, without the economizer, the IT load growing 15 %/yr from 660 kW.

Month [–]	Year [yr]	Loop heat L_m [kW]	$\hat{P}_m$, draws [–]	P_m, exact [–]
0	0.00	682.0	0.0001	0.0001
6	0.50	729.8	0.0054	0.0055
12	1.00	781.0	0.0165	0.0165
18	1.50	835.9	0.0165	0.0165
24	2.00	894.8	0.0165	0.0165
30	2.50	958.0	0.0165	0.0165
36	3.00	1025.8	0.0165	0.0165
42	3.50	1098.4	0.2998	0.3107
48	4.00	1176.3	1.0000	1.0000
54	4.50	1259.9	1.0000	1.0000
60	5.00	1349.5	1.0000	1.0000
66	5.50	1445.6	1.0000	1.0000
72	6.00	1548.6	1.0000	1.0000

11.3.1 The shortfall probability month by month

Planner's question. How does the shortfall probability grow along the path? The question asks for the probability of a shortfall on each year's design day, at every month of the load path.

General form. A prediction under a threshold is a tail of a derived quantity under the posterior, equation (2.12),

$$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) \geq f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

read here at each node of a scenario chain, with the probability that a unit is down summed inside the integral.

Specialized form. The question is equation (2.12) with the event a shortfall in the units' state, the load L_m of the path (11.5) against the capacity C_s of the state; the state of the units is summed by the enumeration (11.3) and the plant's state integrated by the draws,

$$P_m \approx \hat{P}_m = \frac{1}{N} \sum_{i=1}^{N} \pi_m(\mathbf{z}_i, q_i), \qquad P_m^{\text{exact}} = \sum_{i=1}^{N} w_i\, \pi_m(\mathbf{z}_i, q_i), \tag{11.12}$$

which is equation (11.2) by the draws. The enumeration is exact for each draw; only the integral over the plant is sampled.

Building blocks. The draws and their weights (Section 11.2.2); the capacities (11.4); the enumeration (11.3) and its regimes (11.6); the path (11.5).

The path. Table 11.2 gives the shortfall probability at every sixth month: each row is a node of the chain, the columns its month and year, the loop heat there, and the probability by the draws and by the exact posterior. The same probabilities, with the draws' value at every month from 36 to 48 added, are plotted against the month in Figure 11.1, on a logarithmic scale so that the plateau and the step stand on one axis.

Today the shortfall probability is 0.0001: the loop heat of 682.0 kW lies below every one-unit-out capacity, so only a state with two units down falls short, the first line of equation (11.6).

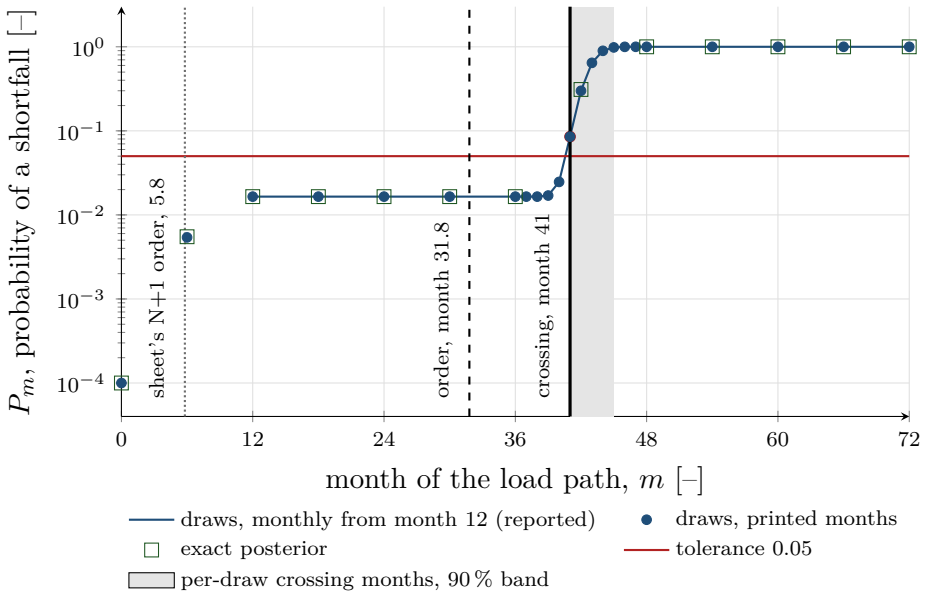


Figure 11.1: The probability of a shortfall on the design day along the load path, Table 11.2 with every month from 36 to 48 added, on a logarithmic axis. Dots: the draws at the printed months; squares: the exact posterior. Blue line: the draws month by month from month 12, flat at the plateau 0.0165 to month 38 (the probability cannot fall with the month, and it is 0.0165 at months 12 and 38) and rising through 0.0247 at month 40 to one by month 47; months 1 to 11 are not printed and carry no line. Red line: the tolerance; red dot: month 41, 0.0853, the first month at or above it. Solid vertical: the crossing month; dashed vertical: the order date $41 - 9.2 = 31.8$; dotted vertical: the sheet's N+1 order date. Shaded: months 41 to 45, the central 90 % band of the per-draw crossing months.

At month 6, 729.8 kW, the load has passed the 721.5 kW the plant carries with its largest unit out, and the probability has risen to 0.0054, about the probability q of that one state. From month 12 to month 36 it stays at 0.0165, the plateau: by month 9 the load has passed 749.8 kW, the larger of the one-unit-out capacities at the MAP state, so every state with a unit down falls short and only the all-up state carries. Over two years in which the load grows from 781.0 to 1025.8 kW the probability does not move, because the all-up capacity is above the load in nearly every draw and the units' chance of being down does not depend on the load. Then it steps, month by month in the figure: 0.0170 at month 39, 0.0247 at month 40, 0.0853 at month 41, 0.2998 at month 42 and 1.0000 from month 47, as the load passes the all-up capacity in one draw after another. The exact posterior agrees to the fourth decimal on the plateau and moves the step slightly earlier, 0.3107 against 0.2998 at month 42.

The plateau is the redundancy risk the planner already carries. The sheet's N+1 rule would order the fourth unit by month 5.8 to avoid it; on the as-built units it is broken even earlier, in month 9. But a shortfall probability of 1.65 % on the design day is below the 5 % the planner has said is acceptable. This is the planner's stated tolerance doing its work: a plant that has lost N+1 is not thereby short, and the tolerance says by how much the planner is willing to rely on the three units all running.

11.3.2 The crossing month and the order date

Planner's question. When does the next unit have to be ordered so that it is in before the probability of a shortfall passes 5 %? The question asks for the first month in which the shortfall probability reaches the tolerance, less the lead time.

General form. A prediction under a threshold, equation (2.12),

$$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) \geq f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

read along the chain until it first reaches a level: the first node at which the tail of the shortfall reaches α.

Specialized form. The question is equation (2.12) along the chain, with the probability of each node that of equation (11.12), and the date read by equation (11.1),

$$\hat{m}^{\dagger} = \min\{m : \hat{P}_m \geq 0.05\}, \qquad \hat{m}_o = \hat{m}^{\dagger} - 9.2, \tag{11.13}$$

in months, and the same with P_m^{exact} for the exact posterior. The shortfall probability rises with the month, since the load does and every indicator of equation (11.3) can only switch on, so the first month above the tolerance is the only crossing. Beside the averaged probability stands the route on the binding element alone: the all-up capacity's band mapped through the path. The month whose load first reaches the 5 % point of C_{all} is the month by which 5 % of the draws have crossed, by equation (11.7), the quantile map of Chapter 6 applied to the capacity.

Building blocks. The probabilities $\hat{P}_m$ and P_m^{exact} at every month (Section 11.3.1); the lead time of 40 weeks; the all-up capacities of the draws and the map (11.7).

The date. By the draws the shortfall probability reaches 5 % in

$$\hat{m}^{\dagger} = \text{month } 41 \text{ (year } 3.417), \qquad \hat{m}_o = 41 - 9.2 = \text{month } 31.8 \text{ (year } 2.650), \tag{11.14}$$

and the exact posterior gives month 41 as well, so the order date is the same by both. The probability is 0.0247 in month 40 and 0.0853 in month 41, and the draws' Monte Carlo standard

Table 11.3: The dates for the fourth chiller: the design sheet's two, the reported date with the exact posterior and the route on the all-up capacity beside it, the band of the per-draw crossing months, and the stated truth. Order month = crossing month less the 9.2-month lead time.

Source	Crossing month [–]	Order month [–]
Design sheet, N+1 rule (800 kW of nameplate with one unit out)	15	5.8
Design sheet, nameplate all up (1,200 kW)	50	40.8
Draws, $\hat{P}_m \geq 0.05$ (reported)	41	31.8
Exact posterior, $P_m^{\text{exact}} \geq 0.05$	41	31.8
All-up capacity's 5 % point, 1083.5 kW, mapped	41	31.8
Per-draw crossing month, 5 % point	41	31.8
Per-draw crossing month, median	43	33.8
Per-draw crossing month, 95 % point	45	35.8
Stated truth	42	32.8

error at month 41 is 0.0018: month 41's value stands nearly twenty standard errors above the tolerance and month 40's about half the tolerance below it, so the month is not a sampling accident. The all-up capacity's 5 % point is 1083.5 kW, and the loop heat first reaches it in month 41: the route on the capacity alone gives the same month. The dates are set side by side in Table 11.3: each row is a source of a date, the columns the month in which the load is to be carried and the month by which the unit must be ordered.

The reported date sits between the sheet's two, and close to neither. The N+1 rule dates an event that the as-built units reach in month 9, six months before the sheet's month 15, and that the tolerance does not ask to avoid; its order date is $31.8 - 5.8 = 26$ months earlier than the tolerance asks. The nameplate date is nine months late, because the three as-built units carry less than their nameplate and the load reaches what they carry in month 41, not month 50. The four routes of the table that read the averaged probability, the exact posterior, the capacity's 5 % point and the per-draw months' 5 % point all give month 41. The capacity is a sum of three exponentials of nearly normal entries, not a ratio, and its band is nearly symmetric about its median, 22.7 kW below and 23.1 kW above; the skew that set the routes apart in Chapter 5 is absent here, and the delta form is not formed.

11.3.3 The band of the date

Planner's question. How sure is the date? The question asks for the band of the month in which the plant crosses the tolerance, read state by state over the posterior.

General form. An estimation question about a derived quantity asks for its posterior distribution, equation (2.9),

$$\Pr(f(\mathbf{z}) \leq f^* \mid \mathbf{y}) = \int \mathbf{1}\big[f(\mathbf{z}) \leq f^*\big]\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

summarized by its median and its central 90 % credible interval, equation (2.11).

Specialized form. The question is equation (2.9) with $f = m^{\dagger}(\mathbf{z}, q)$, the per-draw crossing month of equation (11.7), read from the draws,

$$\hat{F}(m) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[m^{\dagger}(\mathbf{z}_i, q_i) \le m], \qquad m_p^{\dagger} = \min\{m : \hat{F}(m) \ge p\}, \quad p = 0.05,\ 0.5,\ 0.95, \tag{11.15}$$

and with the weights w_i for the exact posterior. Each draw's month is the month its own conditional probability $\pi_m(\mathbf{z}_i, q_i)$ first reaches the tolerance: the date the plant would cross if its capacity and its outage rates were known to be that draw's. The order months are the same quantiles less the lead time. By equation (11.8) the reported month of Section 11.3.2 is no later than $m_{0.05}^{\dagger}$.

Building blocks. The rows $\pi_m(\mathbf{z}_i, q_i)$ of the draws and the weights (Section 11.2.2); the per-draw months (11.7); the lead time.

The band. The per-draw crossing months have

$$\text{median } m^{\dagger} = 43, \qquad m^{\dagger} \in [41,\ 45] \text{ with } 90\,\% \text{ probability}, \tag{11.16}$$

and the exact posterior gives the same median and band. Less the lead time, the order month has a median of 33.8 and a band from 31.8 to 35.8. The share of the draws that cross before month 41 is 0.0089. The per-draw crossing months are counted in Figure 11.2: the horizontal axis is the crossing month, the vertical axis the share of the draws that cross in it; the bars inside the band, months 41 to 45, are dark and those outside it light; the solid line marks the reported month, the dashed line the median and the dotted line the stated truth, a check marker for Section 11.4.4. A further 0.0007 of the draws, too few to draw, cross in months 5 to 10: they are the states with an unreliable chiller fleet, $q \ge 0.017$, whose plateau alone passes 5 %, and they cross as soon as the load passes their one-unit-out capacities.

The reported month is the early end of the band, and this is not a coincidence of these numbers: it is the inequality (11.8). The averaged probability counts every draw that has crossed at its full weight of one, so it reaches 5 % no later than the month by which 5 % of the draws have crossed, and here even a little sooner, since the draws that have not crossed add their plateau. On the plateau the average is $P_m \approx F_m + (1 - F_m)\,\bar{p}$ with $\bar{p} = 0.0165$, which reaches 0.05 when the crossed share is

$$F_m = \frac{0.05 - 0.0165}{1 - 0.0165} = 0.034,$$

a share between month 40's, 0.0089, and month 41's, 0.0706, the sum of the shares of Figure 11.2 and of the early crossings to month 41. The same reading gives $P_{40} \approx 0.0089 + 0.9911 \times 0.0165 = 0.025$ and $P_{41} \approx 0.0706 + 0.9294 \times 0.0165 = 0.086$, against the draws' 0.0247 and 0.0853. The reported month and the band's early end fall in the same whole month. The planner who reads "order by month 31.8, band 31.8 to 35.8" should read the band as the months in which the plant itself may cross, and the reported date as the month by which the probability averaged over all of them reaches the tolerance: a tolerance on the averaged probability is, here, a quantile of the date. A date read at the median of the per-draw months, month 43, would leave the plant short on the design day with a probability far above the tolerance in the months before it, 0.2998 already at month 42.

11.3.4 What sets the date

Planner's question. What sets the date? The question asks which quantity of the plant the date and its band are made of: the units' chance of being down, or their capacity.

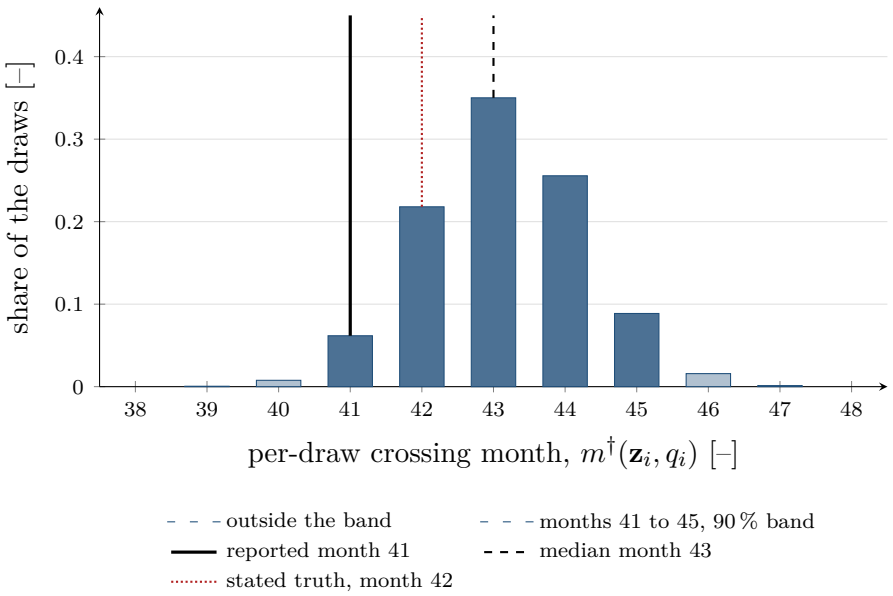


Figure 11.2: The per-draw crossing months of the existing three chillers, equation (11.7): the share of the draws that cross in each month, months 39 to 47. Dark bars: months 41 to 45, the central 90 % band; light bars: outside it. The share in months 5 to 10, 0.0007 in all, is the draws whose plateau alone passes the tolerance and is not drawn. Vertical lines, above the bars: the reported month, the median of the per-draw months, and the stated truth, a check marker (Section 11.4.4).

General form. The moments of a derived quantity under the posterior, equation (2.10),

$$\mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}] = \int f(\mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}, \qquad \mathrm{Var}(f(\mathbf{z}) \mid \mathbf{y}) = \int \big(f(\mathbf{z}) - \mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}]\big)^2\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

read here for the two quantities the staircase (11.6) turns on.

Specialized form. The question is equation (2.10) for two quantities: the all-up capacity $C_{\mathrm{all}}(\mathbf{z})$ of equation (11.4), with gradient (11.10) on the three log-ratios and its band read from the draws, and the plateau $1 - (1 - q)^3 \approx 3q$, the shortfall probability on the third line of equation (11.6), against the tolerance. By equation (11.7) the per-draw month is the capacity mapped through the path as long as the plateau stays below the tolerance; the band of the date is then the band of the capacity, mapped, and the outage rates move only the height of the plateau.

Building blocks. The all-up capacities of the draws; the unavailabilities q_i; the one-unit-out capacities at $\hat{\mathbf{z}}$; the path (11.5).

The two quantities. The all-up capacity of the existing three chillers at the design day, without the economizer, has a median of 1106.2 kW and a central 90 % band from 1083.5 to 1129.3 kW. The load reaches the band's 5 % point in month 41, the reported month, and the three points of the band map to months 41, 43 and 45, the band of the date. The plateau, at month 24, is 0.0165 by the draws against three times the mean unavailability, 0.0166; it is a third of the tolerance. With one unit out the plant carries 721.5 to 749.8 kW at $\hat{\mathbf{z}}$, and the load passes the larger in month 9: from then on the units' outages set a floor under the probability, and the capacity sets the month at which it leaves the floor. The date is the as-built capacity's:

Table 11.4: The date under the planner's inputs: the lead time, the economizer, the tolerance, and the shared pipe as a second limit. Each tolerance row is the first month in which the draws' probability reaches that tolerance.

Variation	Crossing month [–]	Order month [–]
As reported: three units, 40-week lead time, $\alpha = 0.05$	41	31.8
Lead time 30 weeks	41	34.1
Lead time 52 weeks	41	29.0
Economizer built, 40-week lead time	44	34.8
Tolerance $\alpha = 0.02$	40	30.8
Tolerance $\alpha = 0.10$	42	32.8
Tolerance $\alpha = 0.20$	42	32.8
Shared pipe's limit, loop heat 1148.8 kW (median)	46	–

its band is the band the capacity tests leave on the three log-ratios, carried through the path, and a change in the outage record would raise or lower the plateau without moving the month, unless it raised q threefold.

11.4 The checks (step 5)

The answers of Section 11.3 rest on the Laplace draws of the site, on the outage posterior and on the planner's inputs, and each can be tested.

11.4.1 The draws against the exact posterior

The reported probabilities and the exact ones are the same draws under equal and under importance weights, so their difference is the departure of the site's posterior from the Laplace normal, not sampling noise between independent sets. Table 11.2 shows the two equal to the fourth decimal at every month but two: at month 6, 0.0055 against 0.0054, and at month 42, 0.3107 against 0.2998, a gap of +0.0109. Month 42 lies inside the all-up capacity's band, where the probability is the share of the draws whose capacity the load has passed, and the exact posterior puts a little more weight on low capacities. The gap moves the step, not the crossing: both routes reach the tolerance in month 41, and both give the per-draw band $[41, 45]$ with median 43.

11.4.2 The plateau against the outage posterior

On the plateau every draw is on the third line of equation (11.6), and its probability is $1-(1-q_i)^3$, so the plateau's average should be close to three times the mean of q. The draws give 0.0165 at month 24 and $3\bar{q} = 0.0166$: the plateau is the outage posterior alone, and the enumeration reads it exactly. The small difference is the $3q^2$ and q^3 terms that $1-(1-q)^3$ subtracts from $3q$.

11.4.3 The economizer, the lead time and the pipe

The date rests on three inputs the planner sets and on one element the shortfall event leaves out. Table 11.4 varies them: each row is a variation, the columns the crossing month and the order month that follow.

The lead time moves the order date one for one and leaves the crossing where it is: 30 weeks puts the order at month 34.1, 52 weeks at month 29.0. Built, the economizer adds its duty to

Table 11.5: The checks of the expansion date. The draws are those of the site's one fit and its outage posteriors; each reported value stands beside the value it is checked against.

Check	Quantity [unit]	Reported or fitted	Checked against
Exact posterior, month 6	P_6 [–]	0.0054	0.0055
Exact posterior, month 42	P_{42} [–]	0.2998	0.3107
Exact posterior, crossing	$m^{\dagger}$ [–]	41	41
Sampling error at the crossing	P_{41}, MC s.e. [–]	0.0853	0.0018
Exact posterior, per-draw band	$m^{\dagger}_{0.05}$, median, $m^{\dagger}_{0.95}$ [–]	41, 43, 45	41, 43, 45
Plateau against $3\bar{q}$	P_{24} [–]	0.0165	0.0166
Crossing before the reported month	$\hat{F}(40)$ [–]	0.0089	–
Capacity's 5 % point, mapped	month [–]	41	41
Stated truth, all-up capacity	C_{all} [kW]	1106.2 (median)	1088.0
Stated truth, crossing	$m^{\dagger}$ [–]	41	42
Stated truth, order	m_o [–]	31.8	32.8

every state of the enumeration and the crossing comes in month 44, three months later; that duty is the handbook guess the operating record never tested, since the economizer has never run, and Chapter 17 weighs it. The tolerance moves the date little above the plateau and a great deal below it. Above it, a tenfold range of tolerances, 0.02 to 0.20, moves the crossing two months, from month 40 to month 42, because the probability climbs from 0.0247 to 0.2998 in those two months. A tolerance at or below the plateau's 0.0165 is crossed by month 12 at the latest, so its order falls due by month 2.8, and below month 6's 0.0054 it would already be due. The planner's 5 % is the input on which the date most depends, through the side of the plateau it falls on.

The shortfall event of equation (11.3) counts the chillers only. The shared supply pipe reaches its allowable pressure drop at a site IT load of 1126.8 kW at the median, a loop heat of 1148.8 kW, in month 46: five months after the chillers' crossing. A fourth chiller ordered for month 41 carries the plant past the chillers' date and into the pipe's; Chapter 13 sets the two limits side by side with their bands.

11.4.4 The answers against the stated truth

The record of the site is generated from a stated truth, and the last check compares the date with it. Table 11.5 collects the checks: each row is a check, the columns the quantity checked, the reported or fitted value and the value it is checked against. At the truth the all-up capacity is 1088.0 kW, inside the posterior band from 1083.5 to 1129.3 kW and near its lower end; the true plant crosses in month 42, year 3.500, and would be ordered by month 32.8. The reported crossing, month 41, is one month before the truth's, and the true month lies inside the per-draw band $[41, 45]$, one month above its early end. The reported date is on the safe side of the truth by one month, and the median of the per-draw months, 43, would have been late by one.

11.5 The report (step 6)

Order the fourth chiller by month 31.8, two years and eight months from today, not by month 5.8 as the sheet's N+1 rule says nor by month 40.8 as its nameplate says. By then the existing three units, which carry less than their nameplate, are within a lead time of the month in which the shortfall probability on the design day passes the planner's 5 %.

Table 11.6: The report for the site's chiller plant on the 15 %/yr load path, the existing three units without the economizer, on each year's design day. Reported values by the draws of the site's posterior; the exact posterior beside each.

Question	Answer	Band
When does the next unit have to be ordered so that it is in before the probability of a shortfall passes 5 %?	by month 31.8, year 2.650; crossing in month 41 (exact: month 41)	order months 31.8 to 35.8, median 33.8; 34.1 at a 30-week and 29.0 at a 52-week lead time
How does the shortfall probability grow along the path?	0.0001 today, 0.0054 at month 6, 0.0165 from month 12 to month 36, 0.2998 at month 42, 1.0000 from month 48 (exact: 0.3107 at month 42)	the plateau is $3\bar{q} = 0.0166$; with the economizer built the crossing is in month 44
How sure is the date?	per-draw crossing month, median 43 (exact: 43)	months 41 to 45 (exact: 41 to 45); the reported month is the band's early end; 0.0089 of the draws cross before it
What sets the date?	the all-up capacity, median 1106.2 kW; the load reaches its 5 % point in month 41	1083.5 to 1129.3 kW; the plateau of one unit out, 0.0165, is a third of the tolerance

11.5.1 The answers and their bands

Table 11.6 gives the answers in the order of Table 11.1. Each row is a question; the second column gives the reported answer with the exact posterior beside it, and the third its band: the per-draw months for a date, the plateau and the step for the path, and the inputs that move the date.

11.5.2 The design-review entry

The design review's entry reads the crossing of the shortfall probability, not the sheet's rule. On the planner's 15 %/yr path the probability of a shortfall on the design day, with the existing three chillers and without the economizer, reaches 5 % in month 41 by the draws and by the exact posterior, and the fourth chiller, at a 40-week lead time, is ordered by month 31.8. The entry records the band of the per-draw months, 41 to 45, and that the reported month is its early end by construction, so that the date already carries the tolerance across the band. It records that the N+1 rule is broken in month 9 on the as-built units and that the plant then runs on a redundancy risk of 1.65 %, which the planner's tolerance accepts. It records the two inputs the date moves with: the lead time one for one, and the tolerance, which would bring the order forward to month 2.8 or earlier if it were set at or below the plateau. And it records that the shared pipe follows the chillers by five months, in month 46, so that the expansion plan must date the pipe's remedy too.

Common mistake

Dating an expansion by the N+1 rule. The rule asks for the redundancy the planner may not need and ignores the capacity the plant may not have: here it orders twenty-six months early, while the nameplate date it implies for the units all up is nine months late.

11.6 The next measurement (step 7)

The date's band is the as-built capacity's band, mapped through the path, and the outage record only sets the floor the probability rests on. A measurement that narrows the date is one that narrows the capacity of the existing units at the design day.

11.6.1 A capacity test on a hot day

By equation (11.10) the date depends on the three log-ratios alone. The capacity tests that inform them were taken at wet bulbs between 15 and 22°C, and a test away from the rating condition reads the log-ratio and the derate together, its row having an entry $-400\,\mathrm{e}^{\kappa_i}(T-22)$ in the derate's column. A test of each existing unit on a hot day, near the design day's 22°C, is a row on its log-ratio with the derate's entry near zero: it reads the quantity the date is made of. The ranking of such tests by the width each takes off the date, against the other tests and surveys a planner can commission on the site, is Chapter 18's; this chapter computes no preposterior. One more input moves the date by a whole quarter: the economizer, whose built duty would move the crossing three months later and whose conductance is a handbook guess. Before the order is placed at month 31.8, a decision to build the economizer first is a decision on that guess, and Chapter 17 says how much of the headroom's band the guess carries.

11.6.2 What the analysis showed

The design sheet gave two dates three years apart, month 5.8 by its N+1 rule and month 40.8 by its nameplate, and the analysis puts the order at month 31.8, with the plant crossing the planner's tolerance in month 41. The N+1 rule dated the wrong event: losing a unit's worth of redundancy costs the plant a shortfall probability of 1.65 % on the design day, a plateau that the load does not raise and that sits below the tolerance for more than two and a half years. The nameplate dated the right event on the wrong capacity: the as-built units carry a median of 1106.2 kW, not 1,200 kW, and the load reaches their capacity's lower end nine months before it reaches the nameplate. The date is made of capacity, not of outages: each state of the posterior crosses where the load passes its all-up capacity, so the date's band is the capacity tests' band carried through the path. And the reported month is the band's early end by construction: averaging the shortfall probability over the posterior counts every state that has crossed at its full weight, so a tolerance of 5 % on the average is reached no later than the month by which 5 % of the states have crossed. The stated truth, month 42, lies one month inside that end.

11.6.3 Limits of the model

A known path. The growth rate is a stated input and the band carries the plant's uncertainty, not the plan's. An uncertain rate, with a distribution of its own, is Chapter 12's question; a load that arrives in steps, as rows of racks are installed, moves the date to the step.

The plant as it stands. The economizer is drawn and not built, and the date is that of the three chillers alone; built, it moves the crossing three months on a handbook guess. The pump, the pipe and the feed are left out of the shortfall event, and the pipe binds five months after the chillers. The fourth unit's own capacity and its effect on the plateau are not modeled here: the date is the date by which the existing plant needs it, and the size of the unit is Chapter 10's question.

The design day, not the year. The shortfall probability is that of each year's design day, with the units' long-run unavailability applied to it. It is not an annual loss-of-load expectation and it counts no energy unserved; Chapter 23 sums the shortfall over the hours of the year.

Maintenance planned on a hot day would raise the chance of a unit being down above q, and Chapter 22 takes it up.

One rate, independent outages. The three units share one unavailability, read from their pooled record, and fail independently. A unit much less reliable than the others would split the plateau's three states, and a cause shared by the three would raise the two-down states that equation (11.6) neglects; Chapter 15 reads such a shared path.

The month grid. The date is read on whole months, and every crossing is rounded up to the month in which the load first passes the capacity; a weekly grid would move the crossing by a fraction of a month.

The constructed record. The capacity tests, the monthly windows and the outage record are generated from a stated truth with stated errors, not taken from a site, and the comparison with that truth in Section 11.4.4 is a check of the method on its own model, not of the model against a real plant.

The next chapter. Chapter 12 lets the growth rate itself be uncertain and asks how much of what is built may stand idle.

11.7 What to retain

Definition 11.1 (Expansion date by a shortfall tolerance)**.** On a known load path L_m the shortfall probability of month m is the enumeration of the units' states, $\pi_m(\mathbf{z}, q) = \sum_s p(s \mid q)\, \mathbf{1}[C_s(\mathbf{z}) < L_m]$, averaged over the posterior of the plant, $P_m = \mathbb{E}[\pi_m \mid \mathbf{y}]$. The crossing month is the first month with $P_m \geq \alpha$, and the order date is the crossing month less the lead time. Each state of the posterior crosses in its own month, the per-draw crossing month, whose quantiles give the date's band; where the redundancy-risk plateau $1 - (1 - q)^3$ stays below α, the per-draw month is the month the load passes the all-up capacity, and the reported month is no later than the α-quantile of the per-draw months.

Four conclusions carry beyond the site. A redundancy rule is not a tolerance: losing N+1 puts a floor of about nq, for n units each down with probability q, under the shortfall probability, which the planner may well accept, and the rule's date is then too early by the time the plant spends on that floor. A date read from nameplate is read on the wrong capacity; the as-built capacity, with its band, dates the step from the floor to one. The band of a date along a path is the band of what sets it, here the all-up capacity, mapped through the path, and the outage rates move only the floor. And a tolerance on a posterior-averaged probability is a quantile of the date: the reported date sits at the early end of its own band, so the band is not a margin to be added to it.

Table 11.7 lists the terms of the chapter with their expressions, units and the subsection that defines each.

Table 11.7: The terms of expansion timing and where each is defined. Units for the site's chiller plant.

Term	Expression	Units	Meaning	Defined in
Order date	$m_o = m^\dagger - \ell$	month	crossing less lead time	11.1.1
Shortfall probability	$P_m = \mathbb{E}[\pi_m(\mathbf{z}, q) \mid \mathbf{y}]$	–	averaged over the plant	11.1.1
Enumeration	$\pi_m = \sum_s p(s \mid q)\, \mathbf{1}[C_s < L_m]$	–	units up and down at month m	11.1.1
Capacity at the design day	$C_i = 400\,\mathrm{e}^{\kappa_i}$, $C_{\mathrm{all}} = \sum_i C_i$	kW	derate drops out at the rating	11.1.1
Loop heat on the path	$L_m = Q_{\mathrm{IT},0}(1+\gamma)^{m/12} + Q_o$	kW	monthly scenario chain	11.1.1
Plateau	$1 - (1-q)^3 \approx 3q$	–	redundancy risk, one unit out	11.1.1
Per-draw crossing month	$\lfloor \tau(C_{\mathrm{all}}) \rfloor + 1$	month	the plant's own date	11.1.1
Reported against per-draw	$P_m \geq F_m$, so $m^\dagger \leq m^\dagger_\alpha$	–	early end by construction	11.1.1
Rows and posterior				
Gradient of the capacity	$\nabla_{\mathbf{z}} C_{\mathrm{all}} = (C_{C1}, C_{C2}, C_{C3}, 0, \ldots)$	kW	capacity tests set the date	11.1.4
Factored posterior	$p(\mathbf{z} \mid \mathbf{y}_{\mathrm{op}})\, p(\lambda, \mu \mid \mathbf{y}_{\mathrm{out}})$	–	two records, separate entries	11.1.4
Questions and answers				
Path by the draws	$\hat{P}_m = N^{-1} \sum_i \pi_m(\mathbf{z}_i, q_i)$	–	reported probability	11.3.1
Crossing and order	$\min\{m : \hat{P}_m \geq 0.05\}$, less 9.2	month	the reported date	11.3.2
Band of the date	quantiles of $m^\dagger(\mathbf{z}_i, q_i)$	month	how sure the date is	11.3.3
Crossed share at the tolerance	$(\alpha - \bar{p})/(1 - \bar{p})$	–	share crossed when $P_m = \alpha$	11.3.3

11.8 Exercises

Exercise 11.1. *Formulation.* Write the shortfall probability of month 6 as an expectation over the posterior, say which states of the units fall short there at the MAP state, and give its value. ($P_6 = \mathbb{E}[\pi_6 \mid \mathbf{y}]$ with $\pi_6 = \sum_s p(s \mid q)\, \mathbf{1}[C_s < L_6]$ and $L_6 = 729.8\,\mathrm{kW}$. At $\hat{\mathbf{z}}$ the plant carries 721.5 kW with its largest unit, C2, out and up to 749.8 kW with another out, so the load falls between them: the four states with two or three units down fall short, the state with C2 down falls short, and the two other one-out states carry, the second line of equation (11.6) with $k_6 = 2$. The probability is close to q, and the draws give $P_6 = 0.0054$.)

Exercise 11.2. *Derivation.* Show that the plateau is nearly $3q$, and check it against the draws. (On the plateau every state with a unit down falls short and the all-up state carries, so $\pi = 1 - p(\text{all up}) = 1 - (1-q)^3 = 3q - 3q^2 + q^3$; the terms after $3q$ are of order q^2, about half a percent of $3q$ at $\bar{q} = 0.00554$. The draws give 0.0165 at month 24 against $3\bar{q} = 0.0166$.)

Exercise 11.3. *Derivation.* Prove that the reported crossing month is never later than the 5 % point of the per-draw crossing months when every draw's plateau stays below the tolerance, and state both months. (A draw that has crossed has $L_m > C_{\mathrm{all}}$ and $\pi_m = 1$; a draw that has not has $\pi_m \geq 0$. So $P_m \geq F_m$, the crossed share. In the month $m^\dagger_{0.05}$ the crossed share is at least 0.05, so $P_m \geq 0.05$ there and the first month with $P_m \geq 0.05$ comes no later. The reported month is 41 and the 5 % point of the per-draw months is 41.)

Exercise 11.4. *Application.* The chiller vendor quotes a lead time of 52 weeks. Find the order date and say whether the crossing moves. (The lead time is $52 \times 7/365 = 0.997$ of a year, 11.97

months. The crossing month does not depend on the lead time and stays at month 41; the order date is 41 − 11.97, month 29.0.)

Exercise 11.5. *Application.* Show from the plateau and the crossed share why the averaged probability has not reached 5 % in month 40 and has in month 41. (In month 40 the share of the draws that have crossed is 0.0089, and the rest sit on the plateau at about 0.0165, so $P_{40} \approx 0.0089 + (1 - 0.0089) \times 0.0165 = 0.025$, below 0.05. The average reaches 0.05 once the crossed share reaches $(0.05 - 0.0165)/(1 - 0.0165) = 0.034$; in month 41 the crossed share is 0.0706, the sum of the printed shares to month 41, so $P_{41} \approx 0.0706 + 0.9294 \times 0.0165 = 0.086$, above it, and the crossing is month 41. The draws give $P_{40} = 0.0247$ and $P_{41} = 0.0853$.)

Chapter 12

Oversizing and Stranded Capacity

12.1 The build-out drawing at the ten-year load

The ten-year build-out drawing of the site's chilled-water plant shows nine chillers of 400 kW: the three in place, C1–C3, and six more of the same model, 3 600 kW of nameplate at the design day, with the waterside economizer drawn beside them. The plan's forecast takes the site's IT load from 660 kW today at 15 % a year, and with the halls' 22 kW of other heat the loop carries 2 692.1 kW in year ten. The design sheet divides the one by the other, $3\,600/2\,692.1 = 1.337$, and writes beside the drawing: oversized by 33.7 %, past the 30 % that the planning policy tolerates in a plant built for its horizon.

Finance reads the same line differently. A third more plant than load in year ten is capital bought a decade early, and the drawing asks for six chillers at R$ 1 600 000 each. The sheet's 1.337 rests on two numbers neither of which is known: the nameplate, which the existing chillers do not reach, and a growth rate ten years out. Finance's question is the drill question, verbatim,

What is the probability that the plant as drawn is more than 30 % *oversized at the ten-year load?*

It is answered with the method of Chapter 4 on the site plant introduced in Chapter 9, on that chapter's posterior of the site and its assumption for units not yet built. What is new is that the load is uncertain: for the first time in the book the growth rate is not a stated number but a stated distribution, and the question is a tail of the ratio of two uncertain sides.

Jargon: build-out drawing, ten-year load, forecast growth, stated distribution, oversizing ratio, stranded (oversized) capacity, shortfall, breakeven growth, as-built log-ratio, capital recovery factor, annualized cost, parameter risk (variance budget)

A *build-out drawing* is the drawing of the plant at the end of its planning horizon, here every unit the plan expects to install over ten years. The *ten-year load* Q_{10} is the heat the chilled-water loop must carry on the design day of year ten: the IT load grown from today's, plus the halls' other heat. The *forecast growth* γ is the yearly rate at which the IT load grows; a *stated distribution* is a probability distribution supplied as a planning input, like a price, and not fitted to any record. The *oversizing ratio* R is the plant's capacity over the ten-year load; the plant is *more than* 30 % *oversized* when $R > 1.30$. *Stranded (oversized) capacity* is the capacity installed beyond the load, $\max(C - Q_{10}, 0)$, capital that earns nothing while it waits; a *shortfall* is the opposite, the load beyond the capacity. The *breakeven growth* of a threshold is the growth rate at which the ratio sits exactly on it. A unit's *as-built log-ratio* κ compares its capacity at the design day with its nameplate, zero at the nameplate. The *capital recovery factor* turns an installed price into the equal yearly payment that repays it with interest over the unit's life, its *annualized cost*. The *parameter risk* of a derived quantity, its variance budget, splits its variance into one share per uncertain input.

Key idea

Oversizing is a ratio of two uncertain sides, and on a ten-year drawing the two sides are uncertain in very different measure. The plant's capacity is known from its tests to within a few percent; the load ten years out is a forecast compounded ten times. Because the load rises with the growth rate, the event that the ratio exceeds a threshold is the event that the growth falls short of a breakeven rate, and the answer is the forecast's tail at that rate, averaged over the posterior of the capacity. The same ratio has a lower tail, the plant short in year ten, which the oversizing question does not ask about and the planner needs.

12.1.1 Oversizing read as a ratio of two uncertain sides

Chapters 6 and 11 asked when a plant runs out of room on a known load path; the oversizing question asks how much room is left over at the end of the path. Its quantity is the ratio of the build-out's capacity C on the design day to the ten-year load Q_{10},

$$R = \frac{C}{Q_{10}},$$

and the drawing is more than 30 % oversized on the event $R > 1.30$. Both sides are uncertain, for different reasons: the capacity because the chillers do not deliver their nameplate and the record fixes their as-built ratios only to a band, the load because ten years of growth are a forecast.

A ratio of two positive quantities becomes a difference under the logarithm, and when the two sides are uncertain for unrelated reasons the variance of the difference is the sum of the variances,

$$\ln R = \ln C - \ln Q_{10}, \qquad \operatorname{Var}(\ln R) = \operatorname{Var}(\ln C) + \operatorname{Var}(\ln Q_{10}), \tag{12.1}$$

the second relation holding when C and Q_{10} are independent, so that the covariance term $-2\operatorname{Cov}(\ln C, \ln Q_{10})$ vanishes. Each side's variance is that side's share of the ratio's uncertainty, on the scale on which relative errors add. The logarithm is the natural scale for a ratio: a band of ± 10 % on either side is the same band of about ± 0.1 on $\ln R$, whichever side it is on. The split says before any number which side the answer belongs to, and which measurements could narrow it.

Principle 12.1. *An oversizing question is a tail of the ratio of capacity to a future load. On the logarithmic scale the ratio is a difference, and with the two sides independent its variance is the sum of theirs; the side with the larger share owns the answer, and only a measurement on that side can move it.*

12.1.2 The site at its ten-year build-out (step 0)

The plant is the site of Chapter 9: three chillers C1–C3 of 400 kW nameplate at the design day, a wet bulb of 22°C, with as-built log-ratios κ_{C1}, κ_{C2}, κ_{C3} and a derate d per kelvin of wet bulb; one pump, the shared supply pipe and the power feed; and a waterside economizer drawn but not built, whose path conductance is a handbook guess 25 kW/K scaled by e^{g_e}. The site's state has eight entries,

$$\mathbf{z} = \big(\kappa_{C1},\ \kappa_{C2},\ \kappa_{C3},\ d,\ g_e,\ c,\ l_k,\ l_e\big)^{\mathsf{T}}, \tag{12.2}$$

with c the log-efficiency of the chiller plant and l_k and l_e the logarithms of the pipe's flow coefficient and the pump's efficiency; the last three serve the pump, the pipe and the feed, and none of them enters this chapter's answer.

The capacity of the build-out. The build-out adds six new units of the existing model. Following Chapter 9, a new unit's as-built log-ratio is the mean of the existing three plus a between-unit spread of 0.04, the spread of the existing units' as-built ratios,

$$\kappa_j^{\mathrm{new}} = \bar{\kappa}(\mathbf{z}) + 0.04\,\epsilon_j, \qquad \bar{\kappa}(\mathbf{z}) = \tfrac{1}{3}\big(\kappa_{C1}+\kappa_{C2}+\kappa_{C3}\big), \qquad \epsilon_j \sim \mathcal{N}(0,1), \quad j = 1,\dots,6. \tag{12.3}$$

A chiller of nameplate N_i delivers $N_i\,\mathrm{e}^{\kappa_i}\,[1 - d\,(T_{wb} - T_{\mathrm{rate}})]$ at wet bulb T_{wb}, and the economizer, with the chilled-water return at $T_{wr} = 24$°C, carries $25\,\mathrm{e}^{g_e}\,(T_{wr} - T_{wb})$ when the wet bulb is below the return. The nameplates are stated at the design day, $T_{\mathrm{rate}} = 22$°C, so at the design day the derate factor is one, and the capacity of the nine units and the economizer together is

$$C(\mathbf{z}, \boldsymbol{\epsilon}) = 400 \sum_{i=1}^{3} \mathrm{e}^{\kappa_{Ci}} + 400 \sum_{j=1}^{6} \mathrm{e}^{\bar{\kappa}(\mathbf{z}) + 0.04\,\epsilon_j} + 25\,\mathrm{e}^{g_e}\,(24 - 22), \tag{12.4}$$

in kW: three fitted units, six new units that inherit the fitted mean, and the economizer at its guess, 50 kW when $g_e = 0$. The derate d drops out at the design day, so of the eight entries only κ_{C1}, κ_{C2}, κ_{C3} and g_e act on the capacity.

The ten-year load. The IT load today is $Q_{\mathrm{IT}} = 660$ kW and the halls' other heat, the fans and lighting, $Q_o = 22$ kW. At a yearly growth γ held over the decade, the loop heat of year ten is

$$Q_{10}(\gamma) = Q_{\mathrm{IT}}\,(1+\gamma)^{10} + Q_o, \tag{12.5}$$

2 692.1 kW at the forecast's 15 %. The load is an increasing function of the growth, and compounding makes it a convex one: a growth 0.04 above the forecast adds more load than a growth 0.04 below it removes.

The quantities asked about. The ratio, the stranded capacity and the shortfall are

$$R = \frac{C(\mathbf{z}, \boldsymbol{\epsilon})}{Q_{10}(\gamma)}, \qquad S = \max\big(C - Q_{10},\ 0\big), \qquad U = \max\big(Q_{10} - C,\ 0\big), \tag{12.6}$$

S and U in kW; at most one of them is positive, and $S - U = C - Q_{10}$.

Modeling choices at the graph. The growth is the change. Chapters 6 and 11 took the load path as known, a stated rate with, in Chapter 6, stated alternatives beside it, so that their bands were the plant's alone. This chapter gives the growth rate a stated distribution,

$$\gamma \sim \mathcal{N}(0.15,\ 0.04^2)\ \mathrm{yr}^{-1}, \tag{12.7}$$

from the planning inputs. It is a modeling choice, not a finding: the distribution is not inferred from any record of past load, and making a forecast is not this book's subject. The choice is made here because at ten years the forecast, not the plant, is the larger unknown, and a known path would answer a different question (Section 12.3.3 shows how different). It fails where the forecast's own spread is wrong, and Section 12.4.2 varies it. Three further choices make the graph. One rate holds for the whole decade: this makes the ten-year load a function of one number, and it fails if the growth varies from year to year, which would average part of the spread away and narrow the load. The ratio is taken with every unit up: the oversizing question is about what is installed, and a unit's outage, which the enumeration of Chapter 9 counts, does not change what was bought. And the economizer is counted at its handbook guess, as the drawing counts it, with the guess's spread carried into the capacity.

The graph. The rows of the build-out are drawn in Figure 12.1. On the left the three existing chillers, whose ratios the record fits, feed the six new units through their mean, and with the economizer they make the capacity at the design day; below, the growth rate, the only entry of the load side, carries today's IT load to year ten. The two sides meet in the ratio on the right, from which the three quantities of equation (12.6) are read. The growth is drawn in red: it is the one entry with no row of evidence, only its stated distribution.

12.1.3 The planner's questions (step 1)

Finance and the planner ask four questions of the build-out drawing, and step 1 writes each as a probability, a variance or an expectation over the chapter's uncertain quantities: the site's state $\mathbf{z}$, the new units' deviations $\boldsymbol{\epsilon}$ and the growth γ, with the site's record and the stated distributions on the right of the bar. Table 12.1 lists them, in the columns of Table 5.1: the question in the planner's words, the drill question first; the general form of its category from Chapter 2; its mathematical form for the build-out; its category; the computation of Section 2.3.6 that answers it; and the subsection that answers it. The first two are predictions under a threshold, the two tails of the one ratio. The third is estimation, the split of the ratio's variance between its two sides. The fourth is an expected consequence, the stranded capacity and its yearly cost.

12.1.4 The evidence as rows (step 2)

The evidence on the capacity side is the site's record of Chapter 9, row for row: two commissioning capacity tests of each chiller at two wet bulbs, and twelve monthly windows of chiller power, pipe pressure drop, pump power and feed power, fifty-four readings, with a prior row on each of the eight entries of equation (12.2): sixty-two rows in all. The capacity tests read $400\,\mathrm{e}^{\kappa_i}\,[1 - d\,(T_{wb} - 22)]$ directly and fix the three ratios and the derate; the monthly windows read the efficiency, the pipe and the pump. The economizer is in no row: it is not built, so the record has no window in which it carries heat, and g_e keeps its handbook prior, $g_e \sim \mathcal{N}(0,\ 0.35^2)$.

Two inputs are added, neither of them evidence. The new units' deviations have the standard normal of equation (12.3): units not yet built have no record, and their ratios are a statement about the model's spread from unit to unit. The growth has the stated distribution (12.7). Each enters as a row with a single nonzero in its own column, a prior row, and no reading touches those columns.

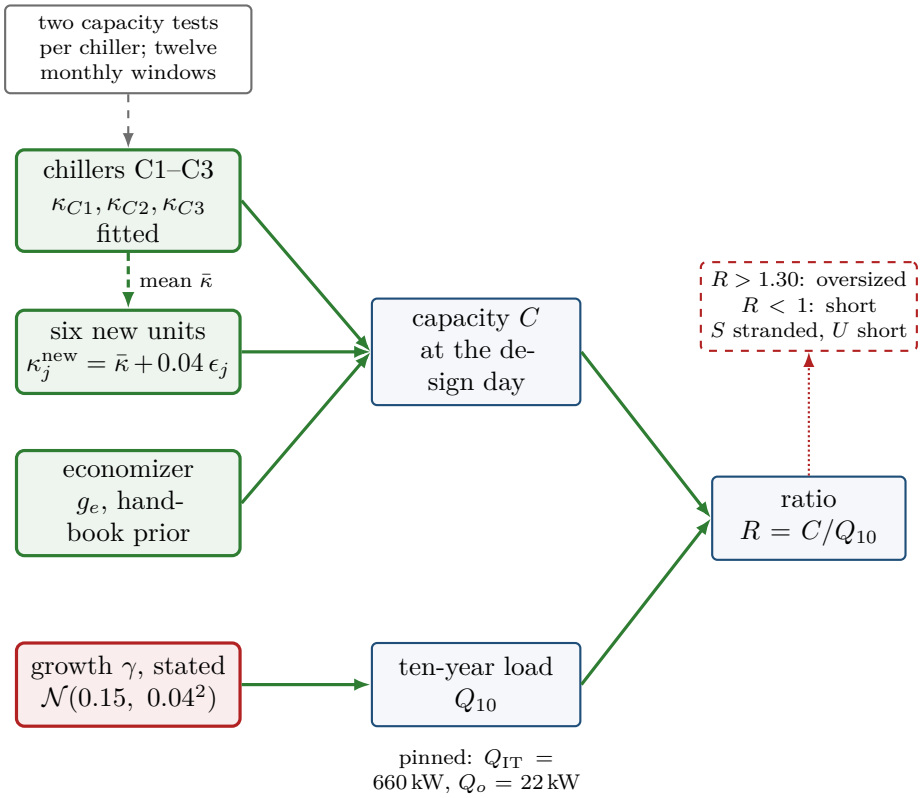


Figure 12.1: The build-out of the site at the design day, wet bulb 22°C. Free (green): the as-built log-ratios of the existing chillers, fitted on the site's record (dashed arrow from the readings); the deviations ϵ_j of the six new units about the existing mean, equation (12.3); the economizer's log-conductance g_e, which keeps its handbook prior. Red: the growth rate, with its stated distribution (12.7) and no row of evidence. Blue: the capacity (12.4), the ten-year load (12.5) and their ratio. The dotted arrow leads to the events and quantities of equation (12.6). Not drawn: the pump, the pipe, the feed and their entries, which do not enter the capacity at the design day.

Table 12.1: The questions asked of the ten-year build-out. $\mathbf{y}$ is the site's record (capacity tests and monthly windows) with its priors, and the stated distributions of the new units' deviations and of the growth; R, S are those of equation (12.6).

Planner's question	General form	Mathematical form	Category	Computation	Answered in
What is the probability that the plant as drawn is more than 30 % oversized at the ten-year load?	$\Pr(f \geq f^* \mid \mathbf{y})$, the upper tail of a ratio	$\Pr(R > 1.30 \mid \mathbf{y})$	prediction under a threshold	posterior	12.3.1
And the probability it is undersized?	$\Pr(f < f^* \mid \mathbf{y})$	$\Pr(R < 1 \mid \mathbf{y})$	prediction under a threshold	posterior	12.3.2
What is the answer made of?	$\text{Var}(f \mid \mathbf{y})$, split by input	$\text{Var}(\ln R) = \text{Var}(\ln C) + \text{Var}(\ln Q_{10})$	estimation	posterior	12.3.3
How much capacity is stranded, and what does it cost?	$\mathbb{E}[L(a, \cdot) \mid \mathbf{y}]$	$\mathbb{E}[S \mid \mathbf{y}]$, annualized	estimation	expected consequence	12.3.4

Which row acts on which entry. The capacity tests act on $(\kappa_{C1}, \kappa_{C2}, \kappa_{C3}, d)$; the monthly windows on (c, l_k, l_e) and, through the chiller power, on the others; nothing acts on g_e, the six ϵ_j or γ but their own prior rows. The gradients of the two sides, by equations (12.4) and (12.5), are

$$\frac{\partial C}{\partial \kappa_{Ci}} = 400\,\mathrm{e}^{\kappa_{Ci}} + \tfrac{1}{3}\sum_{j=1}^{6} 400\,\mathrm{e}^{\kappa_j^{\mathrm{new}}}, \qquad \frac{\partial C}{\partial g_e} = 50\,\mathrm{e}^{g_e}, \qquad \frac{\partial C}{\partial \epsilon_j} = 0.04 \times 400\,\mathrm{e}^{\kappa_j^{\mathrm{new}}},$$
$$\frac{\partial \ln Q_{10}}{\partial \gamma} = \frac{10\,Q_{\mathrm{IT}}\,(1+\gamma)^9}{Q_{10}(\gamma)}, \tag{12.8}$$

every other entry zero: the capacity has no entry in d, c, l_k, l_e or γ, and the load has an entry in γ alone. Three facts are read here before any number. Each existing unit's ratio acts on the capacity with the weight of three units, its own and a third of each of the six new units whose mean it sets, so the capacity tests reach nine units through three. The economizer's entry and the six new units' entries hold only their prior rows, and read as observability they are known only from the handbook and from the spread of the existing units. And the two sides share no entry: the capacity's gradient and the load's have disjoint supports, so under any posterior that does not couple the growth to the site the two sides are independent and the split (12.1) holds.

Modeling choices at the rows. The site's rows carry the choices of Chapter 9: Gaussian, unbiased and independent readings, the capacity tests read as the units' delivered capacity, one year of monthly windows. Two choices are this chapter's own. The new units are of the existing model and scatter about its mean as the existing units do: this lets the record of three units speak for nine, and it fails if a later production run of the model differs, a difference no row could see. And the growth is independent of the plant: no row couples them, which is right for a forecast made from the business and wrong if the plant's own constraints shape the load it is offered.

12.2 The posterior of the state (step 3)

The questions of step 1 owe probabilities and moments over the chapter's uncertain quantities, and this section derives what each owes and computes it. The derivation starts from the probability of the drill question and works down to the posterior of the site, which is Chapter 9's, and the two stated distributions.

12.2.1 What the answer needs from the state

Write $\boldsymbol{\xi} = (\mathbf{z}, \boldsymbol{\epsilon}, \gamma)$ for the fifteen uncertain quantities: the eight entries of the site, the six deviations of the new units and the growth. The drill question owes the probability of the event $R > 1.30$ under their joint distribution given the evidence,

$$\Pr(R > 1.30 \mid \mathbf{y}) = \int \mathbf{1}\big[C(\mathbf{z}, \boldsymbol{\epsilon}) > 1.30\,Q_{10}(\gamma)\big]\, p(\boldsymbol{\xi} \mid \mathbf{y})\, \mathrm{d}\boldsymbol{\xi}. \tag{12.9}$$

The record informs the site's state alone, and the two stated distributions are independent of it and of each other, so the joint distribution factors,

$$p(\boldsymbol{\xi} \mid \mathbf{y}) = p(\mathbf{z} \mid \mathbf{y}) \prod_{j=1}^{6} \phi(\epsilon_j)\ \ \frac{1}{0.04}\,\phi\!\left(\frac{\gamma - 0.15}{0.04}\right), \tag{12.10}$$

with ϕ the standard normal density. The posterior of the site is the only factor that inference supplies; the other two are what the planning inputs state, unchanged by the record.

Table 12.2: The entries of the site's state that act on the build-out's capacity at the design day, from the one fit of the site's record (Chapter 9). The derate is fixed by the record but does not act at the design day, where the derate factor is one.

Entry [unit]	MAP [entry's]	Posterior sd [entry's]	Prior sd [entry's]	Ratio [–]	Truth [entry's]	Informed by
κ_{C1} [–]	−0.0909	0.0158	0.08	0.197	−0.1054	capacity tests and windows
κ_{C2} [–]	−0.0394	0.0175	0.08	0.218	−0.0619	capacity tests and windows
κ_{C3} [–]	−0.1157	0.0189	0.08	0.236	−0.1278	capacity tests and windows
d [K^{-1}]	0.0152	0.0029	0.005	0.577	0.0180	capacity tests; no effect here
g_e [–]	0.0000	0.3500	0.35	1.000	−0.2231	handbook only

The integral over the growth can be done first and in closed form. The load (12.5) rises with γ, so at fixed capacity the ratio exceeds a level r exactly when the load falls below C/r, that is, when the growth falls short of the rate at which the load equals C/r. Solving equation (12.5) for that rate,

$$\gamma_r(C) = \left(\frac{C/r - Q_o}{Q_{\mathrm{IT}}} \right)^{1/10} - 1, \qquad R > r \iff \gamma < \gamma_r(C), \tag{12.11}$$

the breakeven growth of the level r at capacity C. Given the capacity, the probability of the event is the normal distribution function of the growth at the breakeven rate,

$$\Pr(R > r \mid C) = \Phi\!\left(\frac{\gamma_r(C) - 0.15}{0.04} \right), \qquad \Pr(R < r \mid C) = 1 - \Phi\!\left(\frac{\gamma_r(C) - 0.15}{0.04} \right), \tag{12.12}$$

and averaging over the capacity's distribution is the law of total probability of equation (3.2),

$$\Pr(R > 1.30 \mid \mathbf{y}) = \int \Phi\!\left(\frac{\gamma_{1.30}\big(C(\mathbf{z}, \boldsymbol{\epsilon})\big) - 0.15}{0.04} \right) p(\mathbf{z} \mid \mathbf{y}) \prod_j \phi(\epsilon_j)\, \mathrm{d}\mathbf{z}\, \mathrm{d}\boldsymbol{\epsilon}. \tag{12.13}$$

Equations (12.11)–(12.13) are what the answer needs from the load side: nothing but the growth's stated distribution, read at a rate the capacity sets. What it needs from the capacity side is the distribution of $C(\mathbf{z}, \boldsymbol{\epsilon})$: the posterior of $(\kappa_{C1}, \kappa_{C2}, \kappa_{C3}, g_e)$ from the site's fit, and the new units' deviations. The variance split (12.1) needs the same two distributions, and the stranded capacity the expectation of S over all fifteen quantities.

12.2.2 The site's fit, cited

The posterior of the site is the one fit of Chapter 9: the Laplace normal $\mathbf{z} \mid \mathbf{y} \approx \mathcal{N}(\hat{\mathbf{z}}, \boldsymbol{\Sigma})$ of equation (4.38) on the sixty-two rows, fifty-four readings and eight priors, at its MAP state, with twice the objective 59.408 on 54 degrees of freedom, an upper tail of 0.285. Table 12.2 gives the entries that act on the capacity: each row is an entry, the columns its MAP value, posterior spread, prior spread, their ratio and the stated truth from which the record was generated, a check marker for Section 12.4.3.

The three existing chillers are known to about a fifth of their prior spread, and all three sit below their nameplate, by about 4 to 12 % on the log scale. The economizer's guess is untouched by the record, a ratio of 1.000. Two consequences follow for the capacity before any draw. The build-out delivers less than its 3 600 kW of nameplate, because nine units inherit ratios below zero; and the capacity's spread comes from the three ratios, carried through nine units by equation (12.8), from the economizer's guess and from the six new units' scatter.

12.2.3 The draws and the exact posterior

The answers read the joint distribution (12.10) through draws. The $N = 20\,000$ Laplace draws of the site's state of Chapter 9, seed 9, are used as they stand; each is paired with six draws of the new units' deviations and one draw of the growth, from their stated distributions, and pushed through the exact capacity (12.4), the exact load (12.5) and the quantities (12.6). This is the state route. The exact posterior is the same draws under the weights of Chapter 9, each the ratio of the exact posterior of the site to the Laplace normal at the draw, as in equation (5.16); their effective sample size is 19 913 of the 20 000, so the weights are nearly uniform. Only the site's draws are reweighted: the deviations and the growth are drawn from their own distributions exactly, so no weight corrects them. The closed form (12.12) stands beside both: it is exact in the growth and needs a capacity, and pinned at the capacity's median it is the single-side route of this chapter, the analogue of a binding element read alone.

12.3 The answers (step 4)

Every answer is read from the joint draws of Section 12.2.3. The reported value is the state route, the share or the mean of the draws; the exact posterior stands beside it with the gap and the Monte Carlo standard error, and the routes on one side alone stand beside both.

12.3.1 The probability that the build-out is more than 30 percent oversized

Planner's question. What is the probability that the plant as drawn is more than 30 % oversized at the ten-year load? The question asks for the upper tail of the oversizing ratio, with the capacity as built and the load as forecast.

General form. A prediction under a threshold is a tail of a derived quantity under the posterior, equation (2.12),

$$\Pr(f(\boldsymbol{\xi}) \geq f^* \mid \mathbf{y}) = \int \mathbf{1}\big[f(\boldsymbol{\xi}) \geq f^*\big]\, p(\boldsymbol{\xi} \mid \mathbf{y})\, \mathrm{d}\boldsymbol{\xi},$$

here not a requirement to be met but an event to be priced, so no assurance level is set against it.

Specialized form. The question is equation (2.12) with $f = R$ of equation (12.6) and $f^* = 1.30$, over the joint distribution (12.10): the probability (12.9). The reported value is the share of the joint draws in which the event holds,

$$\hat{P}_{1.3} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[C_i > 1.30\, Q_{10,i}\big], \qquad \mathrm{se}(\hat{P}_{1.3}) = \sqrt{\hat{P}_{1.3}\,(1-\hat{P}_{1.3})/N}, \tag{12.14}$$

with C_i and $Q_{10,i}$ the two sides of draw i; the exact posterior sums the weights of the same draws. Beside them stand the routes on one side: the capacity pinned at its median with the growth uncertain, by the draws and by the closed form (12.12); the growth pinned at 15 % with the capacity uncertain; and the sheet's nameplate, 3 600 kW, with the growth uncertain. The delta form stands beside them as well: $\ln R$ taken as normal, with the mean $\ln C - \ln Q_{10}(0.15)$ at the median capacity and the spread of the two sides' first-order variances, equation (4.53).

Building blocks. The site's Laplace draws and their weights (Section 12.2.3); the capacity (12.4); the load (12.5) and the growth's distribution (12.7); the breakeven growth (12.11) and the closed form (12.12).

Table 12.3: The two sides of the oversizing ratio at the design day of year ten, from the joint draws. The capacity is the nine units as built and the economizer; the load is the loop heat of equation (12.5).

Side	Median [kW]	5 % point [kW]	95 % point [kW]	sd [kW]	Set by
Capacity C	3 371.9	3 277.7	3 470.4	58.7	the site's fit, the new units' spread, the economizer's guess
Ten-year load Q_{10}	2 697.8	1 508.9	4 658.1	–	the growth, 0.085 to $0.215\,\mathrm{yr}^{-1}$ at the band's ends

The two sides. Table 12.3 gives the two sides as the draws carry them: each row is a side, the columns its median, the ends of its central 90 % band and its standard deviation, and the last column what sets it. Their densities are drawn on one axis in Figure 12.2: the horizontal axis is the heat on the design day of year ten, the vertical axis the density; the narrow curve is the capacity, the broad one the ten-year load, and the shading marks the loads at which the capacity at its median would be more than 30 % oversized (left) or short (right).

The two sides are of different kinds. The capacity is known to within ±3 % at its band's ends, a narrow peak whose median, 3 371.9 kW, lies 228.1 kW below the nameplate. The ten-year load spreads over a factor of three, from 1 508.9 to 4 658.1 kW, and it is skewed: its band reaches 1 960 kW above its median and 1 189 kW below, the compounding of equation (12.5) at work. Its median, 2 697.8 kW, sits by the sheet's 2 692.1 kW: the load is a monotone function of the growth, so its median is the load at the growth's median, the forecast's 15 %, up to the sampling of the draws. Against that spread the capacity is almost a fixed line, and Figure 12.2 already shows the answer's shape: the gold area left of 2 593.8 kW and the red area right of 3 371.9 kW are both large pieces of the load's density.

The answer. By equation (12.14) the joint draws give the reported value,

$$\Pr(R > 1.30 \mid \mathbf{y}) = 0.4543, \qquad \mathrm{se} = 0.0035, \tag{12.15}$$

and the exact posterior, the same draws reweighted, gives 0.4537, a gap of −0.0006, a sixth of the Monte Carlo standard error. The ratio itself has the median 1.2496 and the central 90 % band [0.7223, 2.2390]. Its density is drawn in Figure 12.3: the horizontal axis is the ratio, the vertical axis its density, with the capacity held at its median; the three shaded regions are the ratio below one, between one and 1.30, and above 1.30, each labeled with its share of the joint draws; the bar above the curve is the reported band with its median.

The density is skewed to the right, as the reciprocal of a convex load must be: the band reaches 0.99 above the median and 0.53 below it, so a symmetric band on the ratio would misplace both tails, and every route here works on the growth, where the distribution is normal, or on the draws. Table 12.4 sets the routes side by side: each row is a route, the columns its probability of the event, its probability of a ratio below one, and the gap of its first column from the reported value.

The build-out is more than 30 % oversized with probability 0.454: about an even chance, not the certainty the sheet's 1.337 implies. Two things separate the sheet from the answer, and the table shows each apart. The first is the as-built capacity. Kept at its nameplate, with the growth uncertain, the drawing is oversized with probability 0.5299; the as-built units deliver 228.1 kW less, and the probability falls by 0.0756. At the forecast's mean growth the as-built median gives a ratio of 3 371.9/2 692.1 = 1.2525, below the threshold the sheet's nameplate

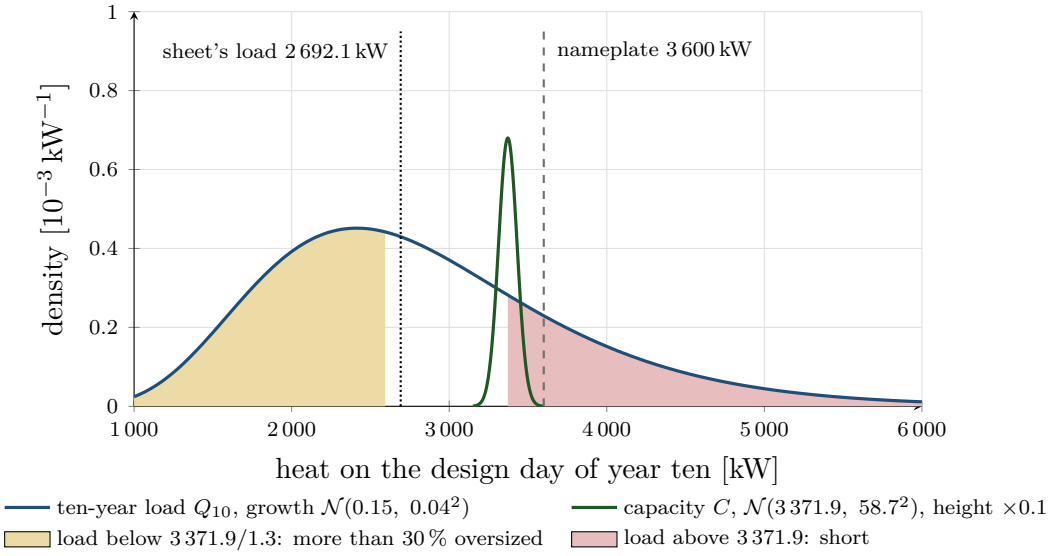


Figure 12.2: The two sides of the oversizing ratio on one axis. Blue: the density of the ten-year load (12.5) with the growth's stated normal, by the change of variables (4.55), $\gamma = ((Q - 22)/660)^{1/10} - 1$. Green: a normal with the capacity's median and standard deviation from Table 12.3, drawn at a tenth of its height so that the load's shape can be read beside it. Shaded: with the capacity at its median, the loads at which the ratio exceeds 1.30 (gold) and falls below one (red). Dotted: the sheet's load at the forecast's mean growth; dashed: the drawing's nameplate.

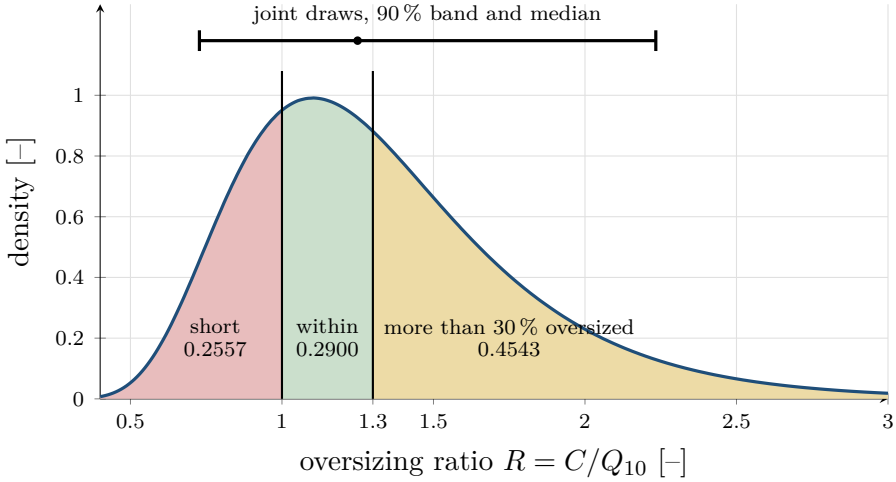


Figure 12.3: The density of the oversizing ratio, with the capacity held at its median 3 371.9 kW and the growth's stated normal carried through $R = 3\,371.9/(660\,(1+\gamma)^{10} + 22)$ by the change of variables. The capacity's own spread, 0.3 % of the variance of $\ln R$ (Section 12.3.3), is left out of the curve. Shaded: the ratio below one (red), between one and 1.30 (green) and above 1.30 (gold), labeled with the shares of the joint draws. Bar: the central 90 % band of the joint draws, [0.7223, 2.2390], and its median 1.2496.

Table 12.4: The probability that the build-out is more than 30 % oversized at the ten-year load, by route. The reported route and the exact posterior take both sides uncertain; the single-side routes pin one side; the delta form takes $\ln R$ as normal. Gaps are route minus reported, in $\Pr(R > 1.30)$.

Route	$\Pr(R > 1.30)$ [–]	$\Pr(R < 1)$ [–]	Gap [–]
Design sheet: nameplate at the forecast's mean, ratio 1.3373	yes, no band	–	–
Nameplate 3 600 kW, growth uncertain	0.5299	–	+0.0756
Joint draws, both sides uncertain (reported)	0.4543	0.2557	–
Exact posterior, both sides uncertain	0.4537	0.2561	−0.0006
Delta form on $\ln R$, both sides uncertain (beside)	0.4571	0.2572	+0.0028
Capacity at its median, growth by the draws	0.4541	0.2551	−0.0002
Capacity at its median, closed form (12.12)	0.4571	0.2548	+0.0028
Growth known at 15 %, capacity uncertain	0.0168	0.0000	−0.4375

clears, so the sheet's own arithmetic, done with the as-built units, would have read the drawing as within the tolerance. The second is the growth, and it is the larger. With the growth known at 15 % the probability is 0.0168, the capacity's narrow tail above $1.30 \times 2\,692.1$ kW; with the growth uncertain it is 0.4543. The capacity at its median with the growth uncertain gives 0.4541 by the same growth draws, within 0.0002 of the reported value: almost the whole answer is the forecast's. The closed form at the median capacity, 0.4571, differs from the draws' 0.4541 at the same capacity by 0.0030, within the Monte Carlo standard error of the growth draws. The delta form, with $\ln R$ normal of mean 0.2252 and spread 0.3454, gives 0.4571 and, on the short side, 0.2572 against the draws' 0.2557. It lands close because the logarithm removes most of the ratio's skew, and what remains, the curvature of $\ln(1+\gamma)$ and the 22 kW of other heat, is what separates it from the draws; it stands beside the report and is not the report.

The closed form also says why the answer sits near one half. By equation (12.11) the median capacity is exactly 30 % oversized at the growth $\gamma_{1.3} = 0.1457\,\mathrm{yr}^{-1}$, and the drawing is more than 30 % oversized whenever the decade's growth falls short of it. The breakeven rate lies 0.0043 below the forecast's mean, a tenth of its spread, so the forecast puts nearly half its probability on each side, $\Phi(-0.0043/0.04)$ by equation (12.12), 0.4571.

12.3.2 The probability that the build-out is short

Planner's question. And the probability it is undersized? The question asks for the lower tail of the same ratio: the drawing's nine units and economizer unable to carry the loop heat of year ten.

General form. The complement of a requirement is a tail of the same kind, equation (2.12), read on the lower side,

$$\Pr(f < f^* \mid \mathbf{y}) = 1 - \Pr(f \geq f^* \mid \mathbf{y}).$$

Specialized form. The question is equation (2.12) with $f = R$ and $f^* = 1$, read on the lower side: $\Pr(R < 1 \mid \mathbf{y})$, the share of the joint draws in which $C_i < Q_{10,i}$. Given the capacity it is the second relation of equation (12.12) with $r = 1$, the growth exceeding the rate $\gamma_1(C)$ at

which the load reaches the capacity. The three regions of the ratio partition the draws,

$$\Pr(R < 1) + \Pr(1 \le R \le 1.30) + \Pr(R > 1.30) = 1. \tag{12.16}$$

Building blocks. The joint draws of Section 12.2.3 and their weights; the capacity (12.4) and the load (12.5) in each; the breakeven growth (12.11) at $r = 1$ and the closed form (12.12).

The short tail. The joint draws give

$$\Pr(R < 1 \mid \mathbf{y}) = 0.2557, \qquad \Pr(1 \le R \le 1.30 \mid \mathbf{y}) = 0.2900,$$

and with the upper tail's 0.4543 the three shares sum to one, as equation (12.16) requires. The exact posterior gives 0.2561 for the short tail, a gap of +0.0004. The single-side routes of Table 12.4 say whose tail it is: with the growth known at 15 % no draw is short, 0.0000; with the capacity at its median the growth draws give 0.2551, and the closed form (12.12) gives 0.2548, the probability that the growth exceeds $\gamma_1 = 0.1764\,\mathrm{yr}^{-1}$, the rate at which the load reaches the median capacity. The drawing that the sheet calls a third too large is short in year ten with probability 0.26, more than one chance in four, and it lands in the band the policy would call right-sized, between one and 1.30, with probability 0.29. It is as likely to be short as to be anywhere in that band.

This is the surprise of the chapter, and the question as asked hides it. The drawing was sized to the forecast's mean, so under an uncertain forecast it inherits both of the forecast's tails, and the upper tail of the ratio, the one finance asked about, is the lower tail of the growth. The lower tail of the ratio is the upper tail of the growth, and compounding makes it the heavier of the two in load: by Table 12.3 the load's band reaches 1 960 kW above its median. The question of oversizing has a lower tail the question did not ask about, and a planner who answers only the question asked reports an even chance of stranded capital and leaves out a one-in-four chance of a plant that cannot carry its load.

12.3.3 What the answer is made of

Planner's question. What is the answer made of? The question asks how the uncertainty of the ratio divides between the plant and the forecast, and so what could narrow it.

General form. The variance of a derived quantity under the posterior, equation (2.10),

$$\mathrm{Var}(f \mid \mathbf{y}) = \mathbb{E}\big[(f - \mathbb{E}[f \mid \mathbf{y}])^2 \mid \mathbf{y}\big],$$

split into one share per uncertain input, the parameter risk of Section 5.3.4.

Specialized form. The question is equation (2.10) with $f = \ln R$, split by equation (12.1) into the capacity's share and the load's. The two sides share no entry, equation (12.8), and the growth is independent of the site, so the covariance term is zero and each share is that side's own variance over the total. To first order the load's share is the growth's variance carried by the last line of equation (12.8),

$$\mathrm{Var}(\ln Q_{10}) \approx \left(\frac{10\, Q_{\mathrm{IT}}\, (1.15)^9}{Q_{10}(0.15)} \right)^2 0.04^2, \tag{12.17}$$

the delta method of equation (4.53): ten years of compounding multiply the growth's spread by about ten on the log scale. The capacity's share is the covariance split (5.25) of $\ln C$ over the site's entries and the new units' deviations, with the gradient of equation (12.8) divided by C. Beside the split stand the answers with one side pinned, which state the shares as probabilities.

Table 12.5: The variance of $\ln R$ split between the two sides, equation (12.1), from the joint draws, and the probabilities of the two tails with each side pinned.

Side (how pinned)	Variance of its log [–]	Share of Var($\ln R$) [%]	$\Pr(R > 1.30)$ with it pinned [–]	$\Pr(R < 1)$ with it pinned [–]
Capacity, $\ln C$ (at its median)	0.00030	0.3	0.4541	0.2551
Ten-year load, $\ln Q_{10}$ (growth at 15 %)	0.11920	99.7	0.0168	0.0000
Ratio, $\ln R$, both uncertain (reported)	0.11959	100	0.4543	0.2557

Building blocks. The joint draws; the logarithms of the two sides in each draw and their variances; the single-side routes of Table 12.4; the breakeven growth.

The budget. Table 12.5 gives the split, each row a side, the columns the variance of its logarithm over the draws, its share of the variance of $\ln R$, and the probabilities of the two tails with that side's uncertainty removed.

The variance of $\ln R$ is 0.11959, of which the capacity contributes 0.00030 and the load 0.11920; the two shares sum to 0.11950, and the remaining 0.00009 is the sample covariance of two sets of draws that are independent by construction. The capacity carries 0.3 % of the ratio's variance and the forecast 99.7 %. The probabilities say the same in the planner's terms: with the capacity known exactly the answer moves from 0.4543 to 0.4541; with the growth known it falls to 0.0168. The short tail reads the same way, 0.2557 to 0.2551 with the capacity known and to 0.0000 with the growth known. The answer is the forecast's, not the plant's. The plant decides only where the breakeven rate sits, $0.1457\,\mathrm{yr}^{-1}$, and the forecast puts probability 0.4571 below it. A plant-side measurement, however good, cannot move a probability made of the forecast, and Section 12.6 returns to that.

12.3.4 The stranded capacity and its cost

Planner's question. How much capacity is stranded, and what does it cost? The question asks for the expected capacity installed beyond the load in year ten and for its yearly cost to the capital budget.

General form. An expected consequence is the posterior expectation of a loss, equation (2.23),

$$\mathcal{R}(a) = \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}] = \int L(a, \mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z},$$

here for the one action on the table, the drawing as it stands.

Specialized form. The question is equation (2.23) with the loss the stranded capacity S of equation (12.6), valued at the installed price of the units it stands for and annualized by the capital recovery factor,

$$\mathbb{E}[S \mid \mathbf{y}] = \frac{1}{N} \sum_{i=1}^{N} \max(C_i - Q_{10,i},\ 0), \qquad K_S = \frac{\mathbb{E}[S \mid \mathbf{y}]}{400\,\mathrm{kW}} \times \mathrm{R\$}\,1\,600\,000 \times 0.131\,\mathrm{yr}^{-1}, \quad (12.18)$$

with the unit price and the factor, 15 years at 10 %, stated planning inputs. The probability that any capacity is stranded is the probability that the ratio exceeds one, $1 - \Pr(R < 1)$, and the expected shortfall $\mathbb{E}[U \mid \mathbf{y}]$ is the same mean taken on the other side.

Table 12.6: The stranded capacity of the build-out in year ten and its annualized cost, from the joint draws; the price and the capital recovery factor are stated inputs.

Quantity	Expression [–]	Value [kW, units, R$/yr or –]
Expected stranded capacity	$\mathbb{E}[\max(C - Q_{10}, 0)]$	733.5
Probability any is stranded	$\Pr(R > 1) = 1 - \Pr(R < 1)$	0.7443
Expected stranded capacity in units	$\mathbb{E}[S]/400\,\text{kW}$	1.83
Its annualized cost	eq. (12.18)	384 354
Expected shortfall	$\mathbb{E}[\max(Q_{10} - C, 0)]$	204.7
Unit price, stated	per 400 kW unit, installed	R$1 600 000
Capital recovery factor, stated	15 years at 10 %	$0.131\,\text{yr}^{-1}$

Building blocks. The joint draws; the quantities (12.6) in each; the unit price and the capital recovery factor.

The stranded capacity. Table 12.6 gives the expectation and its cost, each row a quantity and the columns its expression and its value.

In expectation the build-out strands 733.5 kW in year ten, 1.83 of its nine units, and that idle capacity costs R$ 384 354 a year in capital recovery. Capacity is stranded in 0.7443 of the draws, the complement of the short tail's 0.2557. The expectation is not the stranding at the forecast's mean: it is carried by the draws of slow growth, where the load is far below the capacity, and at the band's lower end, a load of 1 508.9 kW, more than half the build-out is idle. On the other side the plant is short by 204.7 kW in expectation, so the expected excess of capacity over load is $733.5 - 204.7 = 528.8$ kW: a single number that nets a likely surplus against a heavy chance of a deficit, and the reason the two are reported apart. The year-ten figure is a snapshot; a build-out installed early strands more in the years before ten, which the cost above does not count.

12.4 The checks (step 5)

The answers of Section 12.3 rest on the Laplace posterior of the site, on draws, and on a stated spread of the forecast, and each can be tested.

12.4.1 The draws against the exact posterior and the closed form

Reweighted to the exact posterior of the site, the draws give $\Pr(R > 1.30) = 0.4537$ against the reported 0.4543, a gap of -0.0006, a sixth of the Monte Carlo standard error of 0.0035; the posterior of the site departs too little from its Laplace normal to move an answer the site barely enters, and the short tail moves from 0.2557 to 0.2561. The closed form (12.12), exact in the growth at a pinned capacity, gives 0.4571 at the median capacity against the draws' 0.4541 at the same capacity, a gap of 0.0030 that is the sampling of the growth draws. And the three regions of the ratio sum to one, equation (12.16), with the probability that any capacity is stranded, 0.7443, equal to one less the short tail.

12.4.2 Sensitivity to the forecast's spread

The growth's spread is a stated input and the answer is almost wholly made of it, so it is varied: the growth draws are repeated with the spread at 0.02 and $0.06\,\text{yr}^{-1}$, beside the stated 0.04. The two tails are drawn against the spread in Figure 12.4: the horizontal axis is the forecast's

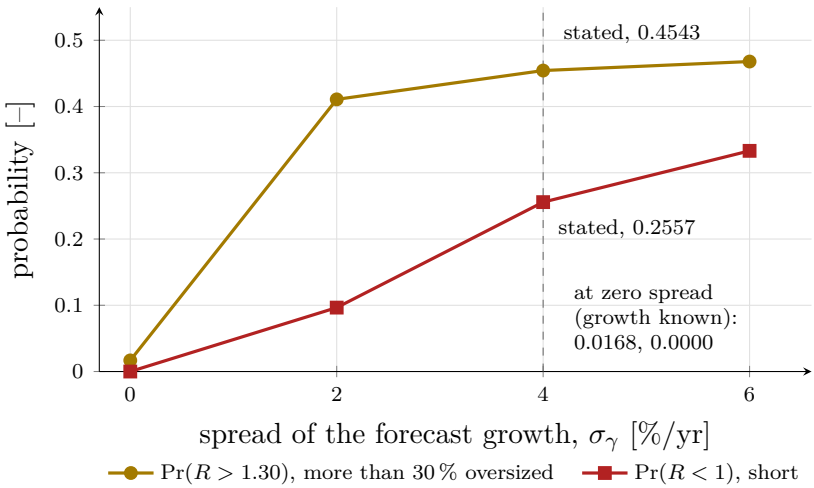


Figure 12.4: The two tails of the oversizing ratio against the spread of the forecast growth, from the joint draws at $\sigma_\gamma = 0.02$, 0.04 (the stated value, dashed) and $0.06\,\mathrm{yr}^{-1}$; the point at zero spread is the growth known at 15 % (Table 12.4), where no draw is short. The lines join the points and carry no values between them.

spread, the vertical axis the probability; the gold line is the upper tail, $R > 1.30$, and the red line the lower tail, $R < 1$, each with its growth-known value at zero spread.

The two tails answer the spread differently. The upper tail rises from 0.0168 with the growth known to 0.4108 at a spread of 0.02, and then barely moves, 0.4543 at 0.04 and 0.4679 at 0.06. The closed form says why: at the median capacity it is $\Phi(-0.0043/\sigma_\gamma)$, which tends to one half as soon as the spread is large against the 0.0043 between the breakeven rate and the forecast's mean. Any forecast uncertain enough to be honest puts the drawing near an even chance of being badly oversized. The lower tail grows with the spread throughout, from 0.0000 with the growth known to 0.0965, 0.2557 and 0.3331: the chance of a plant short in year ten is the quantity the forecast's quality decides. Halving the spread takes it from more than one in four to under one in ten.

12.4.3 The answers against the stated truth

The site's record is generated from a stated truth, and the capacity side can be checked against it; the growth has no truth, since it is a forecast and not part of the record. Table 12.7 collects the checks: each row is a route, a variation or the truth, the columns the capacity, the ratio at the forecast's mean growth, and the two tails. The true build-out, with the new units at the existing units' true mean ratio and the economizer at its true conductance, carries 3 303.2 kW, inside the reported band [3 277.7, 3 470.4] kW and 68.7 kW below its median. The fit puts each existing chiller between 0.6 and 1.3 posterior spreads above its truth (Table 12.2), and the handbook's economizer above its true conductance, and both errors lift the capacity. At the forecast's mean growth the true plant's ratio is 1.2270, below the 1.30 threshold, as the reported median capacity's 1.2525 is: on the capacity side the truth and the answer agree that, at the mean forecast, the drawing is not more than 30 % oversized, and that the sheet's nameplate is what put it over.

12.5 The report (step 6)

The build-out drawing is more than 30 % oversized at the ten-year load with probability 0.45, not with the certainty its sheet implies, and it is short in year ten with probability 0.26. In

Table 12.7: The checks of the oversizing ratio. The reported route is the joint draws; the exact posterior reweights them; the spread variations repeat the draws of the growth; the truth row holds the capacity side at the stated truth of the site's record, with the growth at the forecast's mean.

Route or variation	C [kW]	R at $\gamma = 0.15$ [–]	$\Pr(R > 1.30)$ [–]	$\Pr(R < 1)$ [–]
Joint draws, reported (median capacity)	3 371.9	1.2525	0.4543	0.2557
Exact posterior	–	–	0.4537	0.2561
Gap, exact minus reported	–	–	−0.0006	+0.0004
Monte Carlo standard error	–	–	0.0035	–
Closed form at the median capacity	3 371.9	–	0.4571	0.2548
Delta form on $\ln R$ (beside)	–	–	0.4571	0.2572
Growth spread $0.02\,\mathrm{yr}^{-1}$	–	–	0.4108	0.0965
Growth spread $0.06\,\mathrm{yr}^{-1}$	–	–	0.4679	0.3331
Stated truth of the site, as built	3 303.2	1.2270	–	–
Design sheet, nameplate	3 600	1.3373	–	–

Table 12.8: The report for the ten-year build-out at the design day, with the growth's stated distribution $\mathcal{N}(0.15,\ 0.04^2)\,\mathrm{yr}^{-1}$. Reported values by the joint draws; the exact posterior beside the two tails.

Question	Answer	Band or other routes
What is the probability that the plant as drawn is more than 30 % oversized at the ten-year load?	0.4543 (exact: 0.4537), Monte Carlo standard error 0.0035	ratio median 1.2496, central 90 % band 0.7223 to 2.2390; nameplate 0.5299; growth known 0.0168; spread 0.02 to 0.06: 0.4108 to 0.4679
And the probability it is undersized?	0.2557 (exact: 0.2561); within the policy's band, 0.2900	capacity known 0.2551; growth known 0.0000; spread 0.02 to 0.06: 0.0965 to 0.3331
What is the answer made of?	the forecast: 99.7 % of $\mathrm{Var}(\ln R)$; the capacity 0.3 %	capacity known: 0.4541; breakeven growth $0.1457\,\mathrm{yr}^{-1}$
How much capacity is stranded, and what does it cost?	733.5 kW in expectation, 1.83 units, R$ 384 354/yr	stranded in 0.7443 of the draws; expected shortfall 204.7 kW

expectation it strands 733.5 kW, R$ 384 354 a year. Almost all of that uncertainty is the forecast's.

12.5.1 The answers and their bands

Table 12.8 gives the answers in the order of Table 12.1. Each row is a question; the second column gives the reported answer with the exact posterior beside it where it was computed, and the third its band or its other routes.

12.5.2 The design-review entry

The design review's entry answers finance's question and the one beside it. The build-out drawing of nine 400 kW chillers and the economizer is more than 30 % oversized at the ten-year load with probability 0.4543 by the joint draws, 0.4537 by the exact posterior; the sheet's 33.7 % is a nameplate ratio at the forecast's mean, and with the as-built units at the same mean the drawing is 25 % over, inside the tolerance. The same drawing is short in year ten with

probability 0.2557 (0.2561 exact). Its expected stranded capacity is 733.5 kW, or R$ 384 354 a year at the stated price and capital recovery factor, against an expected shortfall of 204.7 kW. The entry records what the answer rests on: the forecast's stated spread of $0.04\,\mathrm{yr}^{-1}$, which carries 99.7 % of the ratio's variance, and not the plant, whose capacity is known to ±3 %. The second part of the entry follows from it. A drawing whose risk is the forecast's is not made safer by sizing it differently once; it is made safer by not committing all of it at once, so that each unit is ordered when the load has shown itself. The order date of Chapter 11 is the mechanism, a crossing along the load path less the lead time, and the risk–cost curve of Chapter 24 is where the phasing is chosen.

Common mistake

Answering only the tail that was asked about. An oversizing question on an uncertain forecast has two tails, and the drawing sized to the forecast's mean carries both; the report states the probability of being short beside the probability of being oversized.

12.6 The next measurement (step 7)

The report leaves an answer whose variance is 99.7 % the forecast's. A measurement is a row added to the evidence, and what a row on the plant could remove is known before its value is, from the split of Section 12.3.3.

12.6.1 What a measurement on the plant can remove

A candidate measurement a on the plant, a capacity test of an existing chiller, a performance test of a new unit at delivery, a test of the economizer once built, is a row on the capacity's entries, and its preposterior covariance is that of equation (4.67), $\boldsymbol{\Sigma}_a = (\boldsymbol{\Sigma}^{-1} + \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a)^{-1}$, narrowing $\mathrm{Var}(\ln C)$ and nothing else. No row on the plant has a nonzero in the column of γ, equation (12.8), so the best any plant measurement can do is the limit in which the capacity is known exactly, $\mathrm{Var}(\ln C) \to 0$. That limit bounds every candidate: by equation (12.1) it leaves $\mathrm{Var}(\ln R) = 0.11920$ against 0.11959, a standard deviation of $\ln R$ narrower by

$$1 - \sqrt{0.11920/0.11959} = 0.0016,$$

under 0.2 %, and it moves the drill question's answer from 0.4543 to 0.4541. No test on the site narrows this answer, and the ranking of plant-side candidates that Chapter 18 makes by the preposterior has nothing to rank here: every candidate is bounded by the same 0.0002.

12.6.2 The forecast as the next measurement

The quantity that decides the answer is the growth, and the next measurement is a better forecast: a longer record of the site's past load, signed commitments from the tenants who will fill the halls, a staged plan that reveals the growth before the later units are committed. This book does not make forecasts, and it does not value one here. The sensitivity of Section 12.4.2 shows what a better one would buy. Halving the forecast's spread, from 0.04 to $0.02\,\mathrm{yr}^{-1}$, takes the probability of a plant short in year ten from 0.2557 to 0.0965, and moves the probability of being more than 30 % oversized only from 0.4543 to 0.4108. A better forecast narrows the planner's real exposure, the short tail, and leaves the oversizing question near an even chance, because the drawing sits almost exactly at the threshold at the forecast's mean.

12.6.3 The recommendation

Commission no test on the plant for this question: the capacity's 0.3 % of the variance bounds what any test can take off. Spend the effort on the forecast, whose spread decides the short

tail, and on the structure of the commitment, so that the build-out's later units are ordered against the load as it arrives rather than against the forecast of it. Chapter 18 ranks the tests on the existing plant for the questions they do decide, the order date of Chapter 11 among them, whose band is the as-built capacity's.

12.6.4 What the analysis showed

The sheet called the build-out 33.7 % oversized, and the analysis puts the probability that it is more than 30 % oversized at 0.454. Two corrections make the difference. The as-built units deliver 228.1 kW less than their nameplate, so at the forecast's mean the drawing is 25 % over, not 34 %; read by nameplate with the growth uncertain, the probability would be 0.530. And the ten-year load is a forecast compounded ten times, spread over a factor of three, against a capacity known to ±3 %: the forecast carries 99.7 % of the ratio's variance, the breakeven growth sits a tenth of the forecast's spread below its mean, and the answer is close to a coin flip for any honest spread. The surprise is the other tail. The drawing sized to the forecast's mean is short in year ten with probability 0.256, about as likely as landing anywhere in the band the policy calls right-sized, and in expectation it is short by 204.7 kW while it strands 733.5 kW. A ten-year drawing committed at once on an uncertain forecast is exposed on both sides, and no measurement on the plant reduces either exposure.

12.6.5 Limits of the model

A stated forecast. The growth's distribution is a planning input, not inferred; the bands carry it as stated, and an error in its center or its spread moves every answer, the short tail most. A normal on the rate is symmetric in the rate and puts no probability on a contract won or lost at once, a step the normal cannot represent.

One rate for the decade. The growth is one number held for ten years. Growth that varies from year to year, partly independently, would narrow the ten-year load and both tails; growth that slows as the halls fill would bend the path the model compounds.

The design day, all units up. The ratio is taken at the design day with every unit running. Outages, which the enumeration of Chapter 9 counts, would lower the available capacity on some days and raise the short tail's consequence, not the oversizing ratio as defined. The capacity counts the chillers and the economizer only. The shared pipe, which Chapter 11 found binding just after today's three chillers, would have to grow with the build-out, and the pump and the feed with it; the drawing's other elements are the binding constraints of Chapter 13.

The new units and the economizer. The six new units inherit the existing units' mean ratio and spread, which the record of three units cannot test, and the economizer enters at a handbook guess no reading has checked.

A year-ten snapshot. The stranded capacity and its cost are read in year ten. A drawing built at once strands more in every earlier year, and a phased one less; the year-by-year cost is the risk–cost curve's, not this chapter's.

The constructed record. The site's record is generated from a stated truth with the stated sensor errors, not taken from a plant, and the comparison with that truth in Section 12.4.3 is a check of the method on its own model, on the capacity side only.

12.7 What to retain

Definition 12.1 (Oversizing ratio, its two tails and stranded capacity)**.** The oversizing ratio of a drawing is its capacity at the design day over the load at the horizon, $R = C/Q_{10}$. When the load is a monotone function of an uncertain growth rate γ, the event $R > r$ is the event that the growth falls short of the breakeven rate $\gamma_r(C)$ at which the load equals C/r, so that given the capacity the probability is the growth's distribution function there, and the answer is that probability averaged over the posterior of the capacity. The lower tail $R < 1$ is read from the same draws. On the logarithmic scale the ratio is a difference, and with the two sides independent its variance is the sum of theirs. The stranded capacity is $\max(C - Q_{10}, 0)$, reported beside the expected shortfall, never netted against it.

Four conclusions carry beyond the build-out. A ratio of two uncertain sides belongs to the side with the larger variance on the log scale, and the split (12.1) says which before any measurement is bought. A tail of a ratio on a compounded forecast is best read on the growth, where it is normal, through the breakeven rate; a symmetric band on the ratio misplaces both tails of a skewed quantity. A drawing sized to the mean of an uncertain forecast carries both of its tails, and an oversizing question has a shortfall question beside it. And a nameplate ratio overstates both the margin and the stranding: the as-built units are the plant.

Volume I has walked one method through twelve chapters and three plants. Part I taught the questions, the probability rules and the eight steps on one hall; Part II read hall B's headroom, its growth path, its design day and its diversity from one fit of that hall; Part III brought in the site's chilled-water plant and asked how much to build, in what units, when to order and whether the drawing overbuilds. In every chapter the answer was a derived quantity of a fitted state, reported with its band, its budget and the measurement that would narrow it. Volume II keeps the site and the method, and turns to where the plant's limits lie (Part IV, Chapters 13–16), what input uncertainty costs (Part V, Chapters 17–19), how the design holds under failure (Part VI, Chapters 20–22) and what its risk is worth (Part VII, Chapters 23 and 24).

Table 12.9 lists the terms of the chapter with their expressions, units and the subsection that defines each.

Table 12.9: The terms of oversizing and stranded capacity and where each is defined. Units for the site's build-out.

Term	Expression	Units	Meaning	Defined in
Log split	$\mathrm{Var}(\ln R) = \mathrm{Var}(\ln C) + \mathrm{Var}(\ln Q_{10})$	–	each side's share of a ratio	12.1.1
New unit	$\kappa_j^{\mathrm{new}} = \bar{\kappa} + 0.04\,\epsilon_j$	–	as-built ratio of a unit not built	12.1.2
Build-out capacity	eq. (12.4)	kW	nine units and the economizer	12.1.2
Ten-year load	$Q_{10} = Q_{\mathrm{IT}}(1+\gamma)^{10} + Q_o$	kW	loop heat in year ten	12.1.2
Oversizing ratio	$R = C/Q_{10}$	–	capacity over load	12.1.2
Stranded, short	$S = \max(C - Q_{10}, 0)$, $U = \max(Q_{10} - C, 0)$	kW	idle or missing capacity	12.1.2
Stated growth	$\gamma \sim \mathcal{N}(0.15,\ 0.04^2)$	yr^{-1}	a planning input, not inferred	12.1.2
Posterior and draws				
Factored distribution	eq. (12.10)	–	the site's posterior and two stated inputs	12.2.1
Breakeven growth	$\gamma_r(C) = ((C/r - Q_o)/Q_{\mathrm{IT}})^{1/10} - 1$	yr^{-1}	growth at which $R = r$	12.2.1
Tail given capacity	$\Pr(R > r \mid C) = \Phi((\gamma_r(C) - 0.15)/0.04)$	–	closed form in the growth	12.2.1
Joint draws	site draws paired with $\boldsymbol{\epsilon}_i$, γ_i	–	the state route	12.2.3
Questions and answers				
Oversized share	$\hat{P}_{1.3} = N^{-1}\sum_i \mathbf{1}[C_i > 1.30\,Q_{10,i}]$	–	reported probability	12.3.1
Three regions	eq. (12.16)	–	short, within, oversized	12.3.2
Load's log variance	eq. (12.17)	–	compounding of the growth's spread	12.3.3
Stranded cost	eq. (12.18)	R$/yr	annualized idle capital	12.3.4
Plant-side bound	$\mathrm{Var}(\ln C) \to 0$	–	the most a plant test can remove	12.6.1

12.8 Exercises

Exercise 12.1. *Formulation.* Write the planner's second question, whether the build-out is short in year ten, as a probability over the chapter's uncertain quantities, give the event in terms of the growth at fixed capacity, and state its answer. ($\Pr(R < 1 \mid \mathbf{y}) = \int \mathbf{1}[C(\mathbf{z}, \boldsymbol{\epsilon}) < Q_{10}(\gamma)]\, p(\boldsymbol{\xi} \mid \mathbf{y})\, \mathrm{d}\boldsymbol{\xi}$; by equation (12.11) with $r = 1$ the event at fixed capacity is $\gamma > \gamma_1(C) = ((C-22)/660)^{1/10} - 1$, the growth exceeding the rate at which the load reaches the capacity; the joint draws give 0.2557.)

Exercise 12.2. *Derivation.* Derive the breakeven growth $\gamma_{1.3}(C)$ and the probability of the drill question at a pinned capacity, and check them at the median capacity. ($R > 1.30$ means $660\,(1+\gamma)^{10} + 22 < C/1.30$; the load rises with γ, so this holds exactly when $\gamma < ((C/1.30 - 22)/660)^{1/10} - 1$, and with $\gamma \sim \mathcal{N}(0.15,\ 0.04^2)$ its probability is $\Phi((\gamma_{1.3}(C) - 0.15)/0.04)$. At the median capacity 3 371.9 kW the breakeven growth is $0.1457\,\mathrm{yr}^{-1}$, and the probability is 0.4571.)

Exercise 12.3. *Derivation.* Show from the gradients of the two sides why the variance of $\ln R$ is the sum of the two sides' variances, and state the split. (By equation (12.8) the capacity's gradient has entries only in κ_{C1}, κ_{C2}, κ_{C3}, g_e and the ϵ_j, the load's only in γ; the joint distribution (12.10) makes γ independent of all of them, so $\mathrm{Cov}(\ln C, \ln Q_{10}) = 0$ and equation (12.1) holds. The draws give $\mathrm{Var}(\ln R) = 0.11959$, of which the capacity contributes 0.00030 and the load 0.11920.)

Exercise 12.4. *Application.* Finance asks what the expected idle capacity of the build-out costs a year at the stated price of R$ 1 600 000 per 400 kW unit and the capital recovery factor of $0.131\,\mathrm{yr}^{-1}$. (By equation (12.18) the expected stranded capacity of 733.5 kW is $733.5/400 = 1.83$ units, and $1.83 \times$ R$ 1 600 000 $\times 0.131$ with the unrounded unit count gives R$ 384 354 a year.)

Exercise 12.5. *Application.* The planning team offers a forecast with half the spread, $0.02\,\mathrm{yr}^{-1}$. Say which tail it moves and why, and give both. (At the median capacity the upper tail is $\Phi(-0.0043/\sigma_\gamma)$, near one half for any spread large against the 0.0043 between the breakeven rate and the forecast's mean, so it moves little; the lower tail is the growth exceeding $\gamma_1(C)$, well above the mean, and falls fast as the spread narrows. The draws give $\Pr(R > 1.30) = 0.4108$ and $\Pr(R < 1) = 0.0965$.)

Appendix A

Atlas of planning questions

This appendix can be read independently of the chapters. Each row is a question a planner asks. The first column gives the question in the planner's words, as the chapter's questions table gives it; the second the general form of its category, written out or cited by equation number; the third its mathematical form for the plant of that chapter, once the plant graph and its state $\mathbf{z}$ are fixed, in that chapter's symbols; the fourth its category and the fifth its computation, the five categories and the five computations of Section 2.3.6; and the last the section in which the question is answered, or the chapter when a whole chapter answers it. The rows of Chapter 2 are the forms of its classification, Table 2.11. Chapter 3 develops the probability rules the rows rest on and holds no questions table, so it has no rows. The rows grow as chapters are added.

Planner's question	General form	Mathematical form	Category	Computation	Answered in
Chapter 1. Planned Plants as Graphs					
What is the hall air temperature at today's load?	$\mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}]$, $\mathrm{Var}(f(\mathbf{z}) \mid \mathbf{y})$	$p(T \mid \mathbf{y})$, $T = z_1$	estimation	posterior	1.3.1
What is the effective conductance of the air path?	$\mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}]$, $\mathrm{Var}(f(\mathbf{z}) \mid \mathbf{y})$	$p(G \mid \mathbf{y})$, $G = G_{\mathrm{ref}}\, \mathrm{e}^{z_2}$	estimation of a derived quantity	posterior	1.3.1
Is the hall below 24°C now?	$\Pr(f(\mathbf{z}) < f^* \mid \mathbf{y})$	$\Pr(T < 24^\circ\mathrm{C} \mid \mathbf{y})$	prediction under a threshold	posterior	1.3.2
Will the hall carry its 60 kW design load inside 27.5°C?	$\Pr(f(\mathbf{z}) < f^* \mid \mathbf{y})$	$\Pr(T_2 < 27.5^\circ\mathrm{C} \mid \mathbf{y})$, $T_2 = T_c + (Q_2/G_{\mathrm{ref}})\, \mathrm{e}^{-z_2}$	prediction under a threshold	posterior	1.3.2
Would a measurement before the sign-off narrow the answers?	preposterior covariance $(\boldsymbol{\Sigma}^{-1} + \mathbf{J}_a^{\mathsf{T}}\mathbf{J}_a)^{-1}$	$\boldsymbol{\Sigma}_a$ for each candidate row	design	preposterior	1.6.1
Chapter 2. The Language of Capacity Questions					
What is unknown, and how well is it known?	(2.9), (2.10)	$p(z_i \mid \mathbf{y})$, $\mathbb{E}[z_i \mid \mathbf{y}]$	estimation	posterior	2.4.1
Does a requirement hold, with the stated assurance?	(2.12)	$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y}) \geq \kappa$, f^* the requirement	prediction under a threshold	posterior	2.4.2
What happens over the coming peak?	(2.13)	$\int \Pr(A_{\mathrm{future}} \mid \mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, \mathrm{d}\mathbf{z}$	prediction under a threshold	posterior	2.4.4

Planner's question	General form	Mathematical form	Category	Computation	Answered in
Is it rising, and when does it cross its limit?	(2.21)	$\Pr(\gamma > 0 \mid \mathbf{y})$; $\Pr(t^* \le s \mid \mathbf{y}) = \Pr(f_s(\mathbf{z}) \ge f^* \mid \mathbf{y})$	prediction under a threshold for the sign and a crossing by a date; estimation for a forecast and the distribution of a crossing time	trend	2.3.4
Which explanation is correct?	(2.14), (2.15)	$\Pr(M_j \mid \mathbf{y}) \propto p(\mathbf{y} \mid M_j)\,\Pr(M_j)$	hypothesis comparison; diagnosis over a family of faults	evidence	2.5.1, 2.5.2, 2.5.3
What is the expected consequence?	(2.23)	$\mathcal{R}(a) = \mathbb{E}[L(a,\mathbf{z}) \mid \mathbf{y}] = \int L(a,\mathbf{z})\,p(\mathbf{z} \mid \mathbf{y})\,\mathrm{d}\mathbf{z}$	estimation of a derived quantity	expected consequence	2.5.4
What should be built or measured?	(2.23), (2.16)	$\arg\min_a \mathcal{R}(a)$; EVSI	design	expected consequence; preposterior	2.5.5, 2.6.1, 2.6.2
Chapter 4. The Anatomy of Probabilistic Inference					
What is the thermal resistance of the hall's air path?	$\Pr(f(\mathbf{z}) \le f^* \mid \mathbf{y})$, moments	$p(\rho \mid \mathbf{y})$: $\hat{z}_1$, $\sqrt{\Sigma_{11}}$	estimation	posterior	4.3.6
How warm is the hall at the 30 kW test?	$\Pr(f(\mathbf{z}) \le f^* \mid \mathbf{y})$, moments	$p(T_1 \mid \mathbf{y})$: $\hat{z}_2$, $\sqrt{\Sigma_{22}}$	estimation	posterior	4.3.6
How warm will the hall be at 60 kW?	$\Pr(f(\mathbf{z}) \le f^* \mid \mathbf{y})$, moments, quantiles	$p(T_2 \mid \mathbf{y})$: $\hat{z}_3$, $\sqrt{\Sigma_{33}}$, credible interval	estimation	posterior	4.3.6
Can the hall take 60 kW and stay below 28°C?	$\Pr(f(\mathbf{z}) \ge f^* \mid \mathbf{y})$	$\Pr(T_2 < T_{\max} \mid \mathbf{y})$, with $f = -T_2$	prediction under a threshold	posterior	4.3.7
Would the hall stay below 28°C at a load it has not yet carried?	$\Pr(f(\mathbf{z}) \ge f^* \mid \mathbf{y})$ at a scenario node	$\Pr(T^*(Q^*) \le T_{\max} \mid \mathbf{y})$, $T^* = T_c + Q^* z_1$, with $f = -T^*$	prediction under a threshold	posterior	4.3.8
Chapter 5. Thermal Headroom					
How much load can hall B take before an inlet crosses the limit, and how sure is that number?	$\Pr(f(\mathbf{z}) \le f^* \mid \mathbf{y})$, quantiles	$Q^\dagger$ and $H = Q^\dagger - Q_B$: median and 90 % band	estimation	posterior	5.3.1
Which rack binds first?	$\Pr(f(\mathbf{z}) \ge f^* \mid \mathbf{y})$ over a family	$\Pr(\arg\min_k Q_k^\dagger = j \mid \mathbf{y})$, $j = \mathrm{B1}, \ldots, \mathrm{B5}$	prediction under a threshold	posterior	5.3.2
Can hall B carry its 340 kW design load with 95 % assurance?	$\Pr(f(\mathbf{z}) \ge f^* \mid \mathbf{y}) \ge \kappa$	$\Pr(Q^\dagger \ge 340\,\mathrm{kW} \mid \mathbf{y})$ against $\kappa = 0.95$	prediction under a threshold	posterior	5.3.3
What is the band made of?	$\mathrm{Var}(f(\mathbf{z}) \mid \mathbf{y})$, split by entry	the budget of $\mathbf{g}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{g}$	estimation	posterior	5.3.4

Planner's question	General form	Mathematical form	Category	Computation	Answered in
Chapter 6. Load Growth Scenarios					
If the IT load grows 15 % a year, what is the probability the plant is still inside its envelope in year three?	$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y})$ at a scenario node	$\Pr(Q^\dagger \geq Q_3 \mid \mathbf{y})$, $Q_3 = 456.26\,\mathrm{kW}$	prediction under a threshold	posterior	6.3.1
And in each year before and after?	the same, per node	$\Pr(Q^\dagger \geq Q_t \mid \mathbf{y})$, $t = 0, \ldots, 5$	prediction under a threshold	posterior	6.3.2
In what year does the hall leave its envelope?	$\Pr(f(\mathbf{z}) \leq f^* \mid \mathbf{y})$, quantiles	$t^\dagger$: median and 90 % band	estimation	posterior	6.3.3
How much does the answer depend on the growth rate assumed?	the same at stated inputs	$\Pr(Q^\dagger \geq Q_3 \mid \mathbf{y})$ at $\gamma = 12,\ 15,\ 18\,\%/\mathrm{yr}$	prediction under a threshold	posterior	6.3.4
Chapter 7. Design-Day Assessment					
On the design-day ambient, with every unit running, what is the probability of a violation anywhere?	$\Pr(\max_k f_k(\mathbf{z}) > f^* \mid \mathbf{y})$	$\Pr(\max_k T_k^* > 27^\circ\mathrm{C} \mid \mathbf{y})$	prediction under a threshold	posterior	7.3.2
How warm is each inlet on the design day, and how sure?	moments, $\mathbb{E}[f]$ and $\mathrm{Var}(f)$	T_k^*, its sd, and $\Pr(T_k^* > 27^\circ\mathrm{C} \mid \mathbf{y})$ per rack	estimation	posterior	7.3.1
Which rack violates, if one does?	$\Pr(A \mid B)$	$\Pr(T_k^* > 27^\circ\mathrm{C} \mid V,\ \mathbf{y})$	prediction under a threshold	posterior	7.3.3
When does the joint answer differ from the worst rack alone?	the joint tail against its bounds	joint, worst marginal, union bound, independence product; the inlets' correlations	prediction under a threshold	posterior	7.3.4, 7.3.5
Chapter 8. Load Diversity and Coincident Peaks					
The racks never all peak at once; what is the probability the coincident peak exceeds the plant's capacity?	$\Pr(y_{\mathrm{new}} > f(\mathbf{z}) \mid \mathbf{y})$, a future value against a derived quantity	$\Pr(S^* > Q^\dagger \mid \mathbf{y})$	prediction under a threshold	posterior	8.3.1
How correlated are the racks?	$\Pr(z_i \leq z^* \mid \mathbf{y})$, quantiles	ρ: MAP and 90 % band	estimation	posterior	8.3.2
What is the coincidence factor at the design population?	a derived quantity of the state	99th-percentile peak over the sum of the racks' 99th percentiles	estimation	posterior	8.3.3
How much does the covariance move the answer, and how much the capacity band?	conditional probabilities; the spread of a probability	$\Pr(S^* > Q^\dagger \mid \rho, \mathbf{y})$ for $\rho = 0$ to 1; the spread of $\Pr(S^* > Q^\dagger \mid \boldsymbol{\theta}, \mathbf{z})$ with ρ or $Q^\dagger$ known	estimation	posterior	8.3.4
Chapter 9. Redundancy Level					
N+1 or N+2?	$\arg\min_a \mathbb{E}[L(a, \boldsymbol{\theta}) \mid \mathbf{y}]$	annual shed cost saved, $200\,\mathrm{h} \times \mathrm{R\$}\,1\,500/\mathrm{kWh} \times (E_a - E_{a+1})$, against c_u, for N to N+1 and N+1 to N+2	design	expected consequence	9.3.3

Planner's question	General form	Mathematical form	Category	Computation	Answered in
What is each drawing's service availability, and how sure?	$\Pr(f \geq f^* \mid \mathbf{y})$ over an enumeration	$\mathbb{E}[\sum_{\mathbf{s}} P(\mathbf{s})\,\mathbf{1}(\mathrm{Cap}(\mathbf{s}) \geq 1\,000\,\mathrm{kW}) \mid \mathbf{y}]$, with the 90 % band of U_a	prediction under a threshold	posterior	9.3.2
How much capacity does each carry, all up and in each state?	moments	H_a with all units up, median and band; $\mathrm{Cap}(\mathbf{s})$ by the number of units down	estimation	posterior	9.3.1
What is the availability band made of?	conditional variance	$\mathrm{sd}(U_{\mathrm{N}+1})$ with q known and with the capacity known	estimation	posterior	9.3.4
Chapter 10. Unit Sizing					
Two 600 kW units or three 400 kW?	$\arg\min_a \mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$	$\mathbb{E}[L(a, \mathbf{z}) \mid \mathbf{y}]$ for both drawings, and $\Pr(D < 0 \mid \mathbf{y})$	design	expected consequence	10.3.4
Which is more available, and at what load?	$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y})$ over an enumeration	$\mathbb{E}[U(Q^*) \mid \mathbf{y}]$ for $Q^* = 682$ to 1 000 kW; $\Pr(U_{3\times400} < U_{2\times600} \mid \mathbf{y})$	prediction under a threshold	posterior	10.3.1
What headroom does each leave?	$\Pr(f(\mathbf{z}) \leq f^* \mid \mathbf{y})$, quantiles	H_{all} and H_{-1} at 1 000 kW: median and 90 % band	estimation	posterior	10.3.2
How much does efficiency cost?	$\mathbb{E}[f(\mathbf{z}) \mid \mathbf{y}]$ and quantiles of a derived quantity	$p_E\,(E_{3\times400} - E_{2\times600})$: mean and 90 % band	estimation	posterior	10.3.3
Is the difference real?	$\Pr(f(\mathbf{z}) > f^* \mid \mathbf{y})$	$\Pr(D < 0 \mid \mathbf{y})$, with the 90 % band of D	prediction under a threshold	posterior	10.3.5
Chapter 11. Expansion Timing					
When does the next unit have to be ordered so that it is in before the probability of a shortfall passes 5 %?	first crossing of a tail along a scenario chain	$\min\{m : P_m \geq 0.05\} - \ell$	prediction under a threshold	posterior	11.3.2
How does the shortfall probability grow along the path?	$\Pr(f(\mathbf{z}) \geq f^* \mid \mathbf{y})$ at each node	$P_m = \mathbb{E}[\pi_m \mid \mathbf{y}]$, month by month	prediction under a threshold	posterior	11.3.1
How sure is the date?	quantiles of a derived quantity	$m^\dagger(\mathbf{z}, q)$: median and 90 % band	estimation	posterior	11.3.3
What sets the date?	moments of a derived quantity	the all-up capacity's band against the plateau of one unit out	estimation	posterior	11.3.4
Chapter 12. Oversizing and Stranded Capacity					
What is the probability that the plant as drawn is more than 30 % oversized at the ten-year load?	$\Pr(f \geq f^* \mid \mathbf{y})$, the upper tail of a ratio	$\Pr(R > 1.30 \mid \mathbf{y})$	prediction under a threshold	posterior	12.3.1
And the probability it is undersized?	$\Pr(f < f^* \mid \mathbf{y})$	$\Pr(R < 1 \mid \mathbf{y})$	prediction under a threshold	posterior	12.3.2
What is the answer made of?	$\mathrm{Var}(f \mid \mathbf{y})$, split by input	$\mathrm{Var}(\ln R) = \mathrm{Var}(\ln C) + \mathrm{Var}(\ln Q_{10})$	estimation	posterior	12.3.3

How much capacity is stranded, and what does it cost?	$\mathbb{E}[L(a, \cdot) \mid \mathbf{y}]$	$\mathbb{E}[S \mid \mathbf{y}]$, annualized	estimation	expected consequence	12.3.4