Part I

Methods

1

Foundations and algorithms

John Skilling

Why and how – simply – that's what this chapter is about.

1.1 Rational inference

Rational inference is important. By helping us to understand our world, it gives us the predictive power that underlies our technical civilization. We would not function without it. Even so, rational inference only tells us *how* to think. It does not tell us *what* to think. For that, we still need the combination of creativity, insight, artistry and experience that we call intelligence.

In science, perhaps especially in branches such as cosmology, now coming of age, we invent models designed to make sense of data we have collected. It is no accident that these models are formalized in mathematics. Mathematics is far and away our most developed logical language, in which half a page of algebra can make connections and predictions way beyond the precision of informal thought. Indeed, one can hold the view that frameworks of logical connections are, by definition, mathematics. Even here, though, we do not find absolute truth. We have conditional implication: 'If axiom, then theorem' or, equivalently, 'If not theorem, then not axiom'. Neither do we find absolute truth in science.

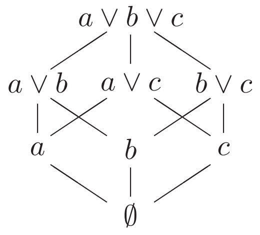
Our question in science is not 'Is this hypothetical model true?', but '*Is this model better than the alternatives?*'. We could not recognize absolute truth even if we stumbled across it, for how could we tell? Conversely, we cannot recognize absolute falsity. If we believe dogmatically enough in a particular view, then no amount of contradictory data will convince us otherwise, if only because the data could be dismissed as evidence of conspiracy to deceive. Yet even a determined sceptic might be sufficiently charitable to acknowledge that a model with demonstrable ability to predict future effects could have practical value.

Let us, then, avoid the philosophical minefields of belief and truth, and pay attention to what we really need, which is predictive ability. We anticipate the Sun will rise tomorrow, not just because it always has done so far, but because this is predicted by models of stellar structure and planetary dynamics, which accord so well with such a variety of data that perceived failure of the Sun to rise might more likely be hallucination.

Rational assessment of different models is the central subject of Bayesian methods, so called after Revd Thomas Bayes, the eighteenth century clergyman generally associated with the beginnings of formal probability theory. We will find that probability calculus is forced upon us as the only method which lets us learn from data irrespective of their order – surely a required symmetry. We will also discover how to use it properly, with the aid of modern computers and algorithms. Inference was held back for a century by technical inability to do the required sums, but that sad era has closed.

1.2 Foundations

Suppose we are given a choice of basic models, $a = apple$, $b = banana$, $c = cherry$, which purport to explain some data. Such models can be combined, so that $apple$-OR-$banana$, written $a \vee b$, is also meaningful. Data that excluded cherries would, in fact, bring us down from the original $apple$-OR-$banana$-OR-$cherry$ combination to just that choice. With n basic models, there are 2^n possible combinations. They form the elements of a lattice, ranging from the absurdity in which none of the models is allowed, up to the provisional truism in which all of them remain allowed. In inference, we need to be able to navigate these possibilities as we refine our knowledge.



The absurdity $\emptyset$ is introduced merely because analysis is cleaner with it than without it, rather as 0 is often included with the positive integers.

The three core concepts of *measure*, *information* and *probability* all have wider scope than inference alone. They apply to lattices in general, whether or not the lattice fills out all 2^n possibilities. By exposing just the foundation that we need, and no more, we can allow wider application, as well as clarifying the basis so that alternative formulations of these concepts become even less plausible than they may have been before.

1.2.1 Lattices

The critical idea we need is ‘partial ordering’. We always have “$=$”: every element equals itself, $x = x$. Sometimes, we have “$<$”, as in $x < y$, meaning that y includes x. In inference, we say that *apple* is included in *apple*-`OR`-*banana*, because the scope of the latter is wider and includes all of the former, but we would not try to include *apple* within *banana*. We don’t need that particular motivation, though. All we need is “$<$” in the abstract. Ordering is to be transitive,

$$x < y \text{ and } y < z \;\text{ implies }\; x < z, \tag{1.1}$$

otherwise it would not make sense.

The other idea we need is ‘least upper bound’. The upper bounds to elements x and y are those elements at or including both x and y. If there is a least such bound, we write it as $x \vee y$ and call it the least upper bound:

$$\left\{ \begin{array}{c} x \leq x \vee y \\ y \leq x \vee y \end{array} \right\}, \text{ and } x \vee y \leq u \text{ for all } u \text{ obeying } \left\{ \begin{array}{c} x \leq u \\ y \leq u \end{array} \right\}. \tag{1.2}$$

In inference, the unique least upper bound $x \vee y$ is that element including all the components of x and y, but no more. There, the existence of least upper bound is obvious.

Technically, a *lattice* is a partially ordered set with least upper bound, so that “$<$” and “$\vee$” are defined. Any pair of elements x and y also has lower bounds, being all those elements at or beneath both. There is a unique greatest lower bound, written $x \wedge y$. (If there were alternatives u and v, then $u \vee v$ could be ambiguously x or y, contradicting uniqueness of their least upper bound.) Mathematicians (Klain & Rota 1997) traditionally define a lattice in terms of $\vee$ and $\wedge$, but our use of $<$ and $\vee$ is equivalent, and (with $=$) underlies their traditional axioms of reflexivity, antisymmetry, transitivity, idempotency, commutativity, associativity and absorption. Of these, the associativity property

$$(x \vee y) \vee z = x \vee (y \vee z) \tag{1.3}$$

is of particular importance to us.

What we now seek is a numerical *valuation* $v(x)$ on our lattice of models, so that we can rank the possibilities. Remarkably, there is only one way of conforming to lattice structure, and this leads us to measure theory, thence to information and probability. Though modernized following Knuth (2003), the approach dates back to Cox (1946, 1961).

1.2.2 Measure

Addition

For a start, we want valuations to conform to "$\leq$", so we require

$$x \leq y \quad \Longrightarrow \quad v(x) \leq v(y). \tag{1.4}$$

Moreover, whatever our valuations were originally, we can shift them to give a standard value 0 to the ubiquitous absurdity $\emptyset$, so that the range of value becomes $0 = v(\emptyset) \leq v(x)$.

We next assume that if x and y are disjoint, so that $x \wedge y = \emptyset$ and they have nothing in common, then the valuation $v(x \vee y)$ should depend only on $v(x)$ and $v(y)$. Write this relationship as a binary operation $\oplus$,

$$v(x \vee y) = v(x) \oplus v(y) \ \text{ when } \ x \wedge y = \emptyset\,. \tag{1.5}$$

To conform with associativity (1.3), we require

$$\big(v(x) \oplus v(y)\big) \oplus v(z) = v(x) \oplus \big(v(y) \oplus v(z)\big). \tag{1.6}$$

This has to hold for arbitrary values $v(x)$, $v(y)$, $v(z)$, and the *associativity theorem* (Azcél 2003) then tells us that there must be some invertable function F of our valuations v such that

$$F\big(v(x \vee y)\big) = F\big(v(x)\big) + F\big(v(y)\big). \tag{1.7}$$

That being the case, we are free to discard the original valuations v and use $m = F(v)$ instead, for which $\vee$ is simple addition: for disjoint x and y we have the *sum rule*

$$\boxed{m(x \vee y) = m(x) + m(y)} \tag{1.8}$$

In other words, valuation can without loss of generality be taken to be what mathematicians call a '*measure*'. They traditionally define measures from the outset as additive over infinite sets, but offer little justification. Mathematicians just do it. Physicists want to know why. Here we see that there's no alternative, and although we start finite we can extend to arbitrarily many elements; it is the same structure. This is *why* measure theory works – it is because of associativity – and we physicists don't have to worry about the infinite.

Assignment

As for actual numerical values, we can build them upwards by addition – except for foundation elements that are not equal to any least upper bound of different elements. Those values alone cannot be determined by the sum rule.

Thus, in the inference example, we can value an apple, a banana and a cherry arbitrarily, but there is a scale on which combinations add. On that scale, if an apple costs 3¢ and a cherry costs 4¢, then their combination costs $3 + 4 = 7$¢, not $3^2 + 4^2 = 25$ or other non-linear construction. Associativity underlies money. Personal assignments may be on a different scale. In economics, for example, personal benefit is sometimes held to be logarithmic in money, $m = \log(\$)$, to reflect the asymmetry between devastating downside risk and comforting upside reward. On that scale, money combines non-linearly, as $\log\$(x \vee y) = \log\$(x) + \log\$(y)$. That's permitted, the point being that there *is* a scale on which one's numbers add. Thus quantification is intrinsically linear – because of associativity.

In inference, $\vee$ behaves as logical OR and $\wedge$ as logical AND, obeying the extra property of distributivity:

$$\begin{aligned}(x \texttt{ OR } y) \texttt{ AND } z &= (x \texttt{ AND } z) \texttt{ OR } (y \texttt{ AND } z)\,,\\ (x \texttt{ AND } y) \texttt{ OR } z &= (x \texttt{ OR } z) \texttt{ AND } (y \texttt{ OR } z)\,.\end{aligned} \tag{1.9}$$

Equivalently, they behave as set union and set intersection of the foundation elements, which can therefore be assigned arbitrary values. In other applications, $\vee$ and $\wedge$ might not be distributive, and the foundation assignments become restricted by the non-equality of combinations that would otherwise be identical. But their calculus would still be additive.

Multiplication

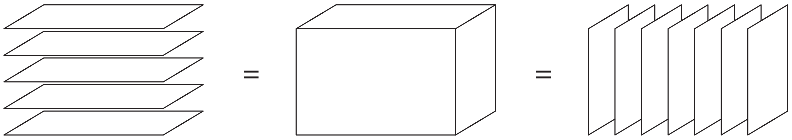
As well as by addition, measures can also combine by multiplication. Here, we consider a direct product of lattices. For example, one lattice might have playing-card foundation elements ($\spadesuit$, $\heartsuit$, $\clubsuit$, $\diamondsuit$) while the other has music-key foundations ($\flat$, $\natural$, $\sharp$). The direct-product lattice treats both together, here with 12 foundation elements like $\heartsuit \times \natural$ and 2^{12} elements overall:

$$\begin{array}{|c|c|c|c|}\hline \spadesuit\times\sharp & \heartsuit\times\sharp & \clubsuit\times\sharp & \diamondsuit\times\sharp \\ \hline \spadesuit\times\natural & \heartsuit\times\natural & \clubsuit\times\natural & \diamondsuit\times\natural \\ \hline \spadesuit\times\flat & \heartsuit\times\flat & \clubsuit\times\flat & \diamondsuit\times\flat \\ \hline \end{array} \;=\; \begin{array}{|c|c|c|c|}\hline \spadesuit & \heartsuit & \clubsuit & \diamondsuit \\ \hline \end{array} \;\times\; \begin{array}{|c|}\hline \sharp \\ \hline \natural \\ \hline \flat \\ \hline \end{array}$$

We now assume that the measure $m(x \times y)$ should depend only on $m(x)$ and $m(y)$. Write this relationship as a binary operation

$$m(x \times y) = m(x) \otimes m(y). \tag{1.10}$$

Now the direct-product operator is associative, $(x \times y) \times z = x \times (y \times z)$,



so

$$\big(m(x) \otimes m(y)\big) \otimes m(z) = m(x) \otimes \big(m(y) \otimes m(z)\big). \tag{1.11}$$

This has to hold for arbitrary values $m(x)$, $m(y)$, $m(z)$, and the associativity theorem then tells us that there must be some invertable function Φ of the measures m such that

$$\Phi(m(x \times y)) = \Phi(m(x)) + \Phi(m(y)). \tag{1.12}$$

We cannot now re-grade to $\Phi(m)$ and ignore m because we have already fixed the behaviour of m to be additive. What we can do is require consistency with that behaviour by requiring the sum rule (1.8) to hold for composite elements, $m(x \times t) + m(y \times t) = m((x \vee y) \times t)$ for any t. The *context theorem* (Knuth and Skilling, in preparation) then shows that Φ has to be logarithmic, so that

$$m(x \times y) = m(x)\, m(y). \tag{1.13}$$

While $\oplus$ is addition, $\otimes$ is multiplication. Combination is intrinsically multiplicative – because of associativity. There is no alternative.

Commutativity

Technically, we have not used the commutative property $x \vee y = y \vee x$ of a lattice. However, the sum rule automatically generates values that are equal, $v(x \vee y) = v(y \vee x)$. So real valuations cannot capture non-commutative behaviour. Quantum mechanics is an example, where states lack ordering so do not form a lattice, and the calculus is complex. Inference is not an example. There, *apple*-OR-*banana* is the same as *banana*-OR-*apple* so $\vee$ is commutative for us, and we are allowed to use the real values that we need.

1.2.3 Information

Different measures can legitimately be assigned to the same foundation elements, as when different individuals value apples, bananas and cherries differently. The difference between source measure μ and destination measure m can be quantified, consistently with lattice structure, as '*information*' $H(m \mid \mu)$.

One way of deriving the form of H is as a variational potential, in which destination m is obtained at the extremal (minimum, actually) of H, subject to whatever constraints require the change from μ. Suppose the playing-card example above has source measure μ, with destination m obtained by some constraint on card suits. Independently, the music-key example has source measure ν, with destination n obtained by some constraint on music keys.

Equivalently, we must be able to analyze the problems jointly. Measures multiply, so the joint element 'card suit i and music key j' has source measure $\mu_i\nu_j$ and destination $m_i n_j$. The latter is to be obtained at the extremal of $H(m_i n_j \mid \mu_i\nu_j)$, under one constraint acting on i and another on j. Temporarily suppressing the fixed source $\mu\nu$, the variational equation for the destination measure is

$$H'(m_i n_j) = \lambda_1(i) + \lambda_2(j), \tag{1.14}$$

where the λ's are the Lagrange multipliers of the i and j constraints. Writing $x = m_i$ and $y = n_j$, and differentiating $\partial^2/\partial x\partial y$, the right-hand side is annihilated, leaving

$$xyH'''(xy) + H''(xy) = 0, \tag{1.15}$$

whose solution is

$$H(z) = A - Bz + Cz\log z. \tag{1.16}$$

Setting $C = 1$ an as arbitrary scale (positive to ensure a minimum), $B = 1$ to place that minimum correctly at $m = \mu$, and $A = \mu$ to make the minimum zero, we reach (Skilling 1988)

$$\boxed{H(m \mid \mu) = \mu - m + m\log\tfrac{m}{\mu}} \tag{1.17}$$

This obeys (1.14), so the potential we seek exists, and is required to be of this unique form. The difference between measures plays a deep rôle in Bayesian analysis.

1.2.4 Probability

Acquiring data involves a reduction of possibilities. Some outcomes that might have happened, did not. In terms of the lattice of possibilities, the all-encompassing top element moves down. To deal with this, we seek a *bi*-valuation $p(x \mid t)$, in which the context t of model x can shrink. Within any fixed context, p is to be a measure, being non-negative and obeying the sum rule. But we want to change the numbers when the context changes.

To find the dependence on context, take ordered elements $x \le y \le z \le t$. As before, we require conformity with lattice ordering, here

$$x \le y \le z \quad\Longrightarrow\quad p(x \mid z) \le p(x \mid y) \tag{1.18}$$

so that a wider context dilutes the numerical value. Ordering such as $x \leq z$ can be carried out in two steps, $x \leq y$ and $y \leq z$. Our bi-valuation should conform to this, meaning that we require a "$\odot$" operator combining the two steps into one:

$$p(x \mid z) = p(x \mid y) \odot p(y \mid z). \tag{1.19}$$

Extending this to three steps and considering passage $p(x \mid t)$ from x to t, via y and z, gives another associativity relationship,

$$\big(p(x \mid y) \odot p(y \mid z)\big) \odot p(z \mid t) = p(x \mid y) \odot \big(p(y \mid z) \odot p(z \mid t)\big), \tag{1.20}$$

representing $(((x \leq y) \leq z) \leq t) = (x \leq (y \leq (z \leq t)))$. As before, this induces some invertable function Φ of our valuations p such that

$$\Phi\big(p(x \mid z)\big) = \Phi\big(p(x \mid y)\big) + \Phi\big(p(y \mid z)\big). \tag{1.21}$$

Again, we require consistency with the sum rule $p(x \vee y \mid t) = p(x \mid t) + p(y \mid t)$ for arbitrary context t. A variant of the context theorem (Knuth and Skilling, in preparation) then shows that Φ has to be logarithmic as before, so $\odot$ was multiplication. Specifically, we recognize p as *probability*, hereafter "pr",

$$\left.\begin{array}{rl} 0 = \mathrm{pr}(\emptyset) \leq \mathrm{pr}(x) \leq \mathrm{pr}(t) = 1 & \text{Range} \\ \mathrm{pr}(x \vee y) = \mathrm{pr}(x) + \mathrm{pr}(y) & \text{Sum rule for disjoint } x, y \\ \mathrm{pr}(x \wedge y) = \mathrm{pr}(x \mid y)\,\mathrm{pr}(y) & \text{Product rule} \end{array}\right\} \quad \| \; t \tag{1.22}$$

(The "$\| \; t$" notation means that all probabilities are conditional on t, and avoids proliferation of "$\mid t$" without introducing ambiguity.)

Just as measure theory was forced for valuations, so probability theory is forced for bi-valuations. We need not be distracted by claimed alternatives because they conflict with very general requirements. It is all very simple. There's only this one calculus for numerical bi-valuations on a lattice. If, say, we seek a calculus for conditional beliefs, then this has to be it. But the calculus itself is abstract and motive-free. We don't have to subscribe to an undefined idea like 'belief' in order to use it. In fact, the reverse holds. It is probability, with its defined properties, that would underpin belief, not the other way round.

Most simply of all, probability calculus can be subsumed in the single definition of probability as a ratio definition of measures:

$$\boxed{\mathrm{pr}(x \mid t) = \frac{m(x \wedge t)}{m(t)}} \tag{1.23}$$

This is the original discredited frequentist definition, as the ratio of number of successes to number of trials, now retrieved at an abstract level, which bypasses the catastrophic difficulties of literal frequentism when faced with isolated non-reproducible situations. The calculus of probability is no more than the calculus of proportions.

1.3 Inference

Henceforward, in accordance with traditional accounts, we take all foundation elements to be disjoint, and work in terms of these. The OR operator $\vee$ can be replaced by the summation to which it reduces, while the AND operator $\wedge$ can be written as the traditional comma. It is also usual to use I for context, and allow the discrete choice x to be continuous θ. The rules of probability calculus then reduce to

$$\left.\begin{array}{cr} \mathrm{pr}(\theta) \geq 0 & \text{Positivity} \\ \int \mathrm{pr}(\theta)\,\mathrm{d}\theta = 1 & \text{Sum rule} \\ \mathrm{pr}(\phi, \theta) = \mathrm{pr}(\phi \mid \theta)\,\mathrm{pr}(\theta) & \text{Product rule} \end{array}\right\} \quad \| \; I \tag{1.24}$$

1.3.1 Bayes' theorem

In inference, we need to consider both parameter(s) θ and data D, all in the overarching context I of all possibilities we are currently considering. By the product law, the joint probability of model and data factorizes:

$$\begin{array}{ccccc} \mathrm{pr}(\theta)\,\mathrm{pr}(D \mid \theta) & = \mathrm{pr}(\theta, D) = & \mathrm{pr}(D)\,\mathrm{pr}(\theta \mid D) & \| \; I \\ \text{Prior} \times \text{Likelihood} & = \quad \text{Joint} \quad = & \text{Evidence} \times \text{Posterior} & \\ \pi(\theta)\,\mathcal{L}(\theta) & = \quad \cdots\cdots \quad = & E\,\mathcal{P}(\theta) & \\ \text{Inputs} & \Longrightarrow & \text{Outputs} & \end{array} \tag{1.25}$$

On the left lies the prior probability $\pi(\theta) = \mathrm{pr}(\theta \mid I)$, representing how we originally distributed the parameters' unit mass of probability. This assignment has provoked legendary argumentation, and we discuss it below. Also on the left is the likelihood $\mathcal{L}(\theta) = \mathrm{pr}(D \mid \theta)$, representing the probability distribution of the data for each allowed input θ. This is less controversial. The instrument acquiring the data can usually be calibrated with known inputs θ to find how often it produces specific outputs D, which effectively fixes the likelihood to any desired precision. If there remain any unknown calibration parameters in the likelihood, they can be incorporated in θ as extra parameters to be determined, leading to extra computation but no difficulty of principle.

On the far right is the posterior $\mathcal{P}(\theta) = \mathrm{pr}(\theta \mid D, I)$, representing our inferred distribution of probability among the models, after using the data. The difference between prior and posterior is the information (1.17)

$$H(\mathcal{P} \mid \pi) = \int \mathcal{P}(\theta) \log\big(\mathcal{P}(\theta)/\pi(\theta)\big)\,\mathrm{d}\theta \tag{1.26}$$

gleaned about θ. Also on the right is the evidence $E = \mathrm{pr}(D \mid I)$, representing how well our original assignments managed to predict the data. E is also known as 'prior predictive' (how it is often used), 'marginal likelihood' (how it is often