NESTED SAMPLING FOR GENERAL BAYESIAN COMPUTATION

John Skilling
Maximum Entropy Data Consultants Ltd.
Killaha East, Kenmare, County Kerry, Ireland
skilling@eircom.net — October 2004, revised February, July, August 2005

Abstract: Nested sampling estimates directly how the likelihood function relates to prior mass. The evidence (alternatively the marginal likelihood, marginal density of the data, or the prior predictive) is immediately obtained by summation. It is the prime result of the computation, and is accompanied by an estimate of numerical uncertainty. Samples from the posterior distribution are an optional by-product, and are obtainable for any temperature. The method relies on sampling within a hard constraint on likelihood value, as opposed to the softened likelihood of annealing methods. Progress depends only on the shape of the "nested" contours of likelihood, and not on the likelihood values. This invariance (over monotonic re-labelling) allows the method to deal with a class of phase-change problems which effectively defeat thermal annealing.

Keywords: Bayesian computation, evidence, marginal likelihood, algorithm, nest, annealing, phase change, model selection.

INTRODUCTION

Our primary task is to evaluate

$$Z = \text{evidence} = \int L\, dX \tag{1}$$

where $L = L(\theta)$ is the likelihood function, $dX = \pi(\theta)d\theta$ is the element of prior mass, and θ represents the unknown parameter(s). The probabilistic context of this is usually in the form of Bayes' theorem, being the product law under background model assumptions $\mathcal{I}$:

$$\begin{array}{rcl} \Pr(D \mid \theta,\mathcal{I}) \times \Pr(\theta \mid \mathcal{I}) & = & \Pr(D \mid \mathcal{I}) \times \Pr(\theta \mid D,\mathcal{I}) \\ \text{Likelihood} \times \text{ Prior} & = & \text{Evidence} \times \text{Posterior} \\ L(\theta) \ \times \ \pi(\theta)d\theta & = & Z \ \times \ p(\theta)d\theta \end{array} \tag{2}$$

Here D are the acquired data which let us modulate our prior belief $dX = \pi(\theta)d\theta$ into posterior $dP = p(\theta)d\theta$. Prior and posterior, as always, are normalised to unit total. Inputs L and π yield outputs Z and p.

Calculating the value of Z allows different model assumptions to be compared through the ratios of evidence values known as Bayes factors. Presenting Z thus lets the results be future-proof, in that future models can be compared with the current one, without having to re-do the current calculation. Giving the value of Z is a courtesy to other workers who may wish to perform model selection, and ought to be a standard part of rational enquiry. It is one half of the output from a Bayesian calculation, the other half being the posterior. Oddly,

the quantity currently lacks a universal name, and the author favours the crisp single word "evidence", which is growing in popularity (MacKay 2003) and for which no other standard technical definition has yet been agreed. "Marginal likelihood" describes how it is (usually but not always) constructed. "Prior predictive" describes how it is (usually but not always) used. "Evidence" denotes what it is.

Historically, following Metropolis *et al.* (1953) and Hastings (1970), algorithms such as Markov chain Monte Carlo (MCMC) have been principally designed for the posterior. Indeed, standard MCMC yields only a set of samples representing the normalised posterior, and fails to give the evidence at all. Obtaining the evidence has required considerable extra work, usually involving a sequence of intermediate distributions that bridge between prior and posterior, as in thermodynamic integration reviewed with generalisations by Gelman & Meng (1998). This is unfortunate, because the extra computational difficulty and the lack of standard terminology suggest that the evidence value is an optional by-product, rather than a quantity of central importance. Nested sampling reverses the historical approach. The evidence is now the prime target, with representative posterior samples available as the optional by-product.

The paper starts with the idea of sorting points θ by their likelihood values, which are then summed to give the evidence. Of course, there are usually far too many points to do this explicitly, so nested sampling simulates the operation statistically. The evidence then becomes accompanied by a corresponding numerical uncertainty. A methodological section argues that nested sampling is Bayesian in nature. With the basic method in place, it is possible to estimate the density of states, to obtain samples from the posterior, and to quantify arbitrary properties of θ. These sections complete the formal development. Nested sampling is then compared with the conventional approach of annealing, and is shown by examples and limiting cases to be wider in scope. The paper concludes with an overview, and an Appendix with a simple 'C' program.

SORTING

The evaluation of $\int L\,dX$ looks like a straightforward problem of numerical analysis. Simplistically, one might raster over underlying coordinates θ to evaluate $\int L(\theta)\pi(\theta)d\theta$. However, this rapidly becomes impractical as soon as θ has more than a very few dimensions. Instead, we will use the prior X directly.

Prior mass X can be accumulated from its elements dX in any order, so define

$$X(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta)d\theta \tag{3}$$

as the cumulant prior mass covering all likelihood values greater than λ. As λ increases, the enclosed mass X decreases from 1 to 0. Writing the inverse function as $L(X)$, *i.e.* $L(X(\lambda)) \equiv \lambda$, the evidence becomes a one-dimensional integral over unit range

$$Z = \int_0^1 L(X)\,dX \tag{4}$$

in which the integrand is positive and decreasing (Figure 1), so it has to be well behaved. Accomplishing this transformation from θ to X involves dividing the unit prior mass into tiny elements, and sorting them by likelihood.

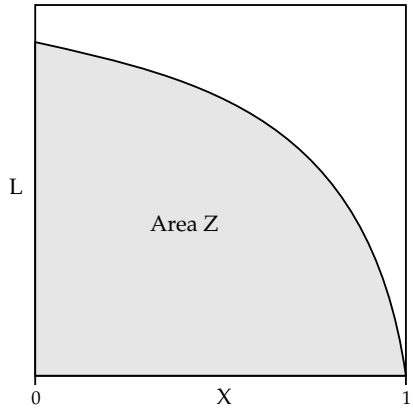

Figure 1: Likelihood function with area Z.

A very simple example, on a 4×4 grid of two-dimensional θ, is shown in the table (5) of likelihood values ascribed to its 16 cells of equal prior mass $\frac{1}{16}$.

$$L \;=\; \begin{array}{|r|r|r|r|} \hline 0 & 8 & 15 & 3 \\ \hline 11 & 24 & 22 & 10 \\ \hline 19 & 30 & 26 & 16 \\ \hline 9 & 23 & 18 & 6 \\ \hline \end{array} \tag{5}$$

Our plan is to proceed as if we could sort these elements by likelihood, in the above example to $L = (30,26,24,23,22,19,18,16,15,11,10,9,8,6,3,0)$, whence $Z = \frac{30}{16}+\frac{26}{16}+\frac{24}{16}+\frac{23}{16}+\frac{22}{16}+\frac{19}{16}+\frac{18}{16}+\frac{16}{16}+\frac{15}{16}+\frac{11}{16}+\frac{10}{16}+\frac{9}{16}+\frac{8}{16}+\frac{6}{16}+\frac{3}{16}+\frac{0}{16} = 15$, to be evaluated right-to-left into domains of progressively greater likelihood. The likelihood corresponding to (say) $X = \frac{1}{5}$, being one fifth of the way along the sequence so falling into the fourth cell out of sixteen, is $L(X=0.2) = 23$.

As a technicality we may need to resolve ties between points of equal L. A point k, which has coordinates θ_k and corresponding likelihood $L_k = L(\theta_k)$, can also be assigned a label ℓ_k, chosen from some library large enough that repeats are not expected. Random samples from `Uniform`(0,1) suffice, as would a cryptographic identification key derived from θ, or almost anything else. Labels parameterise within each likelihood contour, and extend the likelihood to

$$L_k^+ = L_k + \epsilon \ell_k \tag{6}$$

where ϵ is some tiny coefficient that never affects numerical likelihood values (which are always held to finite precision), but nevertheless enables an unambiguous ranking of the points, even where raw likelihoods are equal. For clarity, we mostly ignore this refinement hereafter.

INTEGRATION

Coordinate-dependent complications of geometry, topology, even dimensionality, are all annihilated by the sorting operation, and the remaining task is easy and well understood. Suppose that we knew how to evaluate the likelihood as $L_i = L(X_i)$ at a right-to-left sequence of m points

$$0 < X_m < \cdots < X_2 < X_1 < 1 \tag{7}$$

Any convenient numerical recipe would then estimate Z as a weighted sum

$$Z \longleftarrow \sum_{i=1}^{m} L_i w_i \tag{8}$$

of these values, in which the area in Figure 1 is approximated as a set of columns of height L and width $w \sim \Delta X$.

Because $L(X)$ is non-increasing, it is bounded below by any value evaluated at larger X. Hence $w_i = X_i - X_{i+1}$ with $X_{m+1} = 0$ gives a lower bound

$$Z = \int_0^1 L\, dX \;\geq\; \sum_{i=1}^{m} L_i (X_i - X_{i+1}) \tag{9}$$

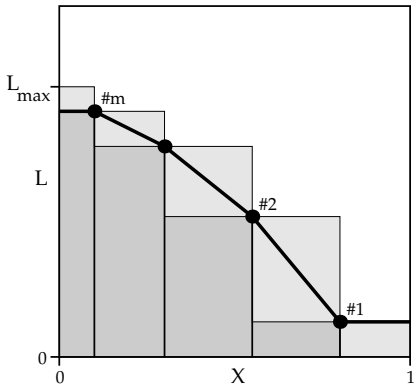

Figure 2: Lower bound (dark shading) and upper bound (all shading) on area. The thick line indicates the trapezoidal rule.

There is a similar upper bound (Figure 2) from $w_i = X_{i-1} - X_i$ with $X_0 = 1$,

$$Z = \int_0^1 L\, dX \;\leq\; \sum_{i=1}^{m} L_i (X_{i-1} - X_i) + L_{\max} X_m \tag{10}$$

where $L_{\max}$ is the maximum likelihood value to be found as $X \to 0$. Technically, $L_{\max}$ is not determined by nested (or any other) sampling. There could always remain some tiny volume containing huge and dominant likelihood values, unless that can be ruled out by some global analysis (as when a Gaussian

likelihood factor cannot exceed $1/\sqrt{2\pi}\sigma$). However, when judging that a run can be terminated, we implicitly assert that any increase in L beyond the highest value yet found is not consequential. With this proviso, the upper limit (10) holds, and errors from numerical integration are at most $\mathcal{O}(N^{-1})$, that being the difference between upper and lower bounds.

The trapezoidal rule $w_i = \frac{1}{2}(X_{i-1} - X_{i+1})$ with reflecting boundary conditions $X_0 = 2 - X_1$ and $X_{m+1} = -X_m$ to avoid awkward behaviour, reduces this to $\mathcal{O}(N^{-2})$ in most cases. The integrand is already quite well behaved, so we do not expect useful improvement over the trapezoid rule (or similar) denoted by the "$\longleftarrow$" arrow.

EVIDENCE

The integral for Z is dominated by wherever the bulk of the posterior mass is to be found. Typically, this occupies a small fraction e^{-H} of the prior, where

$$H = \text{information} = \int \log(dP/dX)\, dP \tag{11}$$

H is (minus) the logarithm of that fraction of prior mass that contains the bulk of the posterior mass, and it may well be of the order of thousands or more in practical problems where the likelihood is concentrated in some obscure corner of the prior domain.

To illuminate the width over X of the posterior, suppose the likelihood function has C approximately-Gaussian principal components, so that L is approximately a rank-C multivariate normal. For this model case, the likelihood can be written $L \propto \exp(-\frac{1}{2}r^2)$, where r is radius in C dimensions, with enclosed prior mass $X \propto r^C$. This posterior $dP \propto L\, dX = LX\, d\log X$ induces a standard deviation $\langle(\delta \log X)^2\rangle^{1/2}$ of $\sqrt{C/2}$. So we expect the posterior mass to be fairly broadly distributed over a range something like $-H \pm \sqrt{C}$ in $\log X$. Generally, each useful principal component of the likelihood significantly restricts the range originally permitted by the prior (otherwise it's not useful), so H should usually exceed C, let alone $\sqrt{C}$, suggesting that locating and reaching the posterior domain is a more difficult task than navigating within it. This qualitative behaviour where the posterior mass is mostly around $\log X \approx -\text{Huge} \pm \text{big}$ (Huge meaning H) is widely seen in practical applications.

To cover such a range, sampling ought to be linear in $\log X$ rather than in X, and we set

$$X_1 = t_1,\ X_2 = t_1 t_2,\ \cdots\cdots,\ \ X_i = t_1 t_2 \ldots t_i,\ \ \cdots\cdots,\ X_m = t_1 t_2 t_3 \ldots t_m \tag{12}$$

where each t_i lies between 0 and 1. It is these ratios $\mathbf{t}$ that control the subsequent calculations. If, for example, we could set $t = 0.99$ each time, then we should reach the bulk of the posterior after something like $100H$ steps, and cross it in a further $100\sqrt{C}$ steps. Any such sequence $\mathbf{t}$ leads to an estimate of Z, which we can make explicit by writing

$$Z(\mathbf{t}) \ \longleftarrow\ \sum_{i=1}^{m} L_i w_i(\mathbf{t}) \tag{13}$$

NESTED SAMPLING IDEA

Although we cannot usually set precise values of t, it turns out that we can often set them statistically, and that is enough. All we need do, at step i, is take a random new point X_i from the prior, subject to $X_i < X_{i-1}$ (starting with $X_0 = 1$). Our knowledge of the new point $X_i = t_i X_{i-1}$ would be specified by $t_i \in \texttt{Uniform}(0, 1)$.

In principle, such a point could be obtained by sampling X_i uniformly from within the corresponding restricted range $(0, X_{i-1})$, then interrogating the original likelihood-sorting to discover what its θ_i would have been. In practice, it would naturally be obtained directly as θ_i, by sampling within the equivalent constraint $L(\theta) > L_{i-1}$ (with $L_0 = 0$ to ensure complete initial coverage), in proportion to the prior density $\pi(\theta)$. This too finds a random point, distributed just the same. The second method is equivalent to the first, but bypasses the use of X. **So we don't need to do the sorting after all!** That's the key.

Successive points are illustrated in Figure 3, in which prior mass is represented by area. Thus, point 2 is found by sampling over the prior within the box defined by $L > L_1$, and so on. Such points will usually be found by some MCMC approximation, starting at some point θ^* known to obey the constraint (if available), or at worst starting at θ_{i-1} which lies on and defines the current likelihood boundary.

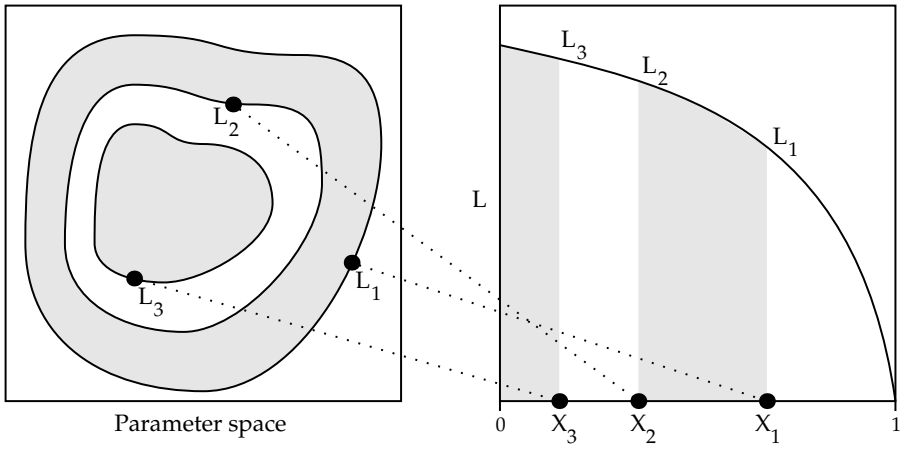

Figure 3: Nested likelihood contours are sorted to enclosed prior mass X.

It is not the purpose of this introductory paper to develop the technology of navigation within such a volume. We merely note that exploring a hard-edged likelihood-constrained domain should prove to be neither more nor less demanding than exploring a likelihood-weighted space. For example, consider a uniform prior weighted by a C-dimensional unit Gaussian likelihood $L(\theta) = \exp(-\frac{1}{2}|\theta|^2)$. Conventional Metropolis-Hastings exploration is simply accomplished with trial moves of arbitrary direction having step-length $|\delta\theta|$ around 1 for efficiency. Most points have $|\theta| \approx \sqrt{C}$, so the relaxation time is about C steps.

In nested sampling, the corresponding hard constraint is the ball $|\theta| < \sqrt{C}$ (or thereabouts). The typical point has $|\theta| \approx \sqrt{C} - 1/\sqrt{C}$, that being the median radius of the ball. Again, the efficient trial step-length is $|\delta\theta| \approx 1$, so the relaxation time per iterate is much the same as before. There are well-developed methods, such as Hamiltonian (or "hybrid") Monte Carlo (Duane *et al.* 1987, Neal 1993), slice sampling (Neal 2003) and more, which learn about more general shapes of L in order to explore the likelihood-weighted space more efficiently. Similar methods ought to work for exploring likelihood-constrained domains, but have not yet been developed.

In terms of prior mass, successive intervals w scan the prior range from $X = 1$ down to $X = 0$. In terms of coordinates θ, the intervals represent nested shells around contours of constant likelihood value, with points exactly on the same contour being ranked by their labels ℓ. More generally, instead of taking 1 point within the likelihood-constrained box, take N of them where N is any convenient number, and select the worst (lowest L, highest X), as the i'th point. This recurrence is

$$X_0 = 1, \qquad X_i = t_i X_{i-1}, \qquad \Pr(t_i) = N t_i^{N-1} \text{ in } (0,1) \tag{14}$$

t_i being the largest of N random numbers from `Uniform`(0,1). The mean and standard deviation of $\log t$ (which dominates the geometrical exploration) are

$$\mathtt{E}(\log t) = -1/N, \qquad \mathtt{dev}(\log t) = 1/N \tag{15}$$

The individual $\log t$ are independent, so after i steps, the prior mass is expected to shrink to $\log X_i \approx -(i \pm \sqrt{i})/N$. Thus we expect the procedure to take about $NH \pm \sqrt{NH}$ steps to shrink down to the bulk of the posterior, and a further $N\sqrt{C}$ or so steps to cross it. For a crude implementation, we can simply proclaim $\log X_i = -i/N$ as if we knew it, though it's more professional to acknowledge the uncertainties.

Actually, it is not necessary to find N points anew at each step, because $N-1$ of them are already available, being the survivors after deleting the worst. Only one new point is required per step, and this θ may be found by any method that draws from the prior subject to $L(\theta)$ being above its constraint L_{i-1}. One method is to replace the deleted point by a copy of a random survivor, evolved within the box by MCMC for some adequate number of trials. Surviving points could be used as stationary guides in such exploration. Another method might be generation of new points by genetic mixing of the survivors' coordinates. All that matters is that the step ends with N usably independent samples within the box.

NESTED SAMPLING PROCEDURE

At each step, the procedure has N points $\theta_1, \ldots, \theta_N$ with corresponding likelihoods $L(\theta_1), \ldots, L(\theta_N)$, augmented to L^+ as in (6) if ties of likelihood are anticipated. The lowest (minimum) such value is the likelihood L_i associated with step i. There are to be j iterative steps.

Start with N points $\theta_1, \ldots, \theta_N$ from prior;
initialise $Z = 0$, $X_0 = 1$.
Repeat for $i = 1, 2, \ldots, j$;
record the lowest of the current likelihood values as L_i,
set $X_i = \exp(-i/N)$ (crude) or sample it to get uncertainty,
set $w_i = X_{i-1} - X_i$ (simple) or $(X_{i-1} - X_{i+1})/2$ (trapezoidal),
increment Z by $L_i w_i$,
then replace point of lowest likelihood by new one drawn from within $L(\theta) > L_i$, in proportion to the prior $\pi(\theta)$.
Increment Z by $N^{-1}(L(\theta_1) + \ldots + L(\theta_N))\, X_j$.

The last step fills in the missing band $0 < X < X_j$ of the desired integral $\int_0^1 L\, dX$ with weight $w = N^{-1} X_j$ for each surviving point, after the iterative steps have compressed the domain to X_j. So the final number of terms in the evidence summation (8) is $m = j + N$. However, there should already have been sufficient steps j to accumulate most of the integral, so this final increment ought to be an unimportant refinement. It is also possible to accumulate the information H along with Z.

Figure 4 illustrates the method, running with $N = 3$ points. Initially, three samples are taken from the unconstrained prior, whose mass is represented by the complete square area on the left. These points could equivalently have been taken randomly from X in (0,1), as shown on the lower line on the right. They have labels 1, 3, 4, as yet unknown. In step 1, the worst (lowest L, highest X) of these points is identified as number 1, with likelihood L_1. It is then replaced by a new point, drawn from inside the contour $L(\theta) > L_1$. Equivalently, it could have been taken randomly from X in $(0, X_1)$. Including the two survivors, there are still three samples, now all uniform in the reduced range $(0, X_1)$. With the particular random numbers used for Figure 4, the new point in step 1 happened to lie outside the two survivors, so became number 2, but in step 2 the new point happened to be the innermost, and was eventually identified as number 5. After the $j = 5$ allotted steps, the five discarded points 1,2,3,4,5 are augmented with the final three survivors 6,7,8 to give the $m = 8$ points $(X_1, \ldots, X_8)$ shown on the top line. It is over these points that the sum $\sum_{i=1}^{8} L_i w_i$ is evaluated to estimate Z.

With $N = 3$ samples, shrinkage is expected to be roughly geometrical, by $\Delta \log X \sim -1/3$ per step. The diagram on the left of Figure 4 shows likelihood contours drawn at levels corresponding to enclosed areas diminishing by this

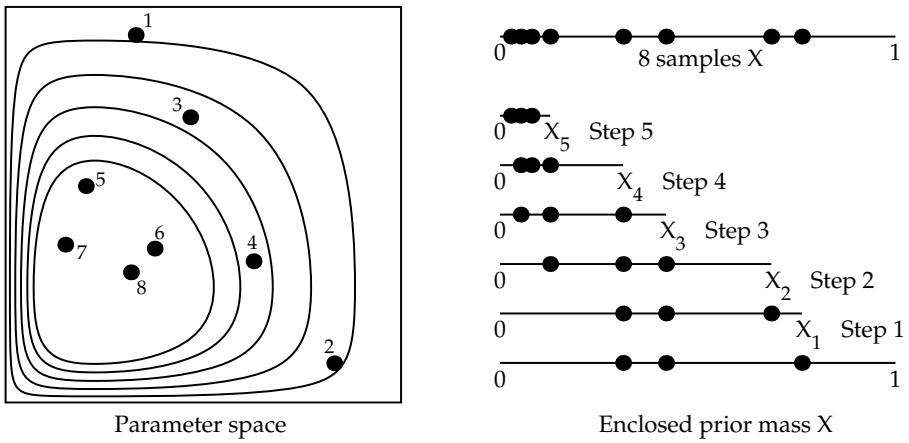

Figure 4: Nested sampling for five steps with a collection of 3 points. Likelihood contours shrink by factors $\exp(-1/3)$ in area and are roughly followed by successive sample points.

factor — *i.e.* the i'th contour encloses prior mass $e^{-i/3}$. Indeed, the first point lies close to the first contour, the second point is not too far outside the second contour, and so on until the fifth point chances to fall inside the fifth contour. If we could arrange exact matching, we would know the X's and have a definitive answer for Z, depending only on the scheme of numerical integration. Since we can't arrange this, we proceed towards deriving a probabilistic estimate instead. Although it is unusual to deal with an integral in which the ordinate L is known and the abscissa X is uncertain, rather than the other way round, the problem remains soluble.

NESTED SAMPLING TERMINATION

Termination of the main loop could simply be after a pre-set number of steps, or could be when even the largest current likelihood, taken over the full current box, would not increase the current evidence by more than some small fraction f;

$$\max(L_1, \ldots, L_N) X_j < f Z_j \quad \Longrightarrow \quad \text{termination.} \tag{16}$$

Plausibly, the accumulation of Z is then tailing off, so the sum is nearly complete.

The usual behaviour of the areas $L_i w_i$ is that they start by rising, with the likelihood L_i increasing faster than the widths w_i decrease. The more important regions are being found. At some point, L flattens off sufficiently that decreasing width dominates increasing likelihood, so that the areas pass across a maximum and start to fall away. Most of the total area is usually found in the region of this maximum, which occurs in the region of $X \approx e^{-H}$. There are counter-examples, but that behaviour is often expected. Remembering $X_i \approx e^{-i/N}$, this

suggests an alternative termination condition

"continue iterating until the count i significantly exceeds NH" (17)

which still expresses the general aim that a nested-sampling calculation should be continued until most of Z has been found. (Of course, H is here the current evaluate from the previous i iterates.) This is the criterion used in the Appendix program.

Unfortunately, we can offer no rigorous criterion based on sampling to ensure the validity of any such termination condition. It is perfectly possible for the accumulation of Z to flatten off, apparently approaching a final value, whilst yet further inward there lurks a small domain in which the likelihood is sufficiently large to dominate the eventual results. Termination remains a matter of user judgment about the problem in hand, albeit with the aim of effectively completing the accumulation of Z. If in doubt, continue upward and inward.

However, if an analytical upper bound $L \leq L_{\max}$ can be found, such as when a Gaussian likelihood factor cannot exceed $1/\sqrt{2\pi}\sigma$, it can be used in (16) to give a firmer termination criterion

$$L_{\max} X_j < f Z_j \quad \Longrightarrow \quad \text{termination}. \tag{18}$$

In this case, all but a fraction f of Z has been found.

NUMERICAL UNCERTAINTY

It is possible to run nested sampling crudely, by assigning each $\log t$ its mean value of $-1/N$, and ignoring its uncertainty. With X_i thereby being set to $e^{-i/N}$, this captures the basic idea by giving a quick picture of the likelihood function $L(X)$. An early example of a similar approach is McDonald & Singer (1967), and it is encoded in the simple Appendix program. However, a fuller and better treatment is also possible.

For a given choice of coefficients $\mathbf{t}$, the estimate of Z would be $\sum_i L_i w_i(\mathbf{t})$ from (13). One such choice of $\mathbf{t}$ will be correct, corresponding to the selected points θ_i, but we do not know which, and it could have been any. In particular, the correct choice of $\log t_i$ is most unlikely to be $-1/N$ for all i. Instead, the "sequence probability" $\Pr(\mathbf{t})d\mathbf{t} = \prod_i N t_i^{N-1} dt_i$ from (14) induces a distribution for our estimates of Z;

$$\Pr(Z) \;\longleftarrow\; \int \delta\big(Z - \sum_{i=1}^{m} L_i w_i(\mathbf{t})\big)\, \Pr(\mathbf{t})\, d\mathbf{t} \tag{19}$$

Analytic expressions for moments $\langle Z \rangle, \langle Z^2 \rangle, \ldots$ are available through means, correlations, and higher moments of t_1, $t_1 t_2$, $t_1 t_2 t_3, \ldots$. However, the dominant uncertainty in Z is usually due to the Poisson variability $NH \pm \sqrt{NH}$ in the number of steps to reach the posterior, implying a geometrical uncertainty factor $\exp(\pm\sqrt{H/N})$ which could be many powers of e. Hence it is $\log Z$ rather than Z

itself that should usually have a fairly symmetrical roughly-normal uncertainty. The distribution of Z may be awkwardly skewed, with its moments dominated by occasional upward outliers in $\log Z$. A reader who doubts this may consider the hypothetical normally-distributed $\log Z = 100 \pm 10$, whose direct mean and standard deviation $Z = e^{150} \pm e^{200}$ are unhelpful, especially since it is known that Z has to be positive. So, rather than proceeding with moment expansions, it seems better to use Monte Carlo, taking a set $\{\mathbf{t}\}$ of several dozen samples from the sequence probability $\Pr(\mathbf{t})$ to simulate the X's and thence obtain the distribution of Z from samples $\{Z\}_{\mathbf{t}}$, from which the statistics of $\log Z$ can be read off with adequate confidence:

$$\log Z \;\longleftarrow\; \text{estimate} \pm \text{uncertainty, from } \{\log Z\}_{\mathbf{t}} \tag{20}$$

The uncertainty accompanying such estimates will usually diminish as the inverse square root of N, the amount of computation that one is prepared to invest in the original exploration. The same approach can be applied to any average over $\mathbf{t}$:

$$\int \cdots \Pr(\mathbf{t})\, d\mathbf{t} \;=\; \big\langle \cdots \big\rangle_{\mathbf{t}} \tag{21}$$

Alongside the uncertainty of about $\pm\sqrt{H/N}$ in $\log Z$, there is also the systematic numerical bias imposed by the integration rule. Thus, the lower- or upper-bound numerical weights (9) or (10) will usually give $\mathcal{O}(1/N)$ bias, which the trapezoidal improvement reduces to $\mathcal{O}(1/N^2)$. Usually, these biasses are overwhelmed by the uncertainty.

Technically, $\Pr(Z)$ is most easily evaluated by Monte Carlo sampling over $\mathbf{t}$. That is highly unlikely to be a significant source of error if several dozen samples are used (which is allowable because no likelihood evaluations are involved). However, if this is thought to be of concern, it is worth noting that Monte Carlo could be evaded by evaluating the integral differently. Write (13) (with lower-bound weights (9) for simplicity) as

$$Z(\mathbf{t}) \;\longleftarrow\; \sum_{i=1}^{m} \lambda_i X_i(\mathbf{t})\,, \quad \lambda_i = L_i - L_{i-1} \tag{22}$$

with $L_0 = 0$, and expand X_i using (12):

$$Z(\mathbf{t}) \;\longleftarrow\; t_1\big(\lambda_1 + t_2\big(\lambda_2 + \cdots + t_{m-1}(\lambda_{m-1} + t_m(\lambda_m))\cdots\big)\big) \tag{23}$$

Working outwards, use the recurrence relation

$$Z_m = \lambda_m, \quad Z_{i-1} = \lambda_{i-1} + t_i Z_i \tag{24}$$

to reach the required $Z = Z_1$. Each step of this takes a computed distribution $\Pr(Z_i)$ and integrates it with $\Pr(t)$ from (14) (a convolution over $\log X$) before shifting it by λ_{i-1} to reach $\Pr(Z_{i-1})$, ending with the required $\Pr(Z)$, from which statistics of Z or $\log Z$ can be read off. In this way, the m-dimensional integral (19) can be reduced to m feasible one-dimensional operations without any methodological qualms.

SIMPLE EXAMPLE

In C dimensions, consider the simple Gaussian problem

$$L(\theta) = \exp\Big(-\frac{r^2}{2\sigma^2}\Big), \qquad r^2 = \sum_{i=1}^{C} \theta_i^2 \tag{25}$$

where θ has flat prior within the unit sphere

$$\text{prior } \pi(\theta) = \frac{(C/2)!}{\pi^{C/2}} \quad (\pi = 3.1415...) \qquad \text{in } r < 1 \tag{26}$$

For convenience, take $\sigma \ll C^{-1/2}$, so that almost all the likelihood is well within the prior domain. The evidence (discarding the tails outside the domain) evaluates to

$$Z = (C/2)!\,(2\sigma^2)^{C/2} \tag{27}$$

We observe that L is a decreasing function of radius r, so that sorting (were we able to perform it) would organise θ into an outward radial sequence of nested shells, having enclosed prior mass

$$X = r^C \tag{28}$$

Hence the likelihood function is

$$L(X) = \exp(-X^{2/C}/2\sigma^2) \tag{29}$$

as plotted in Figure 5 for $C = 10$, $\sigma = 0.01$.

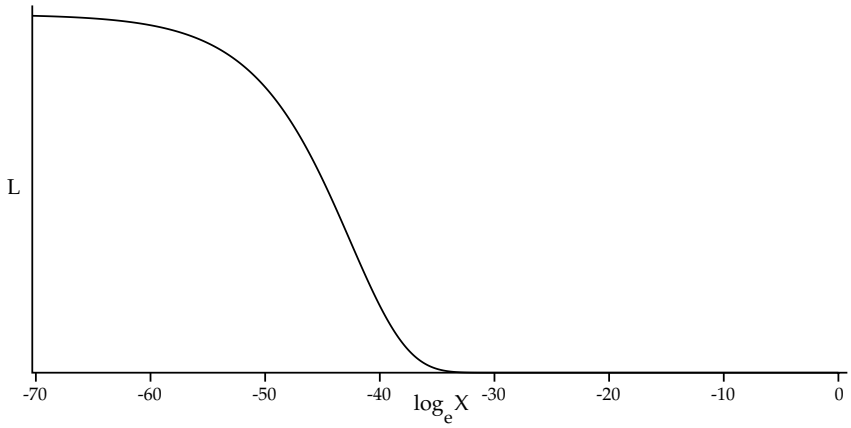

Figure 5: Likelihood function of simple example (on logarithmic abscissa).

The bulk of this mass is around radius $r \sim \sigma\sqrt{C}$, and the information required to reach down into this small patch is

$$H \approx -C \log(\sigma\sqrt{C}) \tag{30}$$

Meanwhile, the true numerical values are $\log Z = -37.81$, $H = 32.80$. (Although Figure 5 is a little misleading visually because of the logarithmic scale, the bulk of the posterior is indeed around $\log X \approx -H \approx -33$.)

It is our task to reproduce this likelihood function, and thence the evidence as its integral, by nested sampling. For simplicity, consider using just one sample $N = 1$.

The initial step is to take a random point θ_1 within the unit sphere, and evaluate its likelihood $L_1 = L(\theta_1)$. The next step is to take a random point θ_2 of greater likelihood, $L(\theta_2) > L_1$. Using our analytical insight into this particular problem, we know that we can obtain such a point by sampling within the sphere $r < r_1$. We have special knowledge, and can use it. However, a nested sampling program would generally have to find θ_2 by some other method. It might use MCMC over the prior, starting at θ_1 and accepting only those points with greater likelihood, until correlation with the original point was deemed to be lost. Or it might use something else. We do not discuss such choices here: we simply assume that sampling from within a likelihood constraint is possible.

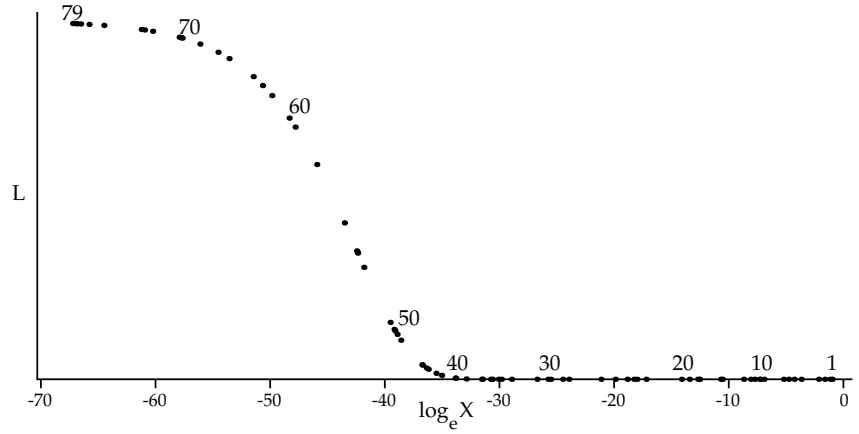

Figure 6: A random sequence of the first 79 nested sampling points, shown at their true X values. With $N = 1$, the i'th point lies at $\log X = -i$ on average, but particular sequences drift randomly away.

The next step is to take a third random point θ_3 of yet greater likelihood, $L_3 = L(\theta_3) > L_2$, and so on indefinitely. If we could use our analytical insight (28), we would evaluate the enclosed prior masses X and hence determine the sequence $(X_1, L_1), (X_2, L_2), (X_3, L_3), \ldots$ properly, as shown in Figure 6 for one particular run. By construction, these points lie exactly on the true likelihood curve. Although the points are quite coarsely spaced so that $\mathcal{O}(1)$ errors are expected, numerical integration (by trapezoid, say) gives a respectable estimate $\log Z \approx -37.60$. This example used just one point ($N = 1$), but for more accuracy we might evolve a collection of $N > 1$ points to obtain a sequence N times more closely spaced and N^2 more accurate. But, without the analytical insight available in this simple example, we would not have exact locations X, so we could not use this scheme.

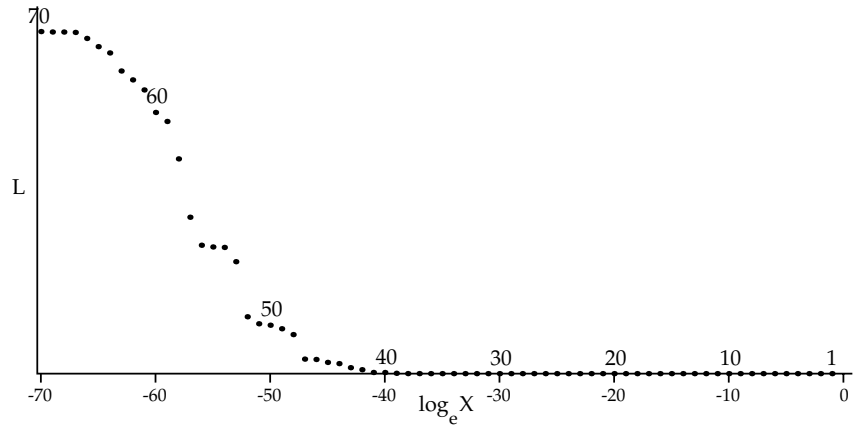

Figure 7: The same sequence of nested sampling points as in Fig. 6, shown at the crude central estimate of their X values (*i.e.* uniformly in $\log X$).

However, we do have crude central estimates $\log X_i = -i/N$ for the locations, and could use them as in Figure 7. The general profile of the likelihood function is preserved, in that likelihood values L and their ordering are the same. However, the $\log X$ abscissa has been distorted to keep the points uniformly spaced in keeping with the central estimates. By the time the bulk of the posterior has been reached, the accumulated Poisson errors have shifted the peak away from $i \approx H$ by $\delta i \sim \pm\sqrt{i}$, so that $\log X$ has been offset, leading to a corresponding (and dominating) uncertainty of $\pm\sqrt{H}$ in $\log Z$ (or $\pm\sqrt{H/N}$ for general N). Numerically, this particular result was $\log Z = -43.6 \pm 5.9$, being these crude estimates of the mean and standard deviation.

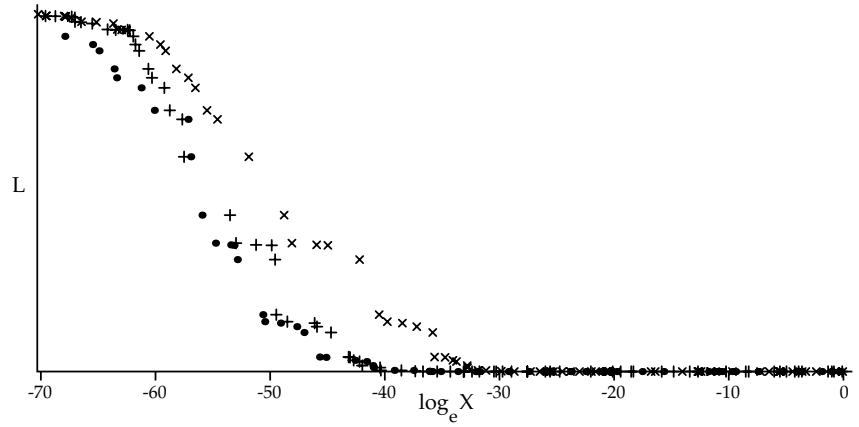

Figure 8: The same sequence of nested sampling points as in Fig. 6, showing three random assignments $(\bullet, +, \times)$ for X values taken from $\Pr(X)$.

A more professional but quite similar estimate $\log Z = -42.7 \pm 5.5$ (again mean and standard deviation) is obtained by sampling (*i.e.* repeatedly guessing)

the mass ratios $\mathbf{t}$ from their pdf (14), and accumulating the values of Z that are thereby produced. Three of these guesses are shown in Figure 8. As expected the curves are mostly the same shape and differ principally by random offset in $\log X$. Averaged over many assignments of $\mathbf{t}$, this better estimate misses the true value $\log Z = -37.81$ by an unremarkable 0.9 standard deviations.

It is, of course, necessary to take enough iterates (at least NH) to reach and then cross the bulk of the posterior mass. Likelihood values can guide the termination decision by indicating when the accumulation of Z appears to be tailing off, perhaps as in (17), but there is seldom a definitive criterion.

Incidentally, the discerning reader will have noticed that the original dimensionality of the example disappeared early on. The problem could just as well have been one-dimensional, described by (29). Nested sampling ignores dimensionality, off-loading such complications to the task of sampling within the likelihood constraint.

MULTIPLE RUNS

In repeated trials of nested sampling, the 50% interquartile range for Z is observed to cover the truth about half of the time, and similarly for other confidence intervals. Logically, though, that frequentist observation is pointless because the test is forced to succeed. It could only fail if there were an error, either in the logic of nested sampling or in the programming. Nevertheless, it raises the question of how multiple runs should best be combined. Loosely, one might invent some *ad hoc* averaging, but there ought to be a better way.

Suppose that several runs $r = 1, 2, \ldots$ are undertaken with numbers $N^{(r)}$ of points, and that run r accumulates likelihood values $L_1^{(r)} < L_2^{(r)} < L_3^{(r)} < \ldots$. Merge these values into a single global sequence

$$\mathcal{L}_1 < \mathcal{L}_2 < \mathcal{L}_3 < \ldots \tag{31}$$

and consider the status of the runs as they pass any likelihood value L^*. Unless run r has already been terminated, it will at that time have its $N^{(r)}$ points uniformly distributed in X (as always), but subject to their likelihoods being above L^*, so that their enclosed prior masses are leftward of $X^* = X(L^*)$ as illustrated in Figure 9. The run's rightmost point (with worst L) is distributed along X as the largest of $N^{(r)}$ random numbers from $\texttt{Uniform}(0, X^*)$.

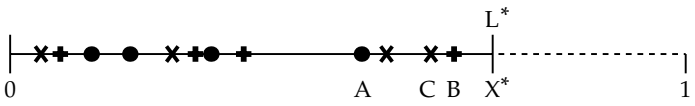

Figure 9: $N = 4$ points each from 3 runs $(\bullet, +, \times)$ as they pass L^*. A,B,C are their rightmost points, respectively. B is the rightmost of all 12 points.

Meanwhile, the rightmost point of all ("B" in Figure 9) is being distributed as the largest of all these $\nu = N^{(1)} + N^{(2)} + N^{(3)} + \ldots$ random numbers from

Uniform$(0, X^*)$. But, if X^* is taken to be the i'th element $\mathcal{X}_i$ of likelihood $\mathcal{L}_i$ in the global sequence, then this rightmost of all is the subsequent element $\mathcal{X}_{i+1}$. The shrinkage factor $t_{i+1} = \mathcal{X}_{i+1}/\mathcal{X}_i$ is distributed as

$$\Pr(t) = \nu t^{\nu-1} \tag{32}$$

Hence the merged combination of individual runs behaves just like a single run with the combined number of points, and can be analyzed as such with no need for any damaging *ad hoc* fixup. Presumably this is how nested sampling might be conveniently implemented on parallel hardware.

PHILOSOPHY

Note that our inferences are properly probabilistic. O'Hagan (1987) has criticised Monte Carlo integration for being frequentist, because results depend on the random generator with its sampling distribution, as well as on the likelihood function. This criticism has entered the folklore and makes it difficult for Bayesians, who have been using frequentist algorithms to generate their results, to argue convincingly against frequentist methodology. Fortunately, the criticism does not apply to nested sampling.

It is true that our results depend on the generator which, by sampling θ within the likelihood constraint, thereby implicitly samples the shrinkage factors t according to (14). Yet nested sampling takes that into account. The numerical integration rule (whether trapezoid or an alternative) imposes a definitive functional form $Z(\mathbf{t})$ on the evidence values, parameterised by the evaluated likelihoods. Accordingly, the known $\Pr(\mathbf{t})$ induces a correspondingly unambiguous $\Pr(Z)$ through the multi-dimensional integral (19).

In the analogy with ordinary Bayesian inference from data, the shrinkage factors $\mathbf{t}$ play the rôle of noise of known distribution, the algorithm plays the rôle of observing equipment, and the likelihood evaluations play data from which we infer $\Pr(Z \mid \text{data})$. Implicit in this is that we knew nothing else of significance about Z or its underlying likelihood function. If we did have some such expectation (Rasmussen and Gharamani 2003), we could factor those probabilities into the calculation, and thereby improve the results. In general it seems unlikely that there would be much to gain, but there might be cases where something useful was known, and the facility is present.

Note that integrating the monotonic function $L(X)$ is quite different from integrating the arbitrary $f(x)$ that O'Hagan had in mind. Suppose, to take a limiting illustration, that we knew $f(0) = f(1) = c$. If f were arbitrary, the uncertainty in its integral would be substantial, and questionable. O'Hagan made that point in greater generality. If f is monotonic, though, it has to be constant, and $\int_0^1 f(x)dx = c$, with no uncertainty. Nested sampling generalises this without any compromise with the ordinary rules of probability theory.

O'Hagan also criticised the Monte Carlo integration of his day for producing results which depended on the sampling distribution. Thus, if the likelihood function factorises as $L = L_1 L_2$, then the result $Z = \int (L_1 L_2)\, dX$ of sampling

over X will differ from the result $Z' = \int L_1 \, (L_2 dX)$ of including L_2 with X and sampling from that combination instead. But that's not a criticism of nested sampling. It's a straightforward feature, to be expected and approved. If we are clever enough to factor part of the likelihood into the prior, and sample from that, we would thereby start closer to the posterior, and can expect to be rewarded with a better estimate having diminished uncertainty. Conversely, if we were foolish enough to retreat away from the posterior by dividing some factor out of the prior, then we should expect to pay for it with increased uncertainty. This is just how a properly constituted algorithm for inference ought to behave. Anything else would be worrisome.

In summary, nested sampling obeys the rules of probability calculus. Accordingly, and as befits an algorithm for Bayesian computation, its nature is Bayesian, not frequentist.

DENSITY OF STATES

The density of states (being the prior mass in a thin likelihood shell — loosely, its area) is often defined with respect to "energy" $E = -\log L$ as $g = dX/dE = -dX/d\log L$, but here it is more convenient to define it in fully logarithmic form as

$$g^*(L) = -\frac{d\log X}{d\log L} \tag{33}$$

Differencing across r steps gives

$$g^*(\overline{L}) \;\longleftarrow\; -\frac{\log X_i - \log X_{i-r}}{\log L_i - \log L_{i-r}} = \frac{-\log t_i - \log t_{i-1} - \ldots - \log t_{i-r+1}}{\log L_i - \log L_{i-r}} \tag{34}$$

for $\overline{L}$ somewhere between L_{i-r} and L_i. The statistics (15) of each $\log t$ are known, and independent, so that in terms of mean and standard deviation

$$g^* \;\longleftarrow\; \frac{(r \pm \sqrt{r})/N}{\log L_i - \log L_{i-r}} \tag{35}$$

As usual in numerical differentiation, the formal uncertainty diminishes as the chosen interval widens, but the difference ratio relates less precisely to the required differential.

Individual steps ($r = 1$) estimate g^* with 100% expected error. Even so, these steps underlie the evidence summation and are the most basic results of the computation. Individual steps can also build properties other than the evidence (known in thermodynamics as the partition function). In particular, the annealed partition function

$$Z(\beta) = \int L^\beta dX \tag{36}$$

is available at any inverse temperature β, provided the computation is carried far enough inward to cover the bulk of the required integral. Nested sampling is not thermal, but can simulate any temperature.

POSTERIOR

Representative samples $\tilde{\theta}$ from the posterior density are defined by sampling from the posterior distribution $p(\theta)$, which is simply the prior weighted by likelihood. Equivalently, they can be obtained by sampling randomly from the area Z under the one-dimensional curve $L(X)$, as shown in Figure 10.

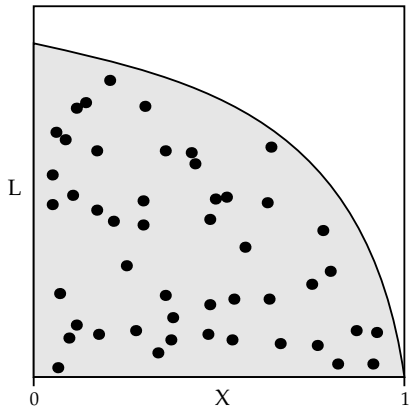

Figure 10: Posterior samples are scattered randomly over the area Z.

Because this area Z is already decomposed (8) into $\sum_i L_i w_i$, we can select our samples by choosing from these sub-areas. In other words, the existing sequence of points $\theta_1, \theta_2, \theta_3, \cdots$ already gives a set of posterior representatives, provided the i'th is assigned the appropriate importance weight $L_i w_i$, normalised by Z to yield a probability with unit total. For a given choice of coefficients $\mathbf{t}$, the posterior probability for point i would be

$$p_i(\mathbf{t}) = L_i w_i(\mathbf{t}) / Z(\mathbf{t}) \tag{37}$$

Taking into account the uncertainty of the coefficients, this posterior probability for point i becomes

$$p_i \longleftarrow \int \frac{L_i w_i(\mathbf{t})}{Z(\mathbf{t})} \Pr(\mathbf{t})\, d\mathbf{t} = L_i \left\langle \frac{w_i(\mathbf{t})}{Z(\mathbf{t})} \right\rangle_{\mathbf{t}} \tag{38}$$

which can be evaluated by Monte Carlo as before.

To obtain equally-weighted posterior representatives, all we need do is accept point θ_i with probability p_i/K, where $K \geq \max_j(p_j)$ must be large enough to avoid duplication. The maximum number of representative posterior samples is given by the entropy (*i.e.* the Shannon channel capacity) as

$$\mathcal{N} = \exp\Big(-\sum_{i=1}^{m} p_i \log p_i \Big) \tag{39}$$

though in practice the available number is somewhat less because of the maximisation limit K. For the approximately rank-C multivariate normal likelihood

conjectured earlier, we may expect $\mathcal{N} \sim N\sqrt{C}$. More precisely, we saw after (11) that this form of likelihood would have a posterior of standard deviation $\sqrt{C/2}$ with respect to $\log X$. The step number i is expected to be roughly $-N \log X$, so p_i should be roughly Gaussian in i, with standard deviation $N\sqrt{C/2}$. In that case, (39) evaluates to $\mathcal{N} = N\sqrt{\pi e C}$. Hence we may assign

$$C \;\longleftarrow\; \mathcal{N}^2/\pi e N^2 \tag{40}$$

as our estimate of the effective rank (number of useful principal components) of the likelihood. It might be nice to know this, and the diagnostic can easily be accumulated along with Z and H.

QUANTIFICATION

Suppose we wish to quantify property $Q(\theta)$, for which point θ_i carries value $Q_i = Q(\theta_i)$. For this, we seek the posterior distribution $\Pr(Q)$. This could be estimated from the posterior samples just obtained, through representative values $Q(\tilde{\theta})$. More accurately, we can use the full sequence of nested points directly, without reducing it to a less informative equally-weighted subset. Each value Q_i is associated with probabilistic weight $p_i(\mathbf{t})$ from (37). In particular, the mean and standard deviation of Q (if they exist) are as usual obtainable from the first and second moments

$$\begin{aligned} \mu(\mathbf{t}) &= \textstyle\sum_{i=1}^{m} Q_i\, p_i(\mathbf{t}) \\ \sigma(\mathbf{t}) &= \Big(\textstyle\sum_{i=1}^{m} Q_i^2\, p_i(\mathbf{t}) \;-\; \mu(\mathbf{t})^2\Big)^{1/2} \end{aligned} \tag{41}$$

for any specified sequence $\mathbf{t}$. Again, we ought to acknowledge uncertainty by invoking the sequence probability. Just as for the evidence in (20), the spread over $\mathbf{t}$ yields mean and deviation

$$\begin{aligned} \mathtt{E}(Q) \;&\longleftarrow\; \text{estimate} \pm \text{uncertainty, from } \{\mu\}_{\mathbf{t}} \\ \mathtt{dev}(Q) \;&\longleftarrow\; \text{estimate} \pm \text{uncertainty, from } \{\sigma\}_{\mathbf{t}} \end{aligned} \tag{42}$$

As was the case for Z, these numerical uncertainties are caused by our limited sequence of nested points, not by any small errors in their evaluation, whether by Monte Carlo or otherwise. With sufficient resources, we could increase N to make these uncertainties arbitrarily small, but of course we would not thereby eliminate $\mathtt{dev}(Q)$, which is caused by the likelihood failing to specify θ precisely, and is part of what we ought to want to know about Q. However, if the numerical uncertainty on the mean $\mathtt{E}(Q)$ exceeded the standard deviation $\mathtt{dev}(Q)$, one might question the utility of the computation, and wish to repeat it more slowly with more points.

ANNEALING

Nested sampling is related to simulated annealing, which uses fractional powers L^β of the likelihood to move gradually from the prior ($\beta = 0$) to the posterior ($\beta = 1$). As the inverse temperature β increases, annealing softly compresses points $\{\theta\}$ sampled from $dP_\beta \propto L^\beta dX$, known as a thermalised ensemble. At stage β, the mean log-likelihood

$$\langle \log L \rangle_\beta = \int \log L \; dP_\beta = \frac{\int L^\beta \log L \; dX}{\int L^\beta dX} = \frac{d}{d\beta} \log \int L^\beta dX \tag{43}$$

is estimated by averaging over the corresponding ensemble. Summing this yields

$$\int_0^1 \langle \log L \rangle_\beta \, d\beta = \log \int L \, dX - \log \int dX = \log Z \tag{44}$$

which is the thermodynamic integration formula. It is not normally accompanied by any estimate of uncertainty, presumably because the uncertainty in $\langle \log L \rangle$ is difficult to assess.

The bulk of the ensemble, with respect to $\log X$, should follow the posterior $dP_\beta \propto L^\beta X \, d\log X$ and be found around the maximum of $L^\beta X$. Under the usual conditions of differentiability and concavity "$\frown$", this maximum occurs where

$$g^* = -\frac{d \log X}{d \log L} = \beta \tag{45}$$

Annealing over β thus tracks the density-of-states g^*, equivalent to $-1/\text{slope}$ on a $\log L / \log X$ plot, whereas nested sampling tracks the underlying abscissa value $\log X$.

As β increases from 0 to 1, one hopes that the annealing maximum tracks steadily up in L, so inward in X (Figure 11a). The annealing schedule that dictates how fast β is increased ought to allow successive posteriors P_β to overlap substantially — exactly how much is still a matter of some controversy. Yet it may not be possible at all.

Suppose that g^* is not an increasing function of $\log X$, so that $L^\beta X$ is not concave (Figure 11b). No matter what schedule is adopted, annealing is supposed to follow the concave hull of the log-likelihood function as its tangential slope flattens. But this will require jumping right across any convex "$\smile$" region that separates ordinary concave "phases" where local maxima of $L^\beta X$ are to be found. At $\beta = 1$, the bulk of the posterior should lie near a maximum of LX, in one or other of these phases. Let us call the outer phase "steam" and the inner phase "water", as suggested by the potentially large difference in volume. Annealing to $\beta = 1$ will normally take the ensemble from the neighbourhood of A to the neighbourhood of B, where the slope is $d\log L / d\log X = -1/\beta = -1$. Yet we actually want samples to be found from the inner phase beyond D, finding which will be exponentially improbable unless the intervening convex valley is shallow. Alternatively, annealing could be taken beyond $\beta = 1$ until, when the ensemble is near the point of inflection C, the supercooled steam

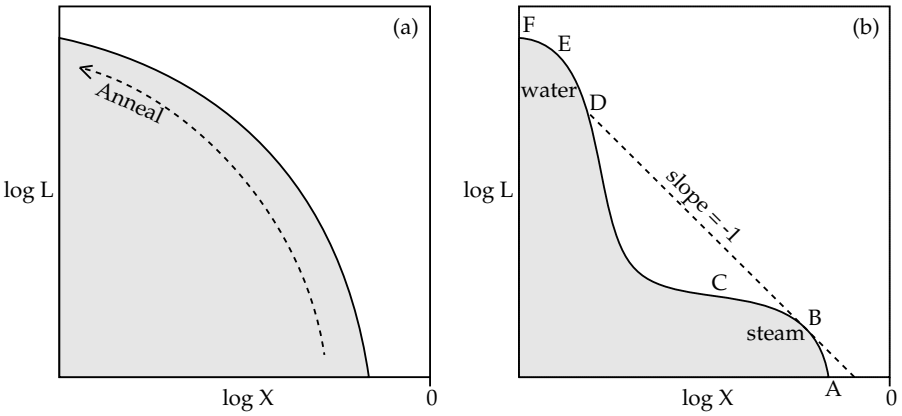

Figure 11: Proper annealing needs log-likelihood to be concave like (a), not (b).

crashes inward to chilled water, somewhere near F. It might then be possible to anneal back out to unit temperature, reaching the desired water phase near E. However, annealing no longer bridges smoothly during the crash, and the value of the evidence is lost. Along with it is lost the internal Bayes factor Pr(states near E)/Pr(states near B) which might have enabled the program to assess the relative importance of water and steam. If there were three phases instead of just two, annealing might fail even more spectacularly. It would be quite possible for supercooled steam to condense directly to cold ice, and superheated ice to sublime directly to hot steam, without settling in an intermediate water phase at all. The dominant phase could be lost in the hysteresis, and inaccessible to annealing.

Phase change problems in general are well known to be difficult to anneal, and especially so when of first order as here. Nested sampling, though, marches steadily down in prior mass X along ABCDEF$\cdots$, regardless of whether the associated log-likelihood is concave or convex or even differentiable at all. There is no analogue of temperature, so there is never any thermal catastrophe. Nested points will pass through the steam phase to the supercooled region, then steadily into superheated water until the ordinary water phase is reached, traversed, and left behind in an optional continued search for ice. All the internal Bayes factors are available, so the dominant phase can be identified and quantified.

PHYSICS EXAMPLE

For an example from physics, consider the following elementary model of order/disorder. Atoms can be in either of two states, 0 or 1. A sequence of n atoms is laid out along a line, so there are 2^n equally-weighted prior states. The atoms define a sequence of clusters c with widths h_c across which the state is constant. For example, the ten atoms 0001111001 have four clusters of width $h_1 = 3$ (zeros), $h_2 = 4$ (ones), $h_3 = 2$ (zeros), $h_4 = 1$ (one). Each cluster has an energy benefit (*i.e.* a log-likelihood gain) proportional to the number

$\frac{1}{2}h(h-1)$ of internal interactions permitted among its members, so that (with specific scaling)

$$\log L = (2/n) \sum_c \tfrac{1}{2} h_c (h_c - 1) \ , \qquad \sum_c h_c = n \tag{46}$$

Of the 1024 states of 10 atoms, the top two (0000000000 and 1111111111) were fully ordered with $\log L = 9$ and shared 49% of the posterior, the next four (0000000001, 0111111111, 1000000000, 1111111110) with $\log L = 7.2$ shared another 16%, the example ten atoms 0001111001 had $\log L = 2$, and so on down to the two lowest states (0101010101 and 1010101010) with $\log L = 0$ which shared 0.006%.

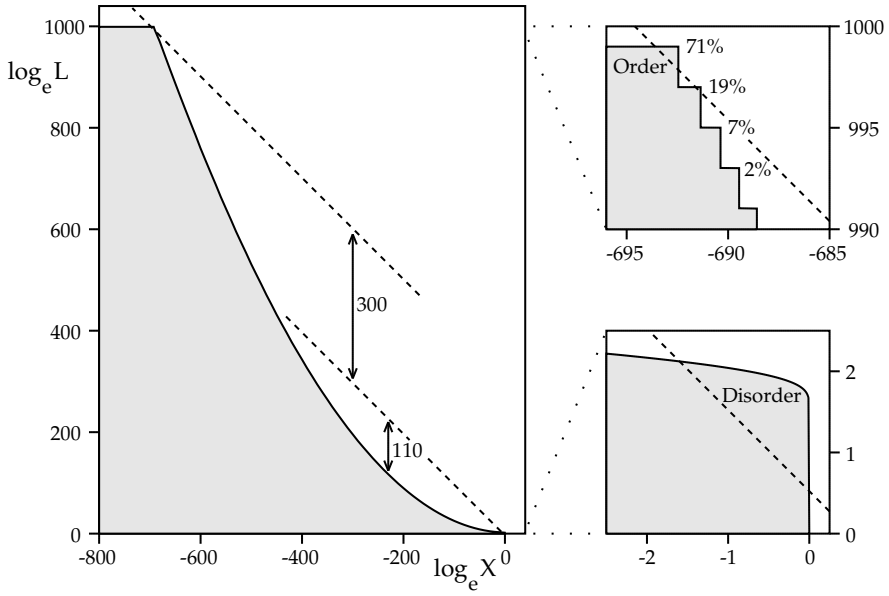

Figure 12: Order/disorder example for 1000 atoms. The upper sub-plot magnifies the "order" phase, and the lower sub-plot magnifies the "disorder" phase. The order phase is favoured by a Bayes factor exp(300) but is hard to find by a factor exp(110). Dashed lines enclose 75% of posterior samples for each phase.

Figure 12 shows the behaviour for $n = 1000$, precisely calculated by recurrence on n. Again, the "order" phase with wide clusters dominates, with the two fully-ordered states $0000\cdots$ and $1111\cdots$ with $\log L = 999$ sharing 71% of the posterior, the next four with $\log L = 997.002$ sharing 19%, and so on. With n being large, the "disorder" phase with most clusters narrow is well separated, but the "order" phase is favoured overall by a Bayes factor of e^{300}.

An ensemble annealed to $\beta = 1$, though, has no chance (technically, about e^{-110} chance) of finding the tiny volume occupied by the "order" states. It ought to transition to the ordered phase at the freezing point $\beta = 0.69$ where the two phases ought to become equally populated, but it won't. As expected, the

author's simulation (which started with a random state and evolved by inverting atoms at random according to the usual Metropolis detailed balance) failed to move away from the "disorder" phase in an allotted trillion trial inversions. Even if by incredible luck it had succeeded in finding the "order" phase, it could not have determined the order/disorder Bayes factor, or the evidence $Z = e^{306.8878}$.

Yet nested sampling (with a fresh sample within the likelihood constraint approximated by allowing trial MCMC inversions of each atom ten times per iterate) successfully estimated $\log L$ as a function of $\log X$ and reached the fully-ordered states steadily, in the expected $NH \approx 700N$ iterates.

DATA ANALYSIS EXAMPLE

It is not just large problems with awkward likelihood functions that exhibit phase changes. For an example in data analysis, consider a small experiment to measure the single coordinate θ, over which the prior $\pi(\theta)$ is flat in (0,1). Its data D yield the likelihood function

$$L(\theta) = 0.99\, q^{-1} e^{-\theta/q} + 0.01 \quad \text{with, say, } q = 10^{-9}. \tag{47}$$

This is already a decreasing function, and the sorting operation of nested sampling is just the identity, $X = \theta$.

An interpretation of (47) is that the experiment was anticipated to work with 99% reliability. If it worked, the likelihood $L = q^{-1} \exp(-\theta/q)$ would have been appropriate, meaning that $\theta \approx 10^{-9}$ was measured. If it failed, which was anticipated 1% of the time, the likelihood would have been the uninformative $L = 1$, because the equipment would just return a random result. Under annealing, the original hot phase is the failure mode. An annealed ensemble limited to $\beta \leq 1$ is most unlikely to find the "working" mode unless it is allowed millions of trials, and will wrongly suggest "failure", with $Z = 0.01$. Only if β is increased far beyond 1 to something above e^{1000} would the ensemble be likely to find the working mode in fewer than millions of trials. Even then, the samples would crash inward and have to be annealed back out through those thousand orders of magnitude. And the evidence value would have been lost.

For nested sampling, which steadily tracks $\log X$ instead of trying to use the slope, these problems are easy. All one needs is the determination to keep going for the NH or so shrinkage steps needed to reach and then cross the dominant mode with a collection of N points. By then, the behaviour of $\log L$ as a function of $\log X$ has been found, so that any distinct phases can be identified along with their Bayes factors, as well as the overall evidence Z.

STATISTICS EXAMPLE

Let the coordinates θ have uniform prior over the 20-dimensional unit cube $[-\frac{1}{2}, \frac{1}{2}]^{20}$, and let the likelihood be

$$L(\theta) = 100 \prod_{i=1}^{20} \frac{1}{\sqrt{2\pi}u} \exp\Big(-\frac{\theta_i^2}{2u^2}\Big) + \prod_{i=1}^{20} \frac{1}{\sqrt{2\pi}v} \exp\Big(-\frac{\theta_i^2}{2v^2}\Big) \tag{48}$$

with $u = 0.01$ and $v = 0.1$. This represents a Gaussian "spike" of width 0.01 superposed on a Gaussian "plateau" of width 0.1. The Bayes factor favouring the spike is 100, and the evidence is $Z = 101$. There is only a single maximum, at the origin, and this should surely be an easy problem. Yet $L(X)$ is partly convex (Figure 13), and an annealing program needs roughly a billion (e^{20}) trials to find the spike, and several times e^{25} to equilibrate properly. On the other hand, H is only 63.2, so nested sampling could reach and cross the spike and cover the whole range of Figure 13 in a mere 100 iterates.

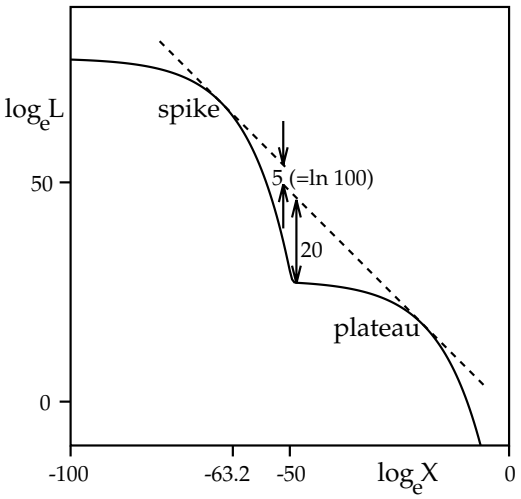

Figure 13: Gaussian spike on plateau. The spike is favoured by a Bayes factor exp(5), but annealing needs exp(20) trials to find it.

Admittedly, about $N = 16$ points would be needed if the uncertainty from

$$\log Z \approx \log(101) \pm \sqrt{63.2/N} \tag{49}$$

(and hence in the spike/plateau Bayes-factor logarithm) needs to be reduced to the ± 2 or so required to identify the favoured (spike) mode with reasonable confidence. That multiplies the computational load to 1600 evaluations, which remains less than a billion.

On the other hand, if the spike was moved off-centre to $(0.2, 0.2, 0.2, \ldots)$, with likelihood

$$L(\theta) = 100 \prod_{i=1}^{20} \frac{1}{\sqrt{2\pi}u} \exp\Big(-\frac{(\theta_i - 0.2)^2}{2u^2}\Big) + \prod_{i=1}^{20} \frac{1}{\sqrt{2\pi}v} \exp\Big(-\frac{\theta_i^2}{2v^2}\Big) \tag{50}$$

then nested sampling too would be in difficulty. There are now two maxima over θ and, at the separatrix contour above which the phases separate, the aperture of the plateau is e^{35} times greater than that of the spike. This means that some huge number of trial points is needed to have a good chance of finding the spike, even though the $\log L / \log X$ plot is indistinguishable from Figure 13. That's impossible in practice. General multi-modality remains difficult.

LIMITING CASES

Two potentially awkward situations merit comment. The first (Figure 14a) is when $\log L$, as a function of $\log X$, has a discontinuity, visualised as a vertical cliff in the plot. Nested sampling is unaffected by this. The likelihood values that accompany a set of nested contours may change the termination condition, but they do not otherwise alter the algorithm's progress. The cliff doesn't matter. For the possible alternative method of multicanonical simulation (Berg

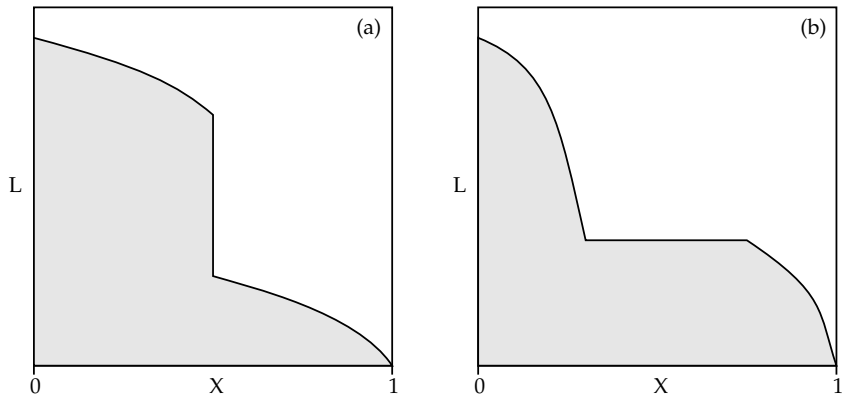

Figure 14: Cliff (a) and plateau (b) likelihood functions.

& Neuhaus 1991), the cliff would matter. Multicanonical simulation incorporates procedures to re-weight points artificially until $n(E) \equiv g = -dX/d\log L$ becomes nearly uniform. Having done that, the results can simulate any temperature. In this, the method is similar to nested sampling, though it lacks such clean numerical uncertainties. However, the discontinuous cliff will be difficult, because the formal aim of multicanonical sampling becomes unattainable. There is a range of L over which there are no prior states at all, so their density $n(\log L)$ can't be re-weighted to the desired constant. Nested sampling's $\log X$ is the abscissa of choice, not multicanonical simulation's $\log L$.

The other potentially awkward situation (Figure 14b) is when $\log L$, as a function of $\log X$, has a plateau. In this case, a finite prior mass is associated with one particular likelihood value. If the plateau covers a wide range, it may be difficult to locate the small interior domain in which L takes larger, possibly dominant, values. After all, the likelihood is offering no guidance, and the domain might have to be found (or not) by random exploration. It might then be very difficult to obtain useful new samples. Even so, it may be possible to generate them efficiently, by choosing labels ℓ so that small values anticipate easy transition to larger likelihoods. In this way, a judicious choice of extended likelihood L^+ (equation (6)) might give guidance even where L does not. Nested sampling would then continue to work without its samples becoming unduly expensive. However, "difficult" shapes, where the posterior breaks into separated islands whose mass is poorly predicted by their volume, will likely remain difficult.

OVERVIEW

Nested sampling reverses the accepted approach to Bayesian computation by putting the evidence first. A conventional collection of posterior samples can be computed as the calculation proceeds, but that is an optional extra, and quantified properties are better accumulated directly. The procedure runs with an evolving collection of N points, where N can be chosen small for speed or large for accuracy. In a specific application, it is the user's task to sample according to the prior density subject to a hard constraint on likelihood value. Helpfully, $N-1$ such points are already available as guides.

Nested sampling proceeds by systematically constraining the available prior mass, steadily decreasing its logarithm according to the shape of the likelihood contours. Its evolution path is independent of the actual likelihood values. This invariance over monotonic re-labelling may make it easier to find analytic properties that might imply almost-certain convergence. Currently, lack of convergence proofs for MCMC procedures is the main missing ingredient in Bayesian calculations, and with nested sampling attention can be focussed cleanly on the shape of the contours, without any distraction from likelihood values.

We do not address the errors that would arise from imperfect sampling of the box within a given likelihood contour. The uncertainties that arise from the method itself, though, are understood and controllable. Numerical uncertainties accompany estimates of evidence and any quantified property. These are properly probabilistic, and not derived from jack-knife or similar frequentist fixup. In short, nested sampling follows the rules of probability calculus, so is Bayesian in nature, as befits an algorithm for Bayesian computation. As usual with probabilistic procedures, uncertainties decrease as $\sqrt{N}$ whereas the computational load is proportional to N.

For some problems, it may be possible to find provably exact samples (Propp & Wilson, 1996) within a likelihood contour, and thus remove any doubt concerning imperfect sampling. Because the probabilities p_i of the nested samples are calculated essentially perfectly, the posterior samples generated through nested sampling would also be exact, as would the quantification statistics Q derived from the complete nested sequence. If exact samples turn out to be easier to find within a likelihood contour than with respect to the full posterior, this would extend the currently small class of problems amenable to exact sampling, with the added benefit of obtaining the evidence value.

Nested sampling has some similarity with annealing in that it works by compressing the available domain. The hard outer constraint on likelihood happens to give a similar restriction on step-length to that applying in standard Metropolis-type detailed balance, so the new method should offer no great gain or loss of computational speed, as compared with annealing under an efficient schedule. Even so, nested sampling is more fundamental than annealing, in that it gives a direct view of the underlying density of states $g^*(L)$ as it steps steadily inward. More importantly, it can deal straightforwardly with convex likelihood functions that exhibit first-order phase changes. Nested sampling has the simplicity and generality that speak of wide-ranging power.

ACKNOWLEDGMENTS

This work was supported by Maximum Entropy Data Consultants Ltd of Cambridge (England). It was Ken Hanson at the 23rd maximum entropy workshop who re-awakened my long-standing interest in probabilistic integration. I also wish to thank David MacKay and the Inference group journal club at Cambridge for useful encouragement and advice, and for hosting nested-sampling papers and programs (currently at www.inference.phy.cam.ac.uk/bayesys). An early account of the method was given in Skilling (2004). This presentation has been considerably improved by the JBA referees, who well succeeded in their task of being critical.

REFERENCES

Berg, B. A. and Neuhaus, T. (1991) "*Multicanonical algorithms for first-order phase transitions*", Phys. Lett. **B267**, 249–253

Duane, S., Kennedy, A.D., Pendleton, B.J. and Roweth, D. (1987) "*Hybrid Monte Carlo*", Phys. Lett. **B195**, 216–222

Gelman, A. and Meng, X.-L. (1998) "*Simulating normalizing constants: from importance sampling to bridge sampling to path sampling*", Statistical Science, **13**, 163–185.

Hastings, W.K. (1970) "*Monte Carlo sampling methods using Markov chains and their applications*", Biometrika, **57**, 97–109.

MacKay, D.J.C. (2003) "*Information Theory, Inference, and Learning Algorithms*", p.379, Cambridge Univ. Press.

McDonald, I.R. and Singer, K. (1967) "*Machine calculation of thermodynamic properties of a simple fluid at supercritical temperatures*", J. Chemical Physics, **47**, 4766–4772.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) "*Equation of state by fast computing machines*", J. Chemical Physics, **21**, 1087–1092.

Neal, R. (1993) "*An improved acceptance procedure for the hybrid Monte Carlo algorithm*", J. Computational Physics, **111**, 194–203.

Neal, R. (2003) "*Slice sampling*", Annals of Statistics, **31**, 705–767.

O'Hagan, A. (1987) "*Monte Carlo is fundamentally unsound*", The Statistician, **36**, 247–249.

Propp, J. G. and Wilson, D. B. (1996) "*Exact sampling with coupled Markov chains and applications to statistical mechanics*", Random Structures and Algorithms **9**, 223–252.

Rasmussen, C. E. and Gharamani, Z. (2003) "*Bayesian Monte Carlo*", in Advances in Neural Information Processing Systems **15**, MIT Press.

Skilling, J. (2004) "*Nested Sampling*", in Bayesian inference and maximum entropy methods in science and engineering, ed. R.Fischer, R. Preuss, U. von Toussaint, Amer. Inst. Phys. Conference Proc. **735**, 395–405.

APPENDIX

```
/*=============================================================
  TOY NESTED SAMPLING PROGRAM IN 'C' by John Skilling, Aug 2005
  =============================================================*/
#include <stdio.h>
#include <stdlib.h>
#include <math.h>
#include <float.h>
#define UNIFORM   ((rand()+0.5) / (RAND_MAX+1.0)) // Uniform(0,1)
#define logZERO   (-DBL_MAX * DBL_EPSILON)        // log(0)
#define PLUS(x,y) (x>y ?  x+log(1+exp(y-x)) :  y+log(1+exp(x-y)))
                      // logarithmic addition log(exp(x)+exp(y))
/* YOU MUST PROGRAM THIS FROM HERE ++++++++++++++++++++++++++++++
typedef struct
{
   ANYTYPE theta;  // YOUR coordinates
   double  logL;   // logLikelihood = ln Prob(data | theta)
   double  logWt;  // ln(Weight), summing to SUM(Wt) = Evidence Z
} Object;
double logLhood(ANYTYPE theta){...}     // logLikelihood function
void Prior  (Object* Obj){...}   // Set Object according to prior
void Explore(Object* Obj, double logLstar){...}
                    // Evolve Object within likelihood constraint
-------------------------------------------------- UP TO HERE */
int main(void)
{
#define N 100             // # Objects
#define MAX 9999          // max # Samples (allow enough)
   Object Obj[N];         // Collection of N objects
   Object Samples[MAX];   // Objects defining posterior
   double logw;           // ln(width in prior mass)
   double logLstar;       // ln(Likelihood constraint)
   double H = 0.0;        // Information, initially 0
   double logZ = logZERO; // ln(Evidence Z, initially 0)
   double logZnew;        // Updated logZ
   int    i;         // Object counter
   int    copy;      // Duplicated object
   int    worst;     // Worst object
   int    nest;      // Nested sampling iteration count
   double end = 2.0; // Termination condition nest = end * N * H
// Set prior objects
   for( i = 0; i < N; i++ )
       Prior( &Obj[i] );
// Outermost interval of prior mass
   logw = log(1.0 - exp(-1.0 / N));
```

```
// Begin Nested Sampling loop ++++++++++++++++++++++++++++++++++
   for( nest = 0; nest <= end * N * H; nest++ )
   {
// Worst object in collection, with Weight = width * Likelihood
       worst = 0;
       for( i = 1; i < N; i++ )
           if( Obj[i].logL < Obj[worst].logL )
               worst = i;
       Obj[worst].logWt = logw + Obj[worst].logL;
// Update Evidence Z and Information H
       logZnew = PLUS(logZ, Obj[worst].logWt);
       H = exp(Obj[worst].logWt - logZnew) * Obj[worst].logL
          + exp(logZ - logZnew) * (H + logZ) - logZnew;
       logZ = logZnew;
// Posterior Samples (optional, care with storage overflow)
       Samples[nest] = Obj[worst];
// Kill worst object in favour of copy of different survivor
       do copy = (int)(N * UNIFORM) % N;  // force 0 <= copy < N
       while( copy == worst && N > 1 );   // don't kill if N=1
       logLstar = Obj[worst].logL;   // new likelihood constraint
       Obj[worst] = Obj[copy];       // overwrite worst object
// Evolve copied object within constraint
       Explore( &Obj[worst], logLstar );
// Shrink interval
       logw -= 1.0 / N;
   } // ------------------------------ end nested sampling loop
// Begin optional final correction, should be small ++++++++++++++
   logw = -(double)nest / (double)N - log((double)N);   // width
   for( i = 0; i < N; i++ )
   {
       Obj[i].logWt = logw + Obj[i].logL;  // width * Likelihood
// Update Evidence Z and Information H
       logZnew = PLUS(logZ, Obj[i].logWt);
       H = exp(Obj[i].logWt - logZnew) * Obj[i].logL
          + exp(logZ - logZnew) * (H + logZ) - logZnew;
       logZ = logZnew;
// Posterior Samples (optional, care with storage overflow)
       Samples[nest++] = Obj[i];
   } // --------------------------- end optional final correction
// Exit with evidence Z, information H, and posterior Samples
   printf("#samples = %d\n", nest);
   printf("Evidence:  ln(Z) = %g +- %g\n", logZ, sqrt(H/N));
   printf("Information:  H = %g nats = %g bits\n", H, H/log(2));
// You can now accumulate results from Samples[0...nest-1]
   return 0;
}
```