Nested Sampling's Convergence

John Skilling
Maximum Entropy Data Consultants Ltd.
Killaha East, Kenmare, County Kerry, Ireland
skilling@eircom.net — November 21, 2007, revised June 11, 2008

Abstract. Nested sampling is an algorithm for computing the Bayesian *evidence* (or prior predictive, marginal likelihood), through a probabilistic method having stated and controllable accuracy. This paper presents the method more clearly than before, and includes convergence proofs that were previously missing.

Keywords: nested sampling, evidence, convergence.

1 Introduction

The basic Bayesian task is to compute the value of *evidence* $\mathrm{pr}(D)$ from a prior $\pi(\theta)$ representing $\mathrm{pr}(\theta)$, and a likelihood $L(\theta)$ representing $\mathrm{pr}(D \mid \theta)$, θ being unknown parameters and D being known data.

$$Z = \mathrm{pr}(D) = \int \mathrm{pr}(D,\theta)\,\mathrm{d}\theta = \int \mathrm{pr}(D \mid \theta)\,\mathrm{pr}(\theta)\,\mathrm{d}\theta = \int L(\theta)\pi(\theta)\,\mathrm{d}\theta$$

This induces the posterior $\mathrm{pr}(\theta \mid D) = \mathrm{pr}(D,\theta)/\mathrm{pr}(D)$, written as

$$P(\theta) = L(\theta)\pi(\theta)/Z$$

which in turn induces the statistics (mean, variance, etc) of θ and its properties. These properties, incidentally, include the prior-to-posterior compression ratio

$$H = \int P(\theta) \log \frac{P(\theta)}{\pi(\theta)}\,\mathrm{d}\theta$$

also known as the *information*. Z has dimensions of inverse data. H is dimensionless, it scales with the size of the problem, and is measured in nats where 1 nat $= \log_2 \mathrm{e}$ bits.

Nested sampling is an algorithm for these calculations, in which information plays a central part. It is simple enough to be straightforwardly accessible at science-undergraduate level. Indeed, implementation could realistically form an informative and educational entry-level statistics project, needing only a flat prior and no awkward Metropolis-Hastings thermal balancing.

This paper starts with the basic method, involving nothing more complicated than the ideas of sorting and sampling. Because our exploration of θ is perforce limited, we are necessarily faced with a problem of inference from incomplete data. Nested sampling faces this problem squarely, and produces a probabilistic result giving the range of what Z might be, given the limited computational effort. This is illustrated with a worked example, using the simplest possible form of likelihood function. It is straightforward to develop the algorithm to improve accuracy. The explanation of that is followed by a proof of mean-square convergence to the correct value of $\log Z$. Finally, a minor simplification is introduced to bypass some of the algorithmic overhead.

Nested sampling was originally introduced in earlier papers Skilling (2006, 2007), but this account has been polished through experience and passage of time, and includes mathematical detail requested by the community. For specific applications, the reader is referred to those papers, and to Sivia and Skilling (2006) where short computer code is published.

2 Nested Sampling

To understand nested sampling, it's simplest to remember that all practical inference is finite, if only because it fits into a computer with a finite number of states. We may use continuum notation as a convenience, but in inference we are always trying to do finite sums. Suppose the parameters θ are coded to occupy ν bits of computer memory. We can think of ordering the 2^ν states by decreasing likelihood value, accumulating prior mass as we do so. To ensure unambiguous ordering when L values coincide, we can append a "key" $\kappa(\theta)$ to the computer's binary coding of L, chosen so that key values don't repeat and all L's become different. This ordering defines a strictly decreasing function

$$L(X) = \text{“enclosing likelihood” function of prior mass } X,$$

with its strictly decreasing inverse

$$X(L) = \int_{L(\theta)>L} \pi(\theta)\,\mathrm{d}\theta = \text{“enclosed prior” of likelihood } L.$$

Note that I adopt the computer-science practice of avoiding symbol proliferation by overloading function names according to argument type: there is no confusion between $L(\theta)$ and $L(X)$ because their argument types differ.

Our task is to find the value of what is now a one-dimensional sum along the states, ordered-by-likelihood as they now are, over unit prior mass.

$$Z = \int_0^1 L(X)\,\mathrm{d}X\,, \qquad (H + \log Z)Z = \int_0^1 L(X)\,\log L(X)\,\mathrm{d}X\,. \tag{1}$$

But the implied summation over 2^ν states is impractical for any sizable application. We are forced to select only a limited number of θ for evaluation, from which we must *infer* the evidence and consequential properties. Let these states be θ_i $(i = 1, 2, \ldots, n)$, with $L_i = L(\theta_i)$ and $X_i = X(L_i)$, and let them be ordered as

$$0 < X_n < \ldots < X_2 < X_1 < X_0 = 1$$

Because L is decreasing over X, its value in any interval is constrained by the end-points, $L_{i+1} \geq L(X) \geq L_i$ in $X_{i+1} \leq X \leq X_i$. Hence Z is constrained within

$$\sum_{i=1}^{n}(L_i - L_{i-1})X_i \leq Z \leq \sum_{i=1}^{n}(L_i - L_{i-1})X_{i-1} + (L_{\max} - L_n)X_n$$

where $L_0 = 0$ and $L_{\max}$ is an upper bound, if one can be found. It's convenient to separate off the last right-hand term as a $O(X_n)$ truncation eror, and consider the remaining range

$$\Delta Z = \sum_{i=1}^{n}(L_i - L_{i-1})(X_{i-1} - X_i)$$

as a $O(n^{-1})$ quadrature error.

If, in practice, we could assign X and compute its $L(X)$, the evaluation of Z would reduce to the above elementary numerical analysis. But we can't. What we *can* do is assign X *at random*, as follows. Each X_i lies beneath its predecessor X_{i-1}, so its likelihood must exceed L_{i-1}. Taking this as a restriction $L(\theta) > L_{i-1}$ on available θ, take a random sample θ_i from the prior π, under that restriction.

$$\Big\| \qquad Pr(\theta_i) \propto \pi(\theta_i) \text{ in } L(\theta_i) > L_{i-1} \qquad \Big\|$$

This construction ensures that X_i is uniformly distributed in $(0, X_{i-1})$. If we could actually do the $\theta \to X$ ordering, we would know exactly what X_i was, but we don't, so we are stuck with this partial knowledge:

$$X_i = t_i X_{i-1}, \text{ where } t_i \sim \text{Uniform}(0,1).$$

However, we do know $L_i = L(\theta_i)$ definitively.

The upshot is that we have a method which generates a correctly-ordered sequence of (L, X) pairs, with L known definitively and X known statistically. This suffices to infer Z statistically. The $O(n^{-1/2})$ statistical uncertainty will dominate the $O(n^{-1})$ quadrature error and the usually-exponentially-small truncation error. This means that it doesn't matter which quadrature formula we use. The lower limit

$$\widehat{Z} = \sum_{i=1}^{n} (L_i - L_{i-1}) X_i \tag{2}$$

is simplest, and avoids the questionable upper bound $L_{\max}$.

The evaluation of evidence now becomes an exercise in probability, and not numerical analysis. Most simply, generate a sequence of t's from Uniform(0,1), and use them to determine estimates $\widehat{X}$, and hence $\widehat{Z}$. Do this a respectable number of times, perhaps 50, to get enough samples of $\widehat{Z}$ to firm up its probability distribution, to be presented as the result $\text{pr}(\widehat{Z})$. Importantly, this includes the procedure's numerical uncertainty.

3 Worked example

Let

$$L(\theta) = \begin{cases} 1 & \text{over a fraction } X^* = \exp(-\xi^*) \text{ of prior mass (* flags the truth)} \\ 0 & \text{over the remaining fraction} \end{cases}$$

From (1), the true evidence is $Z = X^*$ and the true information is $H = \xi^*$. In order to illustrate the behaviour of nested sampling in large applications where the bulk of the posterior covers only an exponentially small fraction of prior mass, think of ξ^* as reasonably large, say a million. As nested sampling proceeds inwards, its exploratory "particle" is almost certain to start at an "outside" location θ with likelihood values of zero, and eventually discover the "inside" with unit values, through a sequence $(0, 0, \ldots, 0, 1, 1, 1, \ldots)$ having $k - 1$ zeros followed by an indefinite remainder of ones.

The unknown underlying parameters in nested sampling are the uniform t's, or equivalently their exponentially distributed logarithms $\tau = -\log t$,

$$\text{pr}(\tau) = \exp(-\tau) \tag{3}$$

After j steps, the probability of its particle reaching $X = \mathrm{e}^{-\xi}$ is Gamma;

$$\mathrm{pr}(\xi \mid j) = \frac{\xi^{j-1}e^{-\xi}}{(j-1)!}$$

It follows that the critical step k, at which $\sum \tau_k$ first exceeds ξ^* and unit L is discovered, is distributed as Poisson;

$$\mathrm{pr}(k \mid \xi^*) = \frac{(\xi^*)^{k-1}e^{-\xi^*}}{(k-1)!}$$

The data sequence has been produced with this degree of variability, basically $k = \xi^* \pm \sqrt{\xi^*}$. We now have to infer $\widehat{Z}$.

Given a sample $\tau_1, \tau_2, \ldots$ from the prior, the value of Z would be given by numerical quadrature (2) as a defined property of the τ's, here

$$\zeta = -\log \widehat{Z} = -\log \widehat{X}_k = \tau_1 + \tau_2 + \ldots + \tau_k \tag{4}$$

Observation of likelihood values does not, unfortunately, influence our knowledge of the τ's. To us limited beings who can't do the sort, that knowledge remains described by (3) because we know no better. Hence our distribution of ζ is Gamma;

$$\mathrm{pr}(\zeta \mid k) = \frac{\zeta^{k-1}\mathrm{e}^{-\zeta}}{(k-1)!}$$

Observation of likelihood values *does* define their property (4) representing evidence. Basically, our inference here about ζ, hence $\widehat{Z}$, is

$$\log \widehat{Z} = -k \pm \sqrt{k}$$

With ξ large, the uncertainty in $\log \widehat{Z}$, which is close to $\pm\sqrt{H}$, spans many orders of magnitude in $\widehat{Z}$, correctly mirroring the variability in k from run to run.

This behaviour is just what is wanted. In general applications, a prior-to-posterior compression ratio $\exp(-H) = \exp(-1000000)$ is quite likely to have come from using a dataset with a million data, involving log-likelihoods of the same order. In terms of chisquared, $\chi^2 \approx 1000000 \pm 1000$, where the range reflects random noise in the data. It may well be that Bayes factors closer than $\exp(1000)$ don't matter much, because the experiment was constructed to make visible the difference between alternative hypotheses. Conveniently, this is just the level of accuracy naturally attained by nested sampling. Moreover, nested sampling tells us what the numerical uncertainty is, because we get a probabilistic inference $\mathrm{pr}(Z)$, and not just some single number. Such quality is new. And, if we need to evaluate the evidence more precisely, we can do so.

4 More accuracy

Runs of nested sampling, with different random seeds, each yield a sequence of likelihood values. As above, each such sequence can be folded with random sequences of t's to yield an estimate $\widehat{Z}$, and N such results can be combined to reduce the statistical uncertainty on $\log \widehat{Z}$ by the usual $N^{-1/2}$. However, that's not optimal. Properly formulated inference uses all the data at once, not relying on ad hoc averaging of partial results.

To infer the evidence properly, merge the likelihood sequences into a single (ordered) sequence of more-closely spaced likelihoods. Then, each inward step involves a N-fold choice of successor particle, all of which are uniformly distributed (over π) within the current constraint, and none of which are intrinsically preferred. So the merged sequence has compression factors t distributed as Beta

$$\mathrm{pr}(t) = N t^{N-1} \tag{5}$$

that being the distribution of the outermost of N candidates, each uniformly distributed on (0,1). There is no longer any need to keep the sequences separate. All N particles can be kept within a single run. At each step, the outermost (lowest L) particle is selected as θ_i, before being discarded in favour of a new one from within $L > L_i$. That new particle could be developed from any of the survivors, and evolved within the constraint to "forget" its origin.

The cost is an extra factor N in computation. The benefit is not just less uncertainty, but greater robustness in case it's difficult to re-sample faithfully from within the required constraint. If a particle "gets stuck", it will soon disappear off the edge provided at least some of the others are able to progress properly, so that a single mistake need not destroy a whole sequence.

5 Convergence

5.1 Strategy

We aim to prove convergence of $\log \widehat{Z}$ in mean square. This requires only a double boundedness assumption

$$Z < \infty \qquad \text{and} \qquad H < \infty\,. \tag{6}$$

Requiring a bound on Z is obvious because results are never infinitely plausible. Requiring a bound on H is also obvious because we never learn infinite information. As for the value of the bound on H, the number of bits in the data-compressed file of data D might be a good place to start. We use N particles and a rather larger number n of iterates. The desired convergence limit is $n \to \infty$, with N chosen to make efficient use of the limited resources n. We start by dividing the $(0, 1)$ range of X at some $X^* = \exp(-\xi^*/N)$, so that

$$Z = Z^- + Z^+\,, \qquad Z^- = \int_0^{X^*} L\,\mathrm{d}X\,, \quad Z^+ = \int_{X^*}^1 L\,\mathrm{d}X\,.$$

Z^+ will be the target of nested sampling, with statistical uncertainty; Z^- the truncation error with bounded uncertainty.

5.2 Statistical uncertainty

Taking the target first, we can write

$$Z^+ = \int_0^1 L^+\,\mathrm{d}X \ \text{ where } L^+(X) = \begin{cases} L(X) & \text{in } X > X^*, \\ 0 & \text{otherwise.} \end{cases}$$

Define $\tau = -N \log t$, so that (from (5) and as in (3)) τ is exponential at each step

$$\mathrm{pr}(\tau) = \mathrm{e}^{-\tau} \tag{7}$$

After n steps, $\xi = \tau_1 + \tau_2 + \ldots \tau_n$ is Gamma distributed with $\mathrm{pr}(\xi) = \xi^{n-1}\mathrm{e}^{-\xi}/(n-1)!$, so the probability of failing to reach ξ^* and cover the support range of L^+ is

$$\mathrm{pr}(\mathrm{fail} \mid n) = \int_{\xi^*}^{\infty} \frac{\xi^{n-1}\mathrm{e}^{-n}}{(n-1)!}\, \mathrm{d}\xi$$

By taking at least twice as many iterates ($n \geq 2\xi^*$) as are, on average, necessary to reach ξ^*, this probability of failing to cover all of L^+ is exponentially small, less than 2^{-n}. Because we aim to prove convergence in the limit as the cost $n \to \infty$, we nelect this trivial probability of failure, and take the n iterates to cover the whole of L^+. Iterates beyond the boundary ξ^*, though, have no effect on the target integral Z^+, because the integrand L^+ is zero there. Hence we can evaluate Z^+ is if n were ∞.

The number k of iterates needed to pass a given value $X = \mathrm{e}^{-\xi/N}$, leaving $k-1$ failed steps behind, is

$$\mathrm{pr}(k \mid \xi) = \frac{\xi^{k-1}\mathrm{e}^{-\xi}}{(k-1)!} \tag{8}$$

Having completed a run, the sequence of likelihood values L_k is available. The corresponding X_k are not available, so have to be inferred as $\widehat{X}_k = \exp(-\zeta_k/N)$ from their distribution. By sampling what the τ's might have been, according to their known distribution (7), the k'th value $\zeta_k = \tau_1 + \tau_2 + \ldots + \tau_k$ is Gamma distributed as

$$\mathrm{pr}(\zeta \mid k) = \frac{\zeta^{k-1}\mathrm{e}^{-\zeta}}{(k-1)!} \tag{9}$$

Putting (8) and (9) together, source location ξ yields recovered location ζ according to

$$\mathrm{pr}(\zeta \mid \xi) = \sum_{k=1}^{\infty} \frac{\zeta^{k-1}\mathrm{e}^{-\zeta}}{(k-1)!}\,\frac{\xi^{k-1}\mathrm{e}^{-\xi}}{(k-1)!}$$

The deviation $\delta = \zeta - \xi$ has mean square

$$\mathrm{E}(\delta^2) = \int_0^{\infty} d\zeta\, (\zeta-\xi)^2 \mathrm{pr}(\zeta \mid \xi) = \sum_{k=1}^{\infty} \frac{\xi^{k-1}\mathrm{e}^{-\xi}}{(k-1)!}\big(k(k+1) - 2k\xi + \xi^2\big) = 2(\xi+1)$$

Next, we use δ to compare Z^+, in its alternative integrated-by-parts form, with its estimate $\widehat{Z^+}$,

$$Z^+ = \int X\, \mathrm{d}L^+ = \int \mathrm{e}^{-\xi/N}\, \mathrm{d}L^+\,, \qquad \widehat{Z^+} = \int \widehat{X}\, \mathrm{d}L^+ = \int \mathrm{e}^{-\zeta/N}\, \mathrm{d}L^+\,.$$

To prove convergence of $\log \widehat{Z^+}$, we proceed to relax the difference. For a given amount of variation, the deviation of $\widehat{Z^+}$ from Z^+ is maximal when δ is fully correlated over L, all perturbations moving up or down together. Further relax the error magnitude to the maximum value δ_n, which occurs at termination step n, giving

$$\mathrm{E}(\delta^2) \;<\; \mathrm{E}(\delta_n^2) = 2(\xi_n + 1)$$

Now ξ_n is large, being Gamma distributed with mean n and variance n. The probability of it lying beyond $2n$, or even $2n-1$, is exponentially small, less than 2^{-n}. We are already neglecting terms of that order, so

$$\mathrm{E}(\delta_n^2) < 4n$$

With that over-estimated deviation, the two integrals would become directly proportional, with $\widehat{Z^+}/Z^+ = \exp(-\delta_n/N)$, so that

$$\log \widehat{Z^+} - \log Z^+ = -\delta_n/N$$

Since the deviation was over-estimated, the statistical mean-square error is bounded as

$$\mathrm{E}\big[(\log \widehat{Z^+} - \log Z^+)^2\big] \; < \; \frac{\mathrm{E}(\delta_n^2)}{N^2} < \frac{4n}{N^2} \equiv \sigma_{\mathrm{stat}}^2 \tag{10}$$

The proof of convergence could stop here, with rms error diminishing as $O(N^{-1/2})$ provided the resources per particle (n/N) remain fixed and adequate to keep the truncation error small. (□)

5.3 Truncation error

However, we can eliminate the doubt about truncation error Z^- by using boundedness of H. Let the upper bound on H be $\mathcal{H}$. The likelihood function $L(X)$ can take arbitrary form, subject only to being non-increasing, and to the conditions (6) on its integrals Z and H. Under these restrictions, the worst possible (maximum) value of Z^- occurs when L is bi-valued;

$$L(x) = \begin{cases} aZ & \text{in } X < X_n, \\ bZ & \text{otherwise.} \end{cases} \tag{11}$$

Both below and above X^*, H is decreased by moving likelihood rightwards, giving more room to manoeuvre — but rightward shift is limited by the non-increasing requirement. With likelihood (11), a and b are determined by

$$X^*a + (1 - X^*)b = 1 \tag{12}$$

$$X^*a \log a + (1 - X^*)b \log b \le \mathcal{H} \tag{13}$$

Now (12) prohibits a and b being either both above 1 or both below 1, and L is non-increasing, so $b \le 1 \le a$. It follows that $-1 < b \log b \le 0$, and clearly $0 \le 1 - X^* \le 1$, so (13) gives

$$a \log a \le \frac{\mathcal{H}+1}{X^*} \equiv y$$

Here, a can never reach $2y/(\log y)$ without breaking the condition. This puts a hard bound on Z^-.

$$Z^- = X^*aZ < \frac{2(\mathcal{H}+1)Z}{\log(\mathcal{H}+1) - \log X^*} = \frac{2N(\mathcal{H}+1)Z}{N \log(\mathcal{H}+1) + \xi^*}$$

We are free to set the boundary ξ^* wherever convenient. The statistical uncertainty (10) held whenever $n \ge 2\xi^*$, so we now set $\xi^* = n/2$ to reach

$$Z^- < \frac{4N(\mathcal{H}+1)Z}{2N \log(\mathcal{H}+1) + n} < \frac{4N(\mathcal{H}+1)}{n} Z$$

Consequently,

$$Z^{+} = 1 - Z^{-} > 1 - \frac{4N(\mathcal{H}+1)}{n} Z$$

which, for sufficiently large n, implies a hard bound on the logarithm

$$-\frac{4N(\mathcal{H}+1)}{n}\log 2 \ < \ \log Z^{+} - \log Z \ < \ 0 \quad \text{provided}\, n > 8N(\mathcal{H}+1). \tag{14}$$

We need a hard bound here because we are now changing coordinates from Z to $\log Z$, and must not naïvely transfer soft mean-square constraints from one to the other. This worst case applies to estimates just as much as to the truth, so

$$-\frac{4N(\mathcal{H}+1)}{n}\log 2 \ < \ \log \widehat{Z^{+}} - \log \widehat{Z} \ < \ 0 \tag{15}$$

also. Putting (14) and (15) together gives

$$-\frac{4N(\mathcal{H}+1)}{n}\log 2 \ < \ (\log \widehat{Z^{+}} - \log Z^{+}) - (\log \widehat{Z} - \log Z) \ < \ +\frac{4N(\mathcal{H}+1)}{n}\log 2$$

which can now be relaxed to a soft mean-square constraint

$$\mathrm{E}\big[\big((\log \widehat{Z^{+}} - \log Z^{+}) - (\log \widehat{Z} - \log Z)\big)^2\big] \ < \ \left(\frac{4N(\mathcal{H}+1)}{n}\right)^2 \equiv \sigma_{\mathrm{trunc}}^2 \tag{16}$$

5.4 Completion

Combining statistical (10) and truncation (16) errors together gives

$$\mathrm{E}\big[(\log \widehat{Z} - \log Z)^2\big] \ \le \ (\sigma_{\mathrm{stat}} + \sigma_{\mathrm{trunc}})^2$$

so that the rms error is bounded as

$$\big\{\mathrm{E}\big[(\log \widehat{Z} - \log Z)^2\big]\big\}^{1/2} \ < \ \frac{2n^{1/2}}{N} + \frac{4N(\mathcal{H}+1)}{n} \tag{17}$$

As resources n increase, the number of iterates N can increase more slowly, allowing ever more accurate estimation of ever more perverse problems. In fact, the most perverse problem of all is that in which L stays constant through the entire run, until jumping unseen to a greater value immediately afterwards, as in (11). Even that problem is accessible with the aid of the constraint on H. Although the numerical coefficient in (14) is exaggerated by the relaxed nature of proof, the requirement $n > 8N(\mathcal{H}+1)$ ensures that there are enough iterates to cover most of the evidence, leaving a mean-square error no greater than 1/2. After that threshold is passed, the choice of N that guarantees least rms error (17) is

$$N = \left(\frac{n^3}{4(\mathcal{H}+1)^2}\right)^{1/4}$$

and the rms error itself is then bounded by

$$\big\{\mathrm{E}\big[(\log \widehat{Z} - \log Z)^2\big]\big\}^{1/2} \ < \ \frac{\big(32(\mathcal{H}+1)\big)^{1/2}}{n^{1/4}} \quad \text{provided} \quad n > 1024(\mathcal{H}+1)^2 \,.$$

Beneath this, lies a negligible quadrature error. The numerical coefficients are unnecessarily large, but adequate to prove the required convergence to the truth of nested-sampling estimates of $\log Z$. As $n \to \infty$, the rms error tends to zero, provided only that the true Z and H are bounded. □

For a fixed *problem*, and fixed resources per particle (n/N), nested sampling's asymptotic uncertainty is the $O(n^{-1/2})$ usual in statistics. However, more resources enable more difficult problems, and the $O(n^{-1/4})$ behaviour reflects a pessimistic view of the worst-case problem given fixed *resources*.

6 Log Z versus Z

There is a view in the community (Evans (2007); Chopin and Robert (2007)) that convergence should relate, not to $\log Z$, but to Z itself. That view is traditional but, actually, it is differences of $\log Z$ that matter, not differences of Z.

Several lines of argument confirm this. (1) Model selection involves Bayes factors, which relate to differences of logarithms, not of raw values. (2) Z is dimensional, having inverse data units. Hence raw differences ΔZ are not meaningful. They have to be divided by Z, as $\Delta Z/Z$, before they take meaningful dimensionless form. But then, why not just consider $\Delta \log Z$ in the first place? (3) In statistical mechanics, arguably our most developed calculus of large systems, the analogue of $\log Z$ is the partition function, from which other thermal properties such as entropy follow by differentiation. The partition function is an extensive variable, scaling with the size of the problem, say $d \sim 10^{24}$. A measurement might yield a range like $602300000000000000000000 \pm 5000000000000000000000$. Exponentiating that to investigate the underlying number of states is not helpful — the means and variances are bizarre.

For example, suppose we are to compare hypotheses A and B on the basis of kilo-dimensional data D. They are differently parameterised so that we can only reach their Bayes factor through separate calculations. We compute evidence estimates

$$\log \widehat{Z}_{\mathrm{A}} = -1500 \pm 40\,, \qquad \log \widehat{Z}_{\mathrm{B}} = -1000 \pm 30\,.$$

Which hypothesis is preferred? The answer, of course, is B, at 10-standard-deviation (500 ± 50) significance. Nevertheless, assuming normal distributions for the logarithms, the means and standard deviations for the naked values are

$$\widehat{Z}_{\mathrm{A}} = \mathrm{e}^{-700} \pm \mathrm{e}^{+100}\,, \qquad \widehat{Z}_{\mathrm{B}} = \mathrm{e}^{-550} \pm \mathrm{e}^{-100}\,.$$

Here, the difference of the means is so far below either of the standard deviations that there seems nothing to choose between A and B. On the other hand, if it was remembered that $\widehat{Z}$'s are positive, these numbers would suggest that A was preferred, on the grounds that the distribution of B is now confined to the leftmost e^{-200} fraction of the distribution of A. So considering moments of naked $\widehat{Z}$ values suggests either indifference or wrong choice, neither of which is helpful.

Differences of $\widehat{Z}$ *can* be used for comparison, but only when both uncertainties have fallen below 100%. In this example, that would involve a thousand-fold ($> 30^2$ or 40^2) computational penalty to improve the accuracies enough. Nevertheless, for the record only, it is not too difficult to follow Evans' lead and prove mean-square convergence of $\widehat{Z}$. A proof is given in the appendix.

7 Less accuracy

Proper, professional use of nested sampling involves sampling the compression factors t enough times to get samples of $\log \widehat{Z}$ and subsidiary properties, sufficient to determine their statistics — typically mean and standard deviation. Although it's relatively inexpensive, this sampling can often be by-passed when a large number of steps is involved. The number of steps required to reach prior mass $X = \mathrm{e}^{-\xi}/N$ is Poisson(ξ), basically $k = \xi \pm \sqrt{\xi}$. When this is large, the uncertainty drops away and the relationship approaches the definitive $X_k = \mathrm{e}^{-k/N}$. This can be used in (2) to get a cheap'n'cheerful approximate estimate

$$\widetilde{Z} = \sum_{k=1}^{n} (L_k - L_{k-1}) \mathrm{e}^{-k/N}$$

This formula, too, converges to the truth in the sense that $\mathrm{E}\big((\log \widetilde{Z} - \log Z)^2\big) \to 0$ as $N \to \infty$. Proof is as above, except that (9) is replaced by $\mathrm{pr}(\zeta \mid k) = \delta(\zeta - k)$. The definitive setting, though, has removed the variability: $\log \widetilde{Z}$ may be a good estimate, but we don't know how good. However, we have seen that the bulk of the evidence is usually found around compression ratio e^{-H}, which occurs after Poisson(NH) steps, basically $NH \pm \sqrt{NH}$. But the typical offset $\pm\sqrt{NH}$ in the number of steps induces a corresponding shift $\pm\sqrt{NH}/N$ in the estimate of $\log Z$. Hence the cheap'n'cheerful approximation, with uncertainty estimated too, is

$$\log Z = \log \widetilde{Z} \pm \sqrt{H/N}$$

Be reminded that this uncertainty estimate assumes that a single, local patch covers the bulk of the evidence. That does not hold if the application is exhibiting first or second order phase transitions.

8 Conclusion

The nested sampling algorithm does not just produce a single value for its target, the evidence. As a well-founded inference method, it also yields the uncertainty surrounding that value, consequent on the limited computational cost involved in any practical evaluation. Proof of convergence to the truth as the cost increases is supplied, and requires the existence of evidence Z and information H, nothing more.

The user is required to sample from the prior distribution *constrained* by likelihood, instead of the traditional modulation. Nested sampling itself imposes no restriction on the topology, continuity, differentiability, dimensionality, concavity or any other such property of θ, its prior or its likelihood. Neither does it make any suggestions about how difficulties encountered there may be overcome. However, the lack of restriction does permit a wider class of applications than is possible with traditional "thermal" algorithms Skilling (2006, 2007); Sivia and Skilling (2006).

The basic computational cost is that of H iterates. The cost of an iterate usually scales as the size d of the application, as does H, so the standard overall cost is $O(d^2)$.

Acknowledgments

This work was supported by Maximum Entropy Data Consultants Ltd of Cambridge (England). I also wish to thank David MacKay and Iain Murray for helpful advice.

Appendix: Formal convergence

The aim is to prove that nested sampling's estimates $\widehat{Z}$ converge to Z in mean square. Once again, proof requires a double boundedness assumption; here

$$\int L\,\mathrm{d}X = Z < \infty\,, \qquad \text{and} \qquad \int X^{-\lambda} L\,\mathrm{d}X = KZ < \infty \text{ for some } \lambda > 0. \tag{18}$$

As before, bounded Z is obvious. As $\lambda \to 0$, the second condition reverts to a constraint on H, but here λ has to be bounded away from 0, which is a stronger condition, less easy to justify.

In this presentation, we ignore truncation, which could be merged as above if required. Accordingly, we assume that we have indefinitely many iterations available. We can't use the previous "$\log Z$" proof because $\mathrm{E}(\delta^2)$ fails to constrain moments of $\mathrm{e}^{-\delta/N}$, which could involve large upward excursions. Consequently, our previous second-moment constraint on $\log \widehat{Z}$ fails to control moments of $\widehat{Z}$. Instead, the analysis now requires calculation of covariances rather than individual variability, so that we need to consider the recovery of source pairs $X = \mathrm{e}^{-\xi/N}$ by estimate pairs $\widehat{X} = \mathrm{e}^{-\zeta/N}$. The numbers of such steps needed to pass a first value ξ_1 followed by a second $\xi_2 > \xi_1$, is distributed as double Poisson

$$\mathrm{pr}(k_1, k_2 \mid \xi_1, \xi_2) = \frac{\xi_1^{k_1-1}}{(k_1-1)!}\,\frac{(\xi_2-\xi_1)^{k_2-k_1}}{(k_2-k_1)!}\,\mathrm{e}^{-\xi_2}$$

When simulating to get ζ_1 followed by ζ_2, the distribution is double Gamma

$$\mathrm{pr}(\zeta_1, \zeta_2 \mid k_1, k_2) = \frac{\zeta_1^{k_1-1}}{(k_1-1)!}\,\frac{(\zeta_2-\zeta_1)^{k_2-k_1-1}}{(k_2-k_1-1)!}\,\mathrm{e}^{-\zeta_2}\,, \qquad \text{with } \frac{x^{-1}}{(-1)!} = \delta(x)\,.$$

The Dirac delta function is explained by the possibility that a single step passes both source locations X_1 and X_2, in which case $k_2 = k_1$ which forces equality of estimates $\zeta_2 = \zeta_1$. Under given source positions, integrating over ζ (0 to ∞) and summing over k (1 to ∞) yields estimates for the mean and covariance;

$$\mathrm{E}(\widehat{X}) = \frac{N}{N+1}\,X^{\frac{N}{N+1}}\,,$$

$$\mathrm{E}(\widehat{X}_1\widehat{X}_2) = \frac{N}{N+2}\,X_1^{\left(\frac{N}{N+1}\;\frac{N}{N+2}\right)}\,X_2^{\frac{N}{N+1}}\,, \qquad (X_2 < X_1).$$

These formulae can be substituted into the evidence, written in its integrated-by-parts form $Z = \int X\,\mathrm{d}L$, to get

$$\mathrm{E}(\widehat{Z}) = \frac{N}{N+1}\int X^{\frac{N}{N+1}}\,\mathrm{d}L$$

$$\mathrm{E}(\widehat{Z}^2) = \frac{2N}{N+2} \iint\limits_{X_1>X_2} X_1^{\left(\frac{N}{N+1}\frac{N}{N+2}\right)} X_2^{\frac{N}{N+1}} \,\mathrm{d}L(X_1)\,\mathrm{d}L(X_2)$$

To obtain an upper bound on the mean-square error $\mathrm{E}((\widehat{Z}-Z)^2)$, relax $\mathrm{E}(\widehat{Z}^2)$ upwards, and $\mathrm{E}(\widehat{Z})$ downwards to $\frac{N}{N+1}\int X\,\mathrm{d}L = \frac{N}{N+1}Z$ to reach the inequality

$$\mathrm{E}\big((\widehat{Z}-Z)^2\big) \;<\; \frac{N}{N+2}\left(\int X^{\left(\frac{N}{N+1}\frac{N}{N+2}\right)}\,\mathrm{d}L\right)^2 - 2\frac{N}{N+1}Z^2 + Z^2 \qquad (19)$$

The error is thus limited by the bracketed integral

$$Q = \int X^{1-\alpha}\,\mathrm{d}L\,, \qquad \alpha = 1 - \frac{N}{N+1}\frac{N}{N+2} \qquad (20)$$

The worst case occurs when the likelihood function is chosen to maximise Q subject to the constraints (18). For N sufficiently large that $\alpha < \lambda$, there is a boundary $\widetilde{X}$ below which Q is increased at constant K by moving likelihood upwards towards $\widetilde{X}$, and above which Q is increased by moving likelihood downwards. Because L is non-increasing, the worst allowable case is when L is constant below $\widetilde{X}$ and zero above. In that case, evaluation of the integrals gives

$$K = \widetilde{X}^{-\lambda}\,, \qquad Q = \widetilde{X}^{-\alpha}Z = K^{\alpha/\lambda}Z\,.$$

Finally,

$$\frac{\mathrm{E}\big((\widehat{Z}-Z)^2\big)}{Z^2} \;<\; \frac{N}{N+2}K^{2\alpha/\lambda} - 2\frac{N}{N+1} + 1$$

As $N \to \infty$, this shrinks as $O(N^{-1})$. In practice, the number of iterates would need to increase in proportion to ensure adequate coverage, but we are here considering indefinitely large resources. This completes the proof of convergence of estimate $\widehat{Z}$ to correct Z as $N \to \infty$, at the usual $O(N^{-1/2})$ rate, with coefficient determined by the boundedness parameters K and λ, and no other assumption. □

Interestingly, if N is sufficiently small that $\alpha > \lambda$, $\widehat{Z}$ fails to converge at all in the sense that its moments don't exist. This is an example where simplistic averaging of separate results with small N is *very* much worse than doing inference properly, all at once with a single combined N.

The boundedness condition is most convincingly derived from an assumption that L is bounded — and it is needed. For example, the likelihood function

$$L(X) = \exp\frac{u^3}{u+1}\,, \qquad u = \sqrt{-\log X}$$

is monotonically decreasing with bounded $Z = 1.0119$ and $H = 0.6146$, yet $\int X^{-\lambda}L\,\mathrm{d}X$ diverges for any $\lambda > 0$. Hence the mean $\mathrm{E}(\widehat{Z})$ diverges for all N, no matter how large. What's happening is that this L has just enough singularity at small X to catch the occasional run which approaches it unusually quickly, rising enough to send $\widehat{Z}$ upwards strongly enough to destroy its moments. As previously proved, though, moments of $\log\widehat{Z}$ remain intact — another demonstration of the superiority of $\log Z$.

With both constraints (18) in place, we can take the large-N asymptotic limit of (20) to reach

$$Q = \int \big(1 - 3(1 + \log X)/N\big)\, L\, \mathrm{d}X$$

If we use this asymptotic form with the more realistic constraints (6) on evidence and information, the worst case occurs when L is a power law $L \propto X^{-h/(1+h)}$ for positive h, in terms of which

$$H = h - \log(1 + h)\,, \qquad Q = (1 + 3h/N)Z\,.$$

In the usual practical case that H is large, the logarithm drops away, leaving $H \approx h$. Substituting back in (19) yields

$$\frac{\mathrm{E}\big((\widehat{Z} - Z)^2\big)}{Z^2} < \frac{6H}{N}$$

and the rms error on $\log \widehat{Z}$ is $O(\sqrt{H/N})$, in line with the proven behaviour ((10) with $n = NH$).

It does, of course, require $N > H$ before the uncertainty in $\widehat{Z}$ becomes less than $\widehat{Z}$ itself, and $N > H/\epsilon^2$ is needed to drive the uncertainty down to a fraction ϵ. Such huge computational penalty compared with $N \sim 1$ would make nested sampling largely useless — *if* such accuracy was necessary. It isn't.

References

Chopin, N. and C. P. Robert. 2007. Discussion of "Nested Sampling for Bayesian computations". In *Bayesian Statistics 8*, eds. J. M. Bernardo, J. M. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 513–515. Oxford: University Press.

Evans, M. J. 2007. Discussion of "Nested Sampling for Bayesian computations". In *Bayesian Statistics 8*, eds. J. M. Bernardo, J. M. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 507–512. Oxford: University Press.

Sivia, D. S. and J. Skilling. 2006. *Data analysis; a Bayesian tutorial (2nd ed.)*, chap. 9. Oxford Univ. Press.

Skilling, J. 2006. Nested Sampling for general Bayesian computation. *J. Bayesian Analysis* 1: 833–860.

—. 2007. Nested Sampling for Bayesian computations. In *Bayesian Statistics 8*, eds. J. M. Bernardo, J. M. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 491–507. Oxford: University Press.