Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics
Benidorm (Alicante, Spain), June 1st–6th, 2006

Nested Sampling for Bayesian Computations

John Skilling
Maximum Entropy Data Consultants Ltd, Kenmare, Ireland.
skilling@eircom.ie

Summary

Nested sampling reverses the usual approach to Bayesian computation by directly targeting the value of the *evidence* (alternatively the marginal likelihood, marginal density of the data, or the prior predictive). Samples from the posterior distribution are an optional by-product. Nested sampling is a simple but general method, and although non-thermal itself, it can simulate thermal results at any temperature. It is invariant over monotonic re-labelling of likelihood values, which allows it to deal with various phase-change problems which effectively defeat thermal methods.

Keywords and Phrases: Bayesian computation, evidence, marginal likelihood, model selection, algorithm, nest, annealing, non-thermal, phase change.

1. INTRODUCTION

1.1. *Bayesian computation from the beginning*

Bayesian inference yields two results. One is the posterior distribution $\Pr(\theta \mid D, H)$ of the parameters θ of interest, in the light of data D and in the context of hypothesis H. The other is the support $\Pr(D \mid H)$ for the data under that hypothesis — variously called the prior predictive (how it's often used), the marginal likelihood (how it's often made), or the evidence (what it is). Good practice suggests giving a crisp name to a centrally important quantity, and I follow the physicists (Mackay, 2003) in using the term "evidence".

The evidence should come first. Model selection relies on it. It is the evidence that guides our choice of model, through the ratios known as Bayes factors. Indeed, why bother calculating a posterior at all if its evidence is poor? At the very least, the value of the evidence ought to be quoted after any probabilistic calculation, as a courtesy to other workers who might wish to analyse the same data differently. The Bayesian's primary task, then, is to evaluate the scalar

$$\text{evidence} = Z = \int L(\theta)\, \pi(\theta)\, d\theta = \int L\, dX \qquad (1)$$

John Skilling was previously at Dept. of Applied Mathematics and Theoretical Physics, University of Cambridge.

where $L = L(\theta)$ is the likelihood function and $\mathrm{d}X = \pi(\theta)\,d\theta$ is the element of mass associated with prior density $\pi(\theta)$.

With this in place, the distribution of posterior mass follows as

$$\text{posterior} = \; dP = p(\theta)\,d\theta = Z^{-1}L(\theta)\,\pi(\theta)\,d\theta \tag{2}$$

In this paper, nested sampling (Skilling 2004, 2006) is developed as a general way of evaluating the integral in (1). The evidence is the prime target, from which representative samples from the posterior (2) follow as an optional by-product.

This methodology reverses the traditional approach dating back to Metropolis *et al.* (1953), in which emphasis was placed on calculating the posterior, usually as a set of random samples. The evidence was relegated to a secondary rôle, usually calculated (if at all) as a by-product (Gelman & Meng, 1998) of algorithms such as simulated annealing which are principally designed to compute the posterior. Now, it comes first.

1.2. *Sorting to one dimension*

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 $X(0) = 1$ to $X(\infty) = 0$.

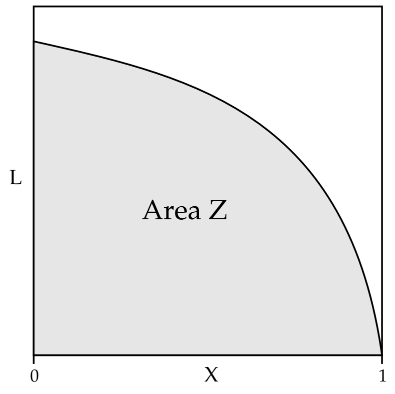

Figure 1: *Sorted likelihood function with area Z.*

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 cannot be too badly behaved. Accomplishing this transformation from θ to X involves dividing the unit prior mass into tiny elements, and sorting them by likelihood.

A very simple example, on a 4×4 grid of two-dimensional θ, is the table (Figure 2a) of likelihood values ascribed to its 16 cells of equal prior mass $\frac{1}{16}$. Our plan is to proceed as if we could sort these elements by likelihood, in this example to $L = (30,29,27,25,23,21,20,18,17,13,12,11,10,9,5,2)$, whence Z is evaluated right-to-left as $\frac{30}{16}+\frac{29}{16}+\frac{27}{16}+\frac{25}{16}+\frac{23}{16}+\frac{21}{16}+\frac{20}{16}+\frac{18}{16}+\frac{17}{16}+\frac{13}{16}+\frac{12}{16}+\frac{11}{16}+\frac{10}{16}+\frac{9}{16}+\frac{5}{16}+\frac{2}{16} = 17$ into domains of progressively greater likelihood as shown in Figure 2b.

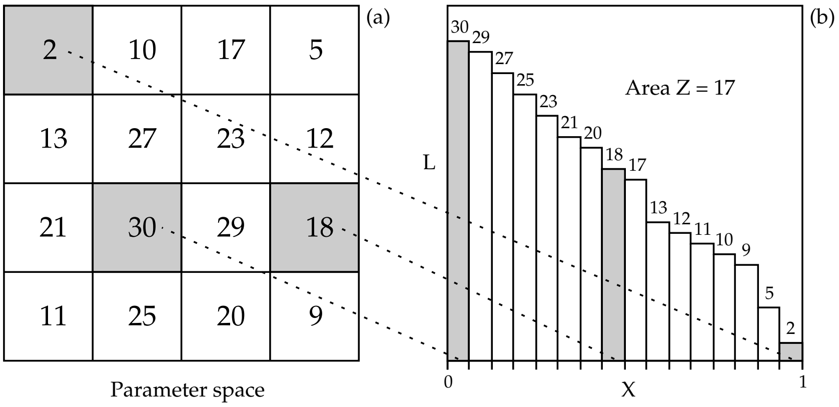

Figure 2: *Likelihood values (a) in parameter space, (b) sorted as $L(X)$.*

Actually doing the sorting would usually be expensive, but it's possible *in principle*. As a technicality, nested sampling requires the likelihood function $L(X)$ to be *strictly* decreasing to make the mapping between θ and X unambiguous. To ensure this, we need to resolve ties between points of equal L. An object 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 parameterize within each likelihood contour, and extend the likelihood to

$$L_k^+ = L_k + \epsilon \ell_k \tag{5}$$

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 objects, even where raw likelihoods are equal. With this refinement understood, we take $L(X)$ to be strictly decreasing.

1.3. *Integration in one dimension*

Coordinate-dependent complications of geometry, topology, even dimensionality, are all annihilated by the sorting operation, and the remaining task of one-dimensional integration 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

$$0 < X_m < \cdots < X_2 < X_1 < 1 \tag{6}$$

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

$$Z = \sum_{i=1}^{m} L_i w_i \tag{7}$$

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

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

$$Z = \int_0^1 L\,dX \;\ge\; \sum_{i=1}^{m} L_i (X_i - X_{i+1}) \tag{8}$$

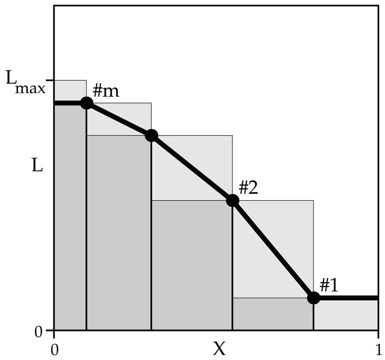

Figure 3: *Lower bound (dark shading) and upper bound (all shading) on area. The thick line indicates the trapezoidal rule.*

There is a similar upper bound (Figure 3) from $w_i = X_{i-1} - X_i$ with $X_0 = 1$,

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

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 (9) is relevant, and errors from numerical integration are at most $\mathcal{O}(N^{-1})$, that being the difference between upper and lower bounds.

The trapezoidal rule $w_i = \frac{1}{2}(X_{i-1} - X_{i+1})$ reduces this to $\mathcal{O}(N^{-2})$ in most cases. The integrand is already well behaved, so further improvement is not expected.

1.4. *Logarithmic sampling*

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

$$H = \int \log(\,dP/\,dX)\,dP = \text{information.} \tag{10}$$

H is (minus) the logarithm of the compression ratio, being that fraction of prior mass that contains the bulk of the posterior mass. It may well be of the order of thousands or more in practical problems where the likelihood is concentrated in some exponentially small corner of the prior domain.

As an example, suppose that the likelihood function has R approximately-Gaussian principal components, so that L is approximately a rank-R multivariate normal. In accordance with the "$\chi^2 = R \pm \sqrt{2R}$" folklore, the shell containing most of the posterior mass would be fairly broadly distributed over a range $\Delta \log X \sim \sqrt{R}$. Moreover, 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 R, let alone $\sqrt{R}$, confirming general experience 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 and big meaning $\sqrt{R}$) is widely seen in practical applications.

To cover such a range, sampling ought to be geometrical rather than linear in X, so we write

$$X_1 = t_1,\ \ X_2 = t_1 t_2,\ \ \cdots\cdots,\ \ \ X_i = t_1 t_2 \ldots t_i,\ \ \cdots\cdots \tag{11}$$

with

$$t_i = X_i / X_{i-1}\,, \quad 0 < t_i < 1\,. \tag{12}$$

It is these ratios $\mathbf{t}$ that control the 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 $100\sqrt{R}$ steps. Any such sequence $\mathbf{t}$ would lead to an estimate of Z, which we would make explicit by writing

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

according to the trapezoidal (or other) rule. This elementary integration scheme *appears* to rely upon explicit and impractical sorting, but actually *it need not.*

2. NESTED SAMPLING

2.1. *The idea*

Although we cannot usually set precise values of t, it turns out that we can often set them statistically, and that is enough. The resulting value of Z will have a corresponding uncertainty, but that is tolerable because we can estimate it. The simplest way of obtaining a random t less than 1 is to set

$$t_i = \texttt{Uniform}(0,1)\,, \text{ from } \Pr(t) = 1\,. \tag{14}$$

In principle, such an object 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.

$$\theta_i = \texttt{Sort}^{-1}\Big(\text{Uniform in } X < X_{i-1}\Big) \tag{15}$$

In practice, it is (much) easier to obtain θ_i directly, by sampling within the equivalent constraint $L(\theta) > L_{i-1}$ in proportion to the prior density $\pi(\theta)$. (At the start, set $L_0 = 0$ to ensure complete initial coverage.)

$$\theta_i = \texttt{Sample}\Big(\text{Prior within } L > L_{i-1}\Big) \tag{16}$$

After all, this constraint on L is equivalent to $X < X_{i-1}$ because $L(X)$ is a decreasing function. Each sampling method yields a random prior element within the common constraint, so the two are equivalent. But the likelihood constraint (16) bypasses explicit use of X. *So we don't need to sort at all!*

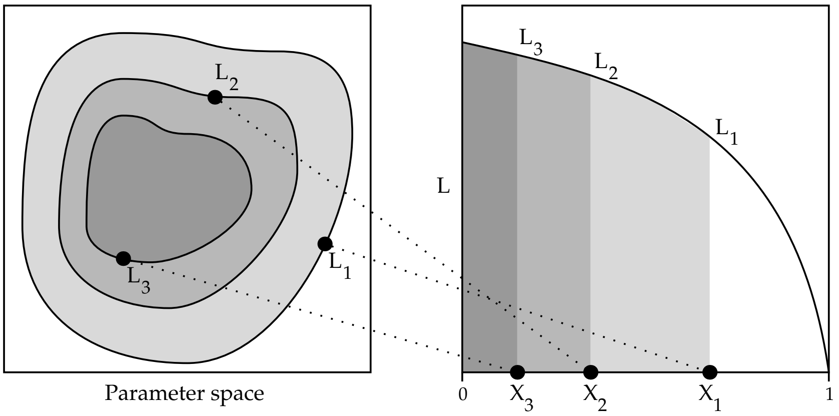

Figure 4: *Nested likelihood contours are sorted to enclosed prior mass X.*

This is illustrated in Figure 4, in which prior mass is represented by area on the left. Thus, object 2 is found by sampling over the prior within the box defined by

$L > L_1$, and so on. Such objects will usually be found by some MCMC approximation, starting at an object $\widetilde{\theta}$ known to obey the constraint (if available), or at worst starting at θ_{i-1} which lies on and defines the current likelihood boundary. We assume that we can do this, noting that sampling within a hard constraint will be (if anything) easier than the traditional Metropolis-Hastings sampling involving likelihood-weighting and detailed-balance.

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 objects exactly on the same contour being ranked by their labels ℓ. More generally, instead of taking 1 object 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. The shrinkage ratio $t_i = X_i/X_{i-1}$ is now distributed as

$$\Pr(t_i) = N t_i^{N-1} \text{ in } (0, 1)\,, \tag{17}$$

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

$$\mathtt{E}(\log t) = -1/N, \qquad \mathtt{dev}(\log t) = 1/N\,. \tag{18}$$

The individual $\log t$ are all 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 $N\sqrt{R}$ 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 objects anew at each step, because $N-1$ of them are already available, being the survivors after deleting the worst. Only one new object 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 object by a copy of a random survivor, evolved within the box by MCMC for some adequate number of trials. Surviving objects could be used as stationary guides in such exploration. Another method might be generation of new objects by genetic mixing of the survivors' coordinates. All that matters is that the step ends with N usably independent objects within the constraint.

2.2. *The procedure*

At each step, nested sampling has N objects $\theta_1, \ldots, \theta_N$ with corresponding likelihoods $L(\theta_1), \ldots, L(\theta_N)$. The likelihood L_i associated with step i is the lowest of these values. There are to be j iterative steps.

```
Start with $N$ objects $\theta_1, \ldots, \theta_N$ from prior;
        initialize $Z = 0$, $X_0 = 1$, and $H = 0$.
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 include its uncertainty,
        set $w_i = X_{i-1} - X_i$ (simple) or $(X_{i-1} - X_{i+1})/2$ (trapezoidal),
        increment $Z$ by $L_i w_i$ and update $H$ likewise, then
        replace object of lowest likelihood by new one drawn
            from within $L(\theta) > L_i$, in proportion to the prior $\pi(\theta)$.
In principle, complete $Z$ with (simply) $N^{-1}(L(\theta_1) + \ldots + L(\theta_N))\,X_j$.
```

The last step uses the surviving objects to fill in the final band $0 < X < X_j$ of the full integral from 0 to 1, after the iterative steps have covered the domain outside X_j. So the final number of terms in the evidence summation (7) becomes $m = j+N$. However, there should already have been sufficient steps j to accumulate most of the integral, so this final increment ought to be unimportant.

Figure 5 illustrates the method, running with $N = 3$ objects. Initially, three objects are taken from the unconstrained prior, whose mass is represented by the complete square area on the left. These objects 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 objects is identified as number 1, with likelihood L_1. It is then replaced by a new object, 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 objects, now all uniform in the reduced range $(0, X_1)$. With the particular random numbers used for Figure 5, the new object in step 1 happened to lie outside the two survivors, so became number 2, but in step 2 the new object happened to be the innermost, and was eventually identified as number 5. After the $j = 5$ allotted steps, the five discarded objects 1,2,3,4,5 are augmented with the final three survivors 6,7,8 to give the $m = 8$ objects $(X_1, \ldots, X_8)$ shown on the top line. It is over these objects that the sum $\sum_{i=1}^{8} L_i w_i$ is evaluated to estimate Z.

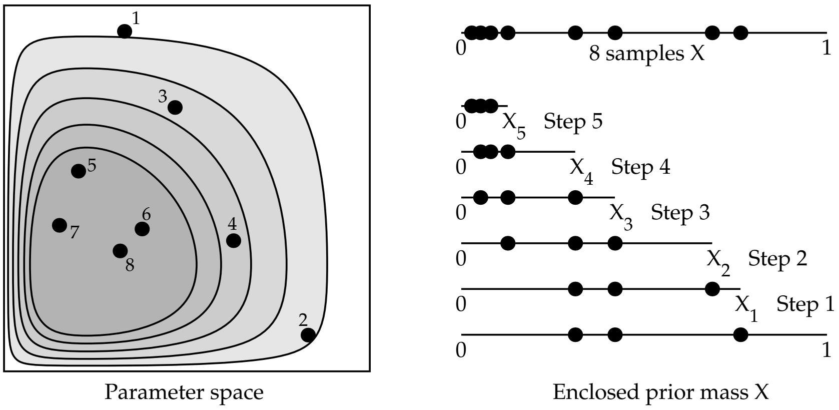

Figure 5: *Likelihood contours shrink by factors* $\exp(-1/3)$ *in area and are roughly followed by successive objects 1,2,3,4,5.*

With $N = 3$ objects, shrinkage is expected to be roughly geometrical, by $\Delta \log X \sim -1/3$ per step. The diagram on the left of Fig. 5 shows likelihood contours drawn at levels corresponding to enclosed areas diminishing by this factor — *i.e.* the i'th contour encloses prior mass $e^{-i/3}$. Indeed, the first object lies close to the first contour, the second object is not too far outside the second contour, and so on until the fifth object 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 will derive a probabilistic estimate instead.

2.3. *Termination*

Termination of the main loop could simply be after a pre-set number of steps, as used for simplicity in the example code (Sivia & Skilling, 2006) of the Appendix. Better, it could be when even the largest current likelihood, taken over the full currently available prior mass, would not increase the current evidence by more than some small fraction f;

$$\max\Big(L(\theta_1), \dots, L(\theta_N)\Big)\, X_j < f Z_j \quad\Longrightarrow\quad \text{termination.} \tag{19}$$

Plausibly, the accumulation of Z is then tailing off, so the sum is nearly complete. If an analytical upper bound $L < L_{\max}$ can be found, such as when a Gaussian likelihood factor cannot exceed $1/\sqrt{2\pi}\sigma$, it can be used in (19) to give a firmer termination criterion

$$L_{\max} X_j < f Z_j \quad\Longrightarrow\quad \text{termination.} \tag{20}$$

In this case, all but a fraction f of Z has (almost-)definitely been found.

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 around this maximum, which occurs in the region of $X \approx e^{-H}$. Remembering $X_i \approx e^{-i/N}$, this suggests an alternative termination condition

$$\text{“continue iterating until the count } i \text{ significantly exceeds } NH\text{”} \tag{21}$$

which still expresses the general aim that a nested-sampling calculation should be continued until most of Z seems to have been found. (Of course, H is here the current evaluate from the previous i iterates.)

Unfortunately, there can be no rigorous criterion based on sampling alone which ensures 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.

2.4. *Numerical uncertainty*

It is possible to run nested sampling crudely, by assigning each compressive $\log t$ 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, we have already seen that if the algorithm takes NH steps to reach the bulk of the posterior, that number will be subject to Poisson uncertainty $\sqrt{NH}$. That translates to a geometrical uncertainty factor $\exp(\pm\sqrt{H/N})$ (which could be many powers of e) in the weights w of the dominating iterates, which scales directly into the corresponding standard-deviation uncertainty

$$\texttt{dev}(\log Z) \approx \sqrt{H/N} \tag{22}$$

For most practical purposes, this estimate suffices. Alongside this uncertainty in $\log Z$, any systematic numerical bias imposed by the integration rule is usually less than a single factor of e, hence negligible.

A more complete 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 objects θ_i, but we do not know which. Instead, the sequence probability $\Pr(\mathbf{t})d\mathbf{t} = \prod_i N t_i^{N-1} dt_i$ from (17) induces a distribution for the estimates of Z:

$$\Pr(Z) = \int \delta\Big(Z - \sum_{i=1}^{m} L_i w_i(\mathbf{t})\Big) \; \Pr(\mathbf{t})\, d\mathbf{t} \tag{23}$$

This can be estimated by Monte Carlo, by 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 the samples $\{Z(\mathbf{t})\}$.

$$\log Z = \text{estimate} \pm \text{uncertainty}, \quad \text{from } \{\log Z(\mathbf{t})\} \tag{24}$$

Just as in (22), the uncertainty accompanying these more refined 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.

2.5. *Posterior and Quantification*

Nested sampling allows posterior samples to be extracted from the evidence calculation, trivially reversing the traditional approach. Representative samples from the posterior density are defined by sampling from the posterior distribution $p(\theta)$, which is simply the prior weighted by likelihood as represented by the area under $L(X)$, illustrated in Figure 6.

In other words, the existing sequence of objects $\theta_1, \theta_2, \theta_3, \cdots$ already gives a set of posterior representatives, provided the i'th is assigned the appropriate importance weight $L_i w_i$, normalized by Z to yield a probability with unit total. For a given choice of coefficients $\mathbf{t}$, the posterior probability for object i would be

$$p_i(\mathbf{t}) = L_i w_i(\mathbf{t}) / Z(\mathbf{t}) \tag{25}$$

with

$$p_i = L_i e^{-i/N} / Z \tag{26}$$

as the standard crude assignment for a run with N objects. There will usually be little reason to adopt the sophistication of averaging (25) over $\mathbf{t}$ (presumably by Monte Carlo), so that (26) should usually suffice. If wanted, equally-weighted samples can be obtained randomly from the area Z, with object θ_i selected proportionally to p_i. About $N\sqrt{R}$ different ones will usually be available, that being the

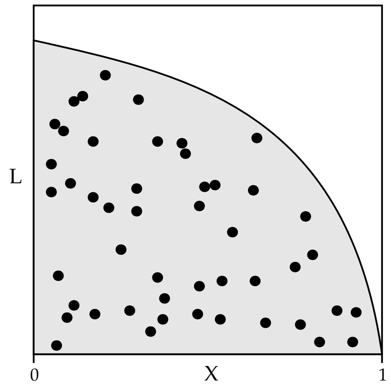

Figure 6: *Posterior samples are scattered randomly over the area Z.*

range covering the bulk of the posterior. They could be used as seeds for traditional Metropolis-Hastings exploration (section 3.2) if yet more were needed.

In order to quantify some property $Q(\theta)$, we seek the posterior distribution $\Pr(Q)$, which comes directly from the weighted sequence (25) or (26). In particular, the mean and standard deviation of Q are as usual obtainable from the first and second moments

$$\mu = \sum_{i=1}^{m} p_i\, Q(\theta_i)\,, \qquad \mu^2 + \sigma^2 = \sum_{i=1}^{m} p_i\, Q(\theta_i)^2 \tag{27}$$

where p_i is set by (26) as standard, or by (25) with subsequent averaging over $\mathbf{t}$ if numerical uncertainties are needed.

The availability of the posterior distribution and its quantification completes nested sampling as a system for Bayesian inference.

3. COMPARISONS

3.1. *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 a thermalized ensemble of objects $\{\theta\}$ sampled from $dP_\beta \propto L^\beta\, dX$. At stage β, the mean log-likelihood

$$\langle \log L \rangle_\beta = \int \log L\; dP_\beta = \frac{\int L^\beta \log L\; dX}{\int L^\beta\, dX} = \frac{d}{d\beta} \log \int L^\beta\, dX \tag{28}$$

is estimated by averaging over the corresponding ensemble of objects at that stage of computation. Summing this yields

$$\int_0^1 \langle \log L \rangle_\beta \, d\beta = \log \int_0^1 L \, dX - \log \int_0^1 dX = \log Z \tag{29}$$

which is the thermodynamic integration formula. The corresponding uncertainty is mostly ignored — presumably because the uncertainty in $\langle \log L \rangle$, though obviously present, is difficult to assess reliably.

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

$$g^* = -\frac{d \log X}{d \log L} = \beta \tag{30}$$

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 $\log X$.

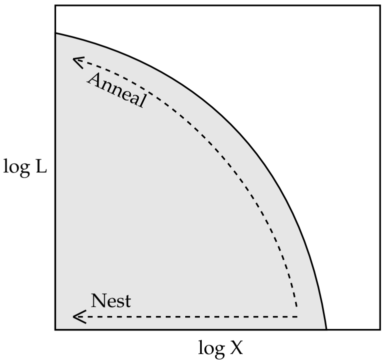

Figure 7: *Annealing tracks the tangent, nested sampling tracks the abscissa.*

As β increases from 0 to 1, one hopes that the annealing maximum tracks steadily up in L, so inward in X (Figure 7). The annealing schedule that dictates how fast the slope $1/\beta$ flattens ought to allow successive posteriors P_β to overlap substantially — exactly how much is still a matter of some controversy. Nested sampling has *no such schedule*: it only needs the assignment of an adequate number N of objects.

3.2. *Metropolis-Hastings*

Nested sampling requires objects θ to be drawn from the prior while obeying a hard constraint $L(\theta) > L_{\min}$. Some form of MCMC will usually be used to approximate this. Assuming a transition scheme $\theta \to \theta'$ that samples the prior faithfully when unconstrained, nested sampling is implemented by the transition rule:

$$\text{accept } \theta' \text{ if and only if } L(\theta') > L_{\min}\,. \tag{31}$$

Of course, the starting point θ will already obey the constraint, so that sufficiently small transitions will normally be accepted.

The traditional Metropolis-Hastings rule (Hastings 1970), aiming for the posterior instead of the evidence, is (or is equivalent to)

$$\text{accept } \theta' \text{ if and only if } L(\theta')^{\beta} > L(\theta)^{\beta} \times \texttt{Uniform}(0,1). \tag{32}$$

Nested sampling is slightly simpler, but the two rules are very similar. At similar stages in their iterations, both methods allow similar step-lengths, so their exploratory speeds are much the same.

However, nested sampling relies *only* on the *shape* of the likelihood contours, and *not* on the likelihood *values*, whether or not these are annealed. This invariance over monotonic relabelling of likelihood contours makes nested sampling independent of the quirks of likelihood value, hence *more robust.*

3.3. *Slice sampling*

Slice sampling (Neal, 2003) is a way of adjusting the MCMC step length "on the fly" to ensure a successful transition. An acceptance level is set, traditionally (32) as for Metropolis-Hastings, and the step length of trial transitions is adjusted either inward or outward (subject to detailed balance) until an appropriate success. Exactly the same technique works with nested sampling, except that the acceptance level is slightly simplified to (31) instead.

3.4. *Importance sampling*

Suppose that some factor $f(\theta)$ is extracted from the likelihood and included in the prior instead. If it is reasonably efficient to sample from the revised prior $f(\theta)\pi(\theta)$, then we can use nested sampling with revised likelihood L/f to compute

$$Z = \int \frac{L(\theta)}{f(\theta)}\, f(\theta)\pi(\theta)\, d\theta \tag{33}$$

using f as an importance factor. This will, of course, be the same evidence value as before, to within the numerical uncertainty. However, the amount of uncertainty will change. If we are clever enough to factor part of the likelihood into the prior, and still manage to normalize that and sample from it, we would thereby start closer to the posterior, with reduced mismatch H, 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.

3.5. *Exact sampling*

For some specialized 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 any 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.

3.6. *Ensembles*

Let nested sampling use N objects instead of just 1. It could be run, as suggested, with each individual $L(\theta_1), \dots, L(\theta_N)$ subject to the common constraint $L_{\min}$. This is illustrated in Figure 8a, using only 2 objects for clarity in a 2-dimensional diagram. In successive iterates, these two objects are constrained within successively small squares of prior mass (regardless of the associated likelihood values).

Alternatively, nested sampling could be implemented with the N objects considered as independent components of a single ensemble with joint likelihood

$$\mathcal{L}(\theta_1, \theta_2, \dots, \theta_N) = L(\theta_1)L(\theta_2)\dots L(\theta_N) \tag{34}$$

subject to the single constraint $\mathcal{L} > \mathcal{L}_{\min}$. But there's a subtle difference, because the constraint on $\mathcal{L}$ will not — indeed can not — yield a simple square. For example, Figure 8b illustrates successive constraints for the likelihood function $L(X) = X^{-1/2}$, for which the curves happen to be hyperbolas. As iterations procced, one of the objects becomes much more constrained than the other, and the dichotomy increases with N, and with iteration. Obviously, exploration may be harder within such awkward shapes, whose details depend on the likelihood function, so the alternative implementation is likely to be inferior.

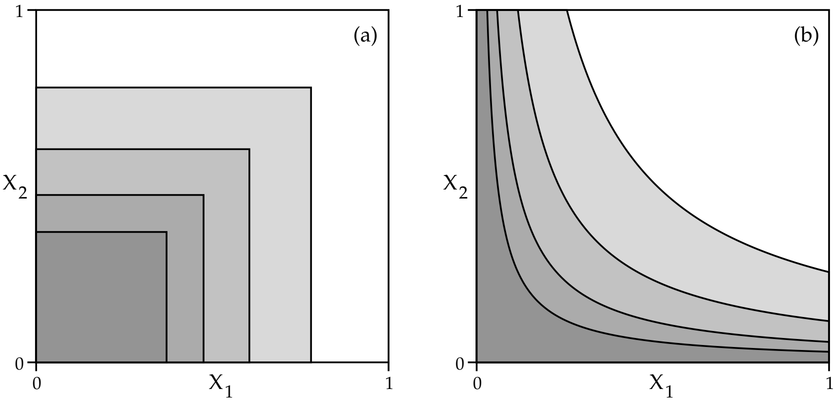

Figure 8: *Nested sampling under (a) separate constraints, (b) joint constraint.*

Traditional exploration imposes likelihood factors L^β on individual objects. This is exactly equivalent to imposing a single factor $\mathcal{L}^\beta$ on the ensemble as a whole. Yet the difference between imposing a temperature $1/\beta$, and imposing a single hard constraint, on $\mathcal{L}$ is usually only 1 part in N. In statistical thermodynamics, the former is called a canonical ensemble whilst the latter is called a microcanonical ensemble, and the two are used interchangeably. So the traditional "thermal" methodology corresponds to an *inferior* implementation of nested sampling.

3.7. *Density of states*

The density of states (being the prior mass in a thin likelihood shell — loosely, its surface 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 in (28) above

$$g^*(L) = -\frac{d\log X}{d\log L} \tag{35}$$

Differencing across r steps gives

$$g^*(\overline{L}) = -\frac{\log X_i - \log X_{i-r}}{\log L_i - \log L_{i-r}} = \frac{-\log t_i - \log t_{i-1} - \ldots - \log t_{i-r+1}}{\log L_i - \log L_{i-r}} \tag{36}$$

for $\overline{L}$ somewhere between L_{i-r} and L_i. The statistics (16) of each $\log t$ are known, and independent, so that in terms of mean and standard deviation

$$g^* = \frac{(r \pm \sqrt{r})/N}{\log L_i - \log L_{i-r}} \tag{37}$$

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^* locally 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

$$Z(\beta) = \int_0^1 L^\beta \, dX \tag{38}$$

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 non-thermal, but can simulate any temperature. The simulation relationship (38) can be written as a Laplace transform.

$$Z(\beta) = \int g(E)\, e^{-\beta E}\, dE \tag{39}$$

Although it is sometimes possible to invert Laplace transforms analytically, numerical inversion is badly conditioned. Basically, Laplace transforms all look so similar that their progenitors can't easily be distinguished. Z can be derived from g, but not the other way round.

This means that nested sampling's density of states is *more fundamental* than annealing's thermal properties. Likewise in physics, thermodynamics is built upon quantum states, not the other way round.

3.8. *Multiple phases*

Suppose, contrary to Figure 7 above, that the logarithmic density of states g^* is not an increasing function of $\log X$, so that $L^\beta X$ is not concave (Figure 9).

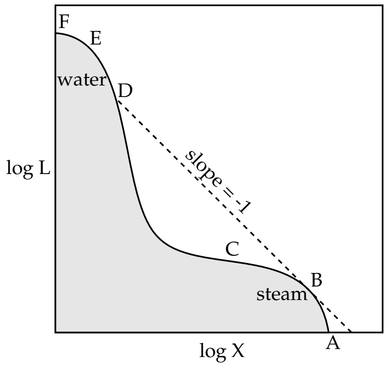

Figure 9: *Annealing has difficulty with convex likelihood.*

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, at B or E 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 also want objects 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 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. Phase change problems in general are 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.

If there were three phases instead of just two, annealing might fail even more spectacularly. It would be quite possible (Figure 10) for steam supercooled to B to condense directly to cold ice at E, and superheated ice at D to sublime directly to hot steam back at A, without settling in an intermediate water phase at all. The dominant phase could be lost in the hysteresis, and inaccessible to annealing.

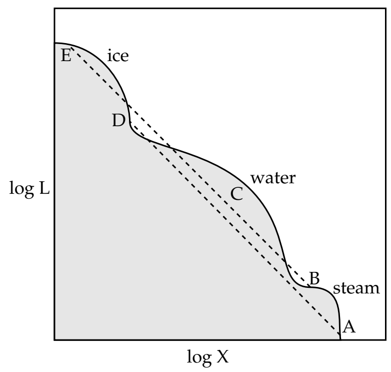

Figure 10: *Dominant "water" phase is lost in hysteresis loop ABED. The desired region BCD can be missed by annealing.*

Nested objects, though, 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.

4. COPYING

Suggested but not forced by nested sampling, the "copy" operation of replacing a rejected object by a duplicated favoured one is useful of itself, particularly when the likelihood is multi-modal. At each iterate, the worst (outermost) object is discarded in favour of a copy of an internal survivor, which is then re-evolved.

The likelihood function need not have the convenient single maximum shown in Figure 4. Consider intead a bi-modal likelihood, having two maxima, such as that shown in Figure 11. One mode is *dominant* because it contains the bulk of the evidence $\int L\,dX$: the other is *subordinate.* There is a critical likelihood *gate* below which the modes are connected, and above which they are separate. Before the gate is reached, MCMC exploration can presumably diffuse freely around the volume enclosing both modes. After the gate is passed, transitions between modes need to jump across to the other domain, which soon becomes hard to find so that transitions are blocked and an object can diffuse only within its own mode.

At the critical likelihood, let the accessible volumes be X_1 for the dominant mode and X_2 for the subordinate. The chance of an exploratory object falling into the dominant mode as the gate closes behind it is the proportional gate width

$$W = \frac{X_1}{X_1 + X_2}\,. \tag{40}$$

Conversely, with chance $1 - W$, it falls into the subordinate mode, where it is essentially trapped. With N independent objects, the chance of "success" with at

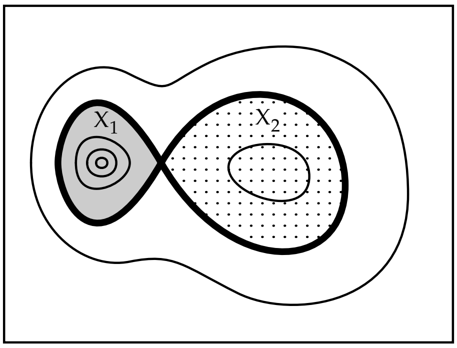

Figure 11: *The thick likelihood contour around the dominant mode X_1 is a "gate". The subordinate mode X_2 is surrounded by a "trap".*

least one object in the dominant mode is

$$\Pr(\text{success} \,|\, N \text{ objects}) = 1 - (1 - W)^N . \tag{41}$$

Basically, we need rather more than W^{-1} objects to be reasonably sure of at least one success. Thus, if the gate width is $W = 1/64$ but we only supply $N = 10$ objects, then the chance of a success is less than 1/6.

It may be that a particular multi-modal problem has just one narrow gate. Eventually, the likelihood will favour the dominant mode by a factor that more than compensates for the narrow opening, but that's not known as the gate is passed, and we don't want the expense of retreating back to the gate afterwards even if we had reason and knew how far to retreat.

It is perhaps more likely that a complicated problem has several gates, perhaps six gates of width 1/2 or so at different likelihood levels, instead of just one of width 1/64. After all, there is a bigger parameter-space associated with this more general framework. If objects are programmed to explore independently, half will fail at the first gate, then half the survivors will fail at the second and so on until the final survival rate is only 1/64, just as for a single narrow gate; this is shown schematically in Figure 12.

With copying, though, the object with lowest (worst) likelihood is eliminated in favour of a copy of one of the (better) others. After a gate is passed and the likelihood constraint continues to climb to more-restrictive heights, the subordinate mode should become progressively less populated. Indeed, after the nested-sampling constraint has climbed above the subordinate maximum, that mode can have no surviving objects at all. So, even if a gate is quite narrow, the dominant mode becomes re-populated provided at least one object manages to find it.

With a gate width of 1/2, the chance of having at least one success from N objects is $1 - 2^{-N}$, nearly certain. Before the next gate is reached, it may well be that the population along the dominant route has increased from the original $N/2$ back up to or nearly to N. After 6 such gates, with a mere 2^{-N} chance of

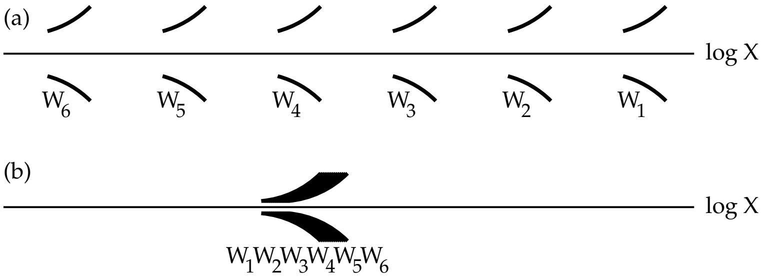

Figure 12: *(a) Several wide gates, and (b) one narrow gate of equivalent aperture.*

failure at each, the chance of success is $(1-2^{-N})^6$. Not only is this more than 99% for $N = 10$, but quite soon afterwards most or all of the objects should be in the dominant mode. Nested sampling's "copy" operation has turned an expectation of total failure into a a high probability of complete success.

Generally, with a series of well-separated gates of widths W_g, the expectations of success are

$$\Pr(\text{success} \mid N \text{ objects}) = \begin{cases} 1 - \left(1 - \prod_g W_g\right)^N & \text{for no copying,} \\ \prod_g \left(1 - (1 - W_g)^N\right) & \text{with copying.} \end{cases} \tag{42}$$

Basically, the number of objects needed to give a good chance of success is

$$N_{\text{minimum}} \approx \begin{cases} \prod W^{-1} & \text{for no copying,} \\ (W_{\max})^{-1} & \text{with copying.} \end{cases} \tag{43}$$

and copying always beats exploration by individually-preserved objects.

5. EXAMPLES

5.1. *Gaussians*

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

$$L(\theta) = 100 \prod_{i=1}^{20} \frac{1}{\sqrt{2\pi}u} \exp\left(-\frac{\theta_i^2}{2u^2}\right) + \prod_{i=1}^{20} \frac{1}{\sqrt{2\pi}v} \exp\left(-\frac{\theta_i^2}{2v^2}\right) \tag{44}$$

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 restricted to $\beta \leq 1$ needs roughly a billion (e^{20}) trials to find the spike, and several times e^{25} to equilibrate properly. On the

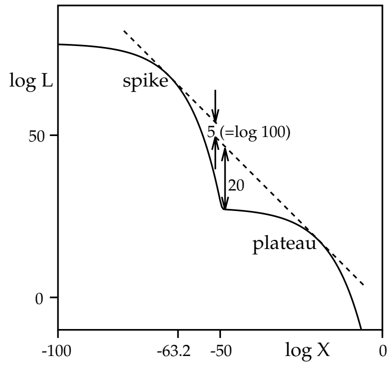

Figure 13: *Gaussian spike on plateau. The spike is favoured by a Bayes factor exp(5), but annealing needs exp(20) trials to find it.*

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.

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

$$\log Z \approx \log(101) \pm \sqrt{63.2/N} \tag{45}$$

(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 something like 1600 evaluations, though this remains comfortably less than e^{25}.

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

$$L(\theta) = 100 \prod_{i=1}^{20} \frac{1}{\sqrt{2\pi}u} \exp\left(-\frac{(\theta_i - 0.2)^2}{2u^2}\right) + \prod_{i=1}^{20} \frac{1}{\sqrt{2\pi}v} \exp\left(-\frac{\theta_i^2}{2v^2}\right) \tag{46}$$

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 trials 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.

Of course, this particular problem is easily soluble by splitting it into its constituent parts and evaluating the two evidence values separately. With evidence values being so easy to compute, it's better to split problems (where possible) than to attempt joint computation. Splitting avoids the dangerous mismatch between initial gate widths and final Bayes factors: it's better to avoid a problem than to have to solve it. More generally, it's preferable to do separate calculations when trying to compare models of different types (as in cosmology: Mukherjee, Parkinson and Liddle (2006)).

5.2. *Data analysis*

It is not just large problems with awkward likelihood functions that exhibit phase changes. In data analysis, it frequently happens that there is an initial "it's all just noise" phase to be overcome before the true interpretation of the signals emerges.

As trivial illustration, 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 (Figure 14)

$$L(\theta) = 0.99\,\frac{2q^2}{(\theta+q)^3} + 0.01 \quad \text{with, say, } q = 10^{-9}. \tag{47}$$

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

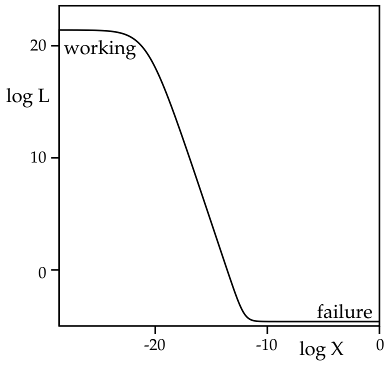

Figure 14: *The "working" phase on the left is hard to find by annealing.*

An interpretation of (47) is that the experiment was anticipated to work with 99% reliability. If it worked, the likelihood $L = 2q^2/(\theta+q)^3$ 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 about a billion trials, and will wrongly suggest "failure", with $Z = 0.01$. Only if β is increased far beyond 1 to a million or so would the ensemble be likely to find the working mode in fewer than thousands of trials. Even then, the samples would crash inward and have to be annealed back out through that factor of a million. And the evidence value would have been lost.

For nested sampling, which steadily tracks $\log X$ instead of trying to use the slope, such 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 objects. 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.

5.3. *Potts model*

The Potts (or Ising) model is a two-dimensional rectangular model of atoms of two changeable "colours" (A and B). The loglikelihood counts the number of bonds between atoms of the same colour (A-A or B-B), with a coefficient set to the critical value.

$$\log L = \log(1+\surd 2) \times (\text{number of A-A } + \text{ number of B-B}) \qquad (48)$$

This is always concave, but a wide swathe of it becomes almost straight as the number of atoms increases. Figure 15 plots it for a coarse (18 × 18) grid, and much of the curve is already nearly straight. Supercooling is never needed, but a tiny change in temperature moves the bulk of the posterior from one end of the straight section to the other. Because there is no actual convexity, this is an example of a "second order" phase transition. It is difficult to anneal it properly, but nested sampling moves steadily inward towards the fully-ordered states without any difficulty, exploring intermediate partially-ordered states on the way.

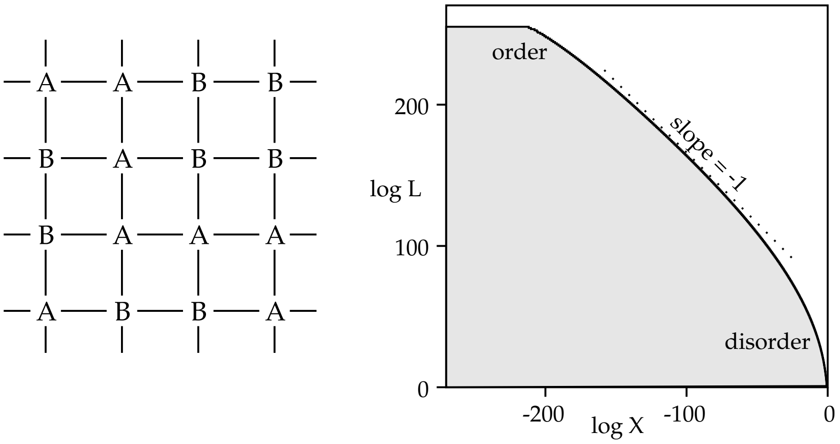

Figure 15: *Potts model (left), on* 18 × 18 *grid (right). The transition region between disorder and order is almost straight.*

Murray *et al.* (2006) have programmed this on a 256×256 grid using the Fortuin-Kasteleyn-Swendsen-Wang exploration strategy (Edwards and Sokal (1988)), and found the anticipated steady increase in order.

6. CONCLUSIONS

Nested sampling is a new algorithm which reverses the traditional approach to Bayesian computation by putting the *evidence* (a.k.a. marginal likelihood, prior predictive) first. A conventional collection of posterior samples can be acquired as the calculation proceeds, but that is an optional extra.

The algorithm proceeds by systematically constraining the available prior mass, shrinking it geometrically under successively tighter lower bounds on likelihood. The evolution path thus depends only on the *shapes* of the likelihood contours, and is independent of the likelihood *values.* This invariance enables nested sampling to deal with convex likelihood functions, which define a class of problems that is effectively denied to standard annealing. Nested sampling won't solve everything, because general multi-modality is and will remain difficult, but it promises to solve more.

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. Usually this will be accomplished by MCMC, where the hard constraint happens to give a similar restriction on step-length to that applying in standard Metropolis-Hastings evolution. Ancillary techniques such as importance sampling and slice sampling transfer straightforwardly to nested sampling. Hence, in those cases where annealing works and has an efficient schedule, the new method should offer no great gain or loss of computational speed. Even so, nested sampling is more fundamental in that it gives a direct view of the underlying density of states $g^*(L)$ as it steps steadily inward. It is not thermal, but can simulate thermal properties at any temperature.

We do not address the errors that would arise from imperfect sampling of the prior 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 well-founded probabilistic estimates, not derived from any frequentist fixup. As usual, uncertainty diminishes as $\sqrt{N}$, where N measures the amount of computation allowed, here quantified as the number of objects being evolved. In short, nested sampling follows the rules of probability calculus, as an algorithm for Bayesian computation should do.

Compared with traditional methods, there seems to be no disadvantage to using nested sampling, but there is demonstrable advantage in range of application, and in its straightforward specification of uncertainty. It's simple, and it's general.

7. ACKNOWLEDGMENTS

I wish to thank David MacKay and the Inference group journal club at Cambridge for valuable encouragement and advice, and for hosting nested-sampling papers and programs (currently at www.inference.phy.cam.ac.uk/bayesys). This work was supported by Maximum Entropy Data Consultants Ltd of Cambridge (England).

REFERENCES

Edwards, R. G. and Sokal, A. D. (1988). Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and the Monte Carlo algorithm. *Phys. Rev. D* **38**, 2009–2012.

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

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **5**7, 97–109.

MacKay, D. J. C. (2003). Information Theory, Inference, and Learning Algorithms (page 379). Cambridge: University Press.

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

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

Mukherjee, P., Parkinson, D. R. and Liddle, A. R. (2006). A nested sampling algorithm for cosmological model selection. *Astrophys. J.*, **638**, L51–L54.

Murray, I., MacKay, D. J. C., Gharahmani, Z. and Skilling, J. (2006). Nested sampling for Potts models. *Advances in Neural Information Processing Systems*, **18** (to appear).

Neal, R. (2003). Slice sampling. *Ann. Statist.* **3**1, 705–767.

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.

Sivia, D. S. and Skilling, J. (2006). *Data Analysis: a Bayesian tutorial* (2nd ed.) Oxford: University Press.

Skilling, J. (2004). Nested sampling. *Amer. Inst. Phys. Conference Proc.* **735**, 395–405.

Skilling, J. (2006). Nested Sampling for General Bayesian Computation. *Bayesian Analysis* **1**, (to appear).

APPENDIX

```
// NESTED SAMPLING MAIN PROGRAM IN 'C' by John Skilling, Aug 2005
//(GNU General Public License software (C) Sivia & Skilling 2006)
//================================================================
#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 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 +++++++++++++++++++++++++++++++
#define n ...      // number of objects
#define MAX ...    // max number of iterates
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
void Results(Object* Samples, int nest, double logZ){...}
             // optional list of samples of weight exp(logWt-logZ)
---------------------------------------------------- UP TO HERE */
```

```
int main(void)
{
   Object Obj[n];         // Collection of n objects
   Object Samples[MAX];   // Objects defining posterior
   double logwidth;       // ln(width in prior mass)
   double logLstar;       // ln(Likelihood constraint)
   double H    = 0.0;     // Information, initially 0
   double logZ =-DBL_MAX; // 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
// Set prior objects
   for( i = 0; i < n; i++ )
       Prior( &Obj[i] );
// Outermost interval of prior mass
   logwidth = log(1.0 - exp(-1.0 / N));
// NESTED SAMPLING LOOP ++++++++++++++++++++++++++++++++++++++++++
   for( nest = 0; nest < MAX; 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 = logwidth + 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)
       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
       logwidth -= 1.0 / N;
   } // --- NESTED SAMPLING LOOP (might be ok to terminate early)
   printf("# iterates = %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));
   Results(Samples, nest, logZ);     // optional
   return 0;
}
```