NESTED SAMPLING

John Skilling (skilling@eircom.net)

Bayesian Inference

$$\begin{pmatrix} \text{Prior}(\theta) \\ \times \\ \text{Likelihood}(\text{Data} \mid \theta) \end{pmatrix} = \text{Joint}(\theta, \text{Data}) = \begin{pmatrix} \text{Evidence}(\text{Data}) \\ \times \\ \text{Posterior}(\theta \mid \text{Data}) \end{pmatrix}$$

Substrate and Modulation

MCMC computation needs two procedures:

① Sample θ from the equilibrium distribution $\pi(\theta)$ of the ...

② ... Transition scheme $\theta \rightleftharpoons \theta'$.

The **substrate** $dX = \pi(\theta)d\theta$ uses as much of Joint(θ) as possible while remaining tractable.

The **modulation** $L(\theta) = \text{Joint}(\theta) \,/\, \pi(\theta)$ is what's left.

Usually,

$$\left.\begin{array}{rcl} \text{Substrate} & = & \text{prior} \\ \text{Modulation} & = & \text{likelihood} \end{array}\right\} \text{ (maybe } \begin{matrix} \times \\ \div \end{matrix} \text{ importance weights).}$$

Bayesian Computation

$$\begin{pmatrix} \text{Substrate } \pi(\theta) \\ \times \\ \text{Modulation } L(\theta) \end{pmatrix} = \text{Joint}(\theta) = \begin{pmatrix} \text{Evidence } Z & (= \int L\pi \, d\theta) \\ \times & \\ \text{Posterior } P(\theta) & (= L\pi/Z) \end{pmatrix}$$

MCMC is programmed by letting the transitions operate under progressively stronger modulation.

Do you have a difficult problem?

Traditional Annealing

This standard algorithm uses progressive cooling $0 \le \lambda \le 1$, replacing L by L^λ and modulating the transitions by Metropolis-Hastings.

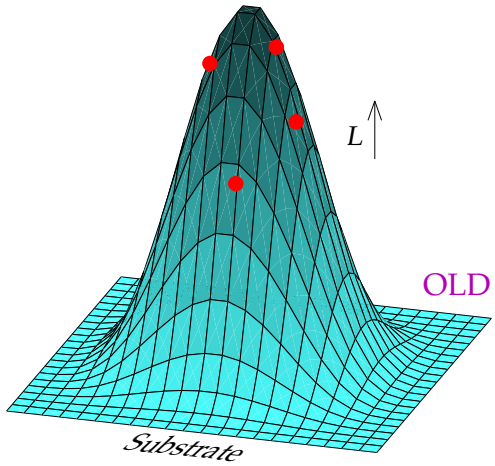
$$\boxed{\text{Accept } \theta' \quad \text{iff} \quad L^\lambda(\theta') \ge L^\lambda(\theta) \times \texttt{Uniform}(0,1)}$$

Every so often, increase the coolness λ according to its *annealing schedule* (usually guessed in advance). We assume that transitions *can* equilibrate θ, or an ensemble of N such points ●, at each new λ.

The posterior is reached when $\lambda = 1$, at which the evidence

$$\log Z = \int_0^1 \left\langle \log L \right\rangle_\lambda d\lambda$$

has been accumulated. After the "burn-in" is finished, an exploration phase explores the posterior.



Perhaps this is the answer.

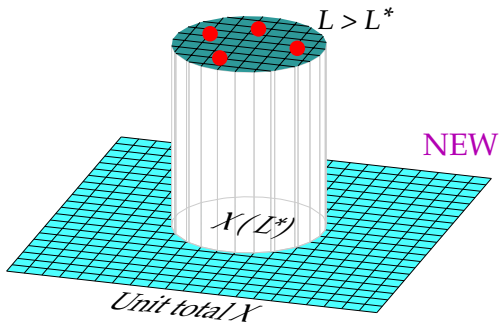
NESTED SAMPLING

This new algorithm uses a progressive lower limit $0 < L^* < \infty$, modulating the transitions by the simpler rule designed to give *uniform* sampling above the limit.

Accept θ' iff $L(\theta') \geq L^*$

Every so often, tighten the bound L^* by re-setting it to the current $L(\theta)$, or to the lowest L in an ensemble of N such points ●. We assume that transitions *can* find the required random new sample.

On average, this shrinks the available substrate by a factor $e^{-1/N}$, so that the k^{th} portion of substrate mass is $\delta X_k \propto e^{-k/N}$, normalised to unit total. A fuller treatment includes the randomness involved.



Nested sampling produces a sequence $\{\theta_k, L_k, \delta X_k\}$, from which the evidence is

$$Z = \sum L_k \, \delta X_k$$

and point θ_k has weight $L_k \, \delta X_k / Z$ in the posterior.

Convergence

Convergence of evidence and posterior in the usual $N^{-1/2}$ way is guaranteed provided only that $H < \infty$, where

$$H = \int \frac{P}{\pi} \log \frac{P}{\pi} \, \pi d\theta = \sum \frac{L_k}{Z} \log \frac{L_k}{Z} \, \delta X_k$$

$$= \log(\text{ substrate-to-posterior compression ratio })$$

How could it be otherwise? In fact, the number of steps needed to "burn through" the posterior mass is about NH, and the uncertainty is

$$\texttt{variance}(\log Z) \approx H/N \, .$$

Dimensionality, continuity, differentiability, topology and so on are not involved. As always, we merely assume that the transition scheme *can* cope with the modulation shape.

Extra generality at no cost

Nested sampling produces a sequence of points θ_k, weighted proportionally to $L_k \, \delta X_k$. These weights can be used to simulate annealing's integrand, formally

$$\left\langle \log L \right\rangle_\lambda = \frac{\int L^\lambda \log L \, dX}{\int L^\lambda \, dX}$$

at coolness λ, as

$$\left\langle \log L \right\rangle_\lambda = \frac{\sum L_k^\lambda \log L_k \, \delta X_k}{\sum L_k^\lambda \, \delta X_k}$$

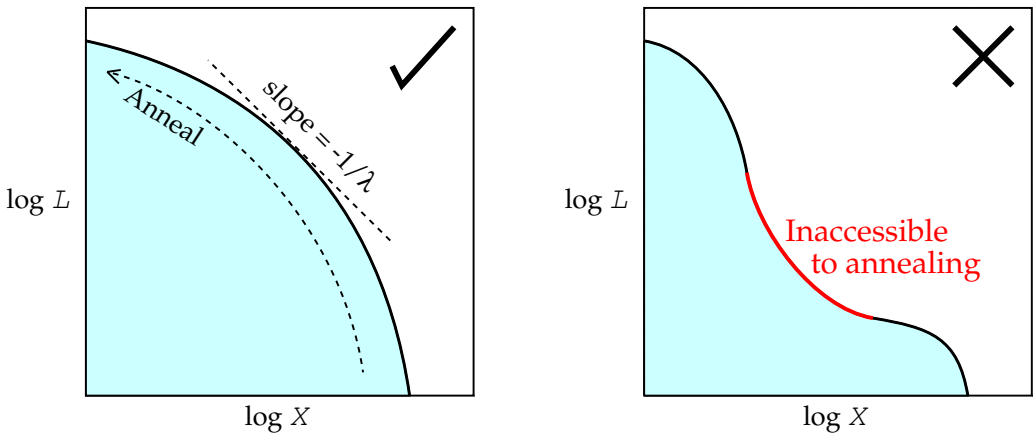
and thence reach annealing's estimate

$$\log Z = \int_0^1 \left\langle \log L \right\rangle_\lambda \, d\lambda$$

which is, of course, the same number as obtained directly from $\sum L_k \, \delta X_k$.

That relationship suggests that the optimal annealing schedule should be slaved to nested sampling's systematic compression in support ($\log X$) – and that seems to be true. Nested sampling's weights δX_k can be recovered from annealing's $\left\langle \log L \right\rangle_\lambda$. The methods can simulate each other and require the *same minimum CPU cost* to reach given accuracy with a given transition scheme.

But *only in simple problems* does λ increase as annealing requires.



Otherwise, *annealing does not work at all!* But nested sampling does.

Advantages of nested sampling

① **Simplicity.**
Exploration is directly over the substrate (= prior). There is no need to restrict the prior to conjugate form. There is no questionable annealing schedule.

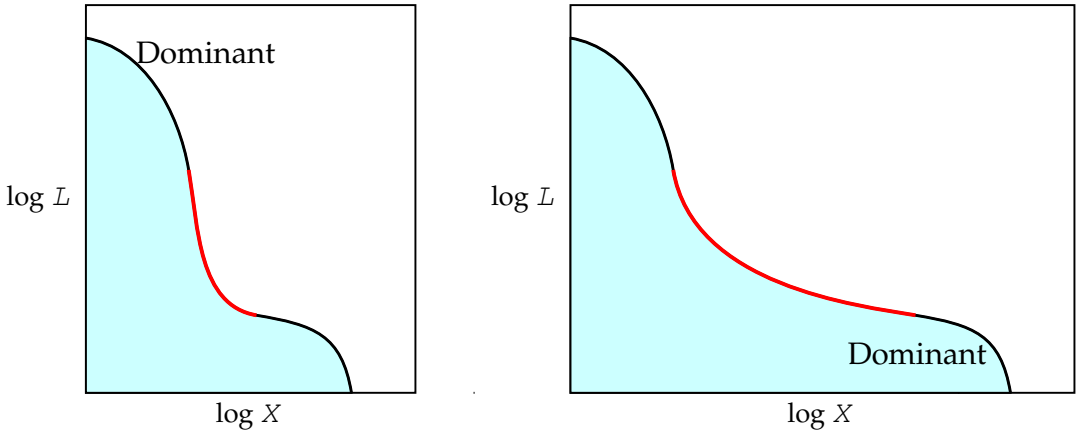
② **Efficiency.**
The posterior needle in the substrate haystack is found by compression at geometrical rate.

③ **Invariance.**
Exploration is invariant to monotonic re-labelling of L (= likelihood). Solve several problems for the price of one.

④ **Fundamental.**
Nested sampling gives a direct view of the problem's density of states, whereas thermal methods view it through a Laplace transform. The controlling "information" H (= log compression) is also the basis of information theory, and the key to automated acquisition of knowledge.

⑤ **Power.**
Nested sampling can cope with partly-convex ("$\smile$") likelihood functions that define *multi-phase problems* — protein folding, climate modelling, materials science, thermodynamic systems — that are currently thought very difficult.

Because it iterates on concave slope λ, annealing cannot distinguish between these situations



. . . which have different dominant phases and different evidence values. Because it iterates directly on the support $\log X$, nested sampling tracks the whole curve, does distinguish, and evaluates the evidence correctly.

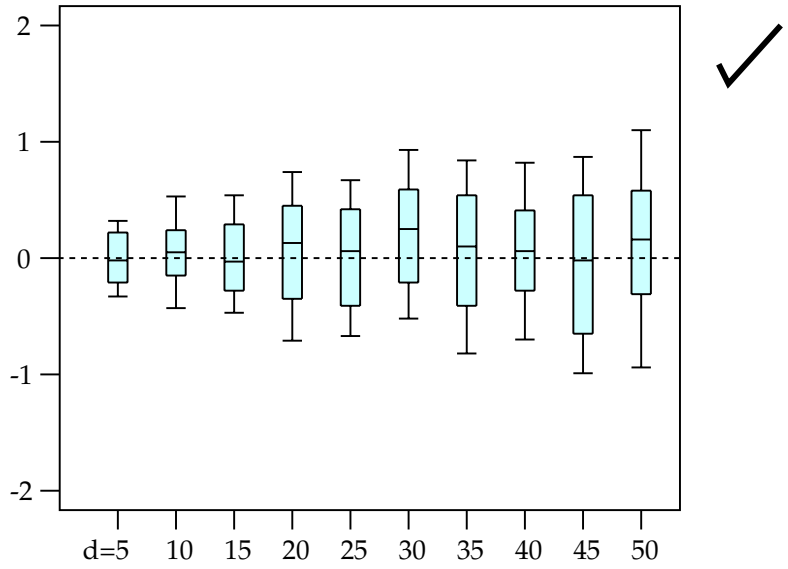
Tests

My own papers (JBA and Valencia8) introduce nested sampling, with some examples. Murray *et al.* have demonstrated its application to the Ising/Potts model, a second-order phase transition on the verge of inaccessibility to thermal methods. Mukherjee *et al.* have used it to compare cosmological models, on the basis of their evidence values. Chopin and Robert's website article ("C&R") publicises three statistics-style tests, which I reproduce here.

① De-centred Gaussian, testing dimension.

A d-dimensional ($d = 5, 10, 15, 20, \ldots$) Gaussian likelihood is offset down the side of a similarly-wide Gaussian prior π. C&R report a "bias increasing exponentially with the dimension", but that's wrong.

Nested sampling knows only the relation between L and enclosed substrate mass X. It has no way of detecting the underlying dimensionality, so cannot respond to it, and doesn't. My correct version of C&R's Figure 1 is below. It shows box-and-whisker plots at 10%, 25%, 50%, 75%, 90% quantiles of $\log Z$ excess (= computed − truth) for 100 runs of $N = 100$ points.

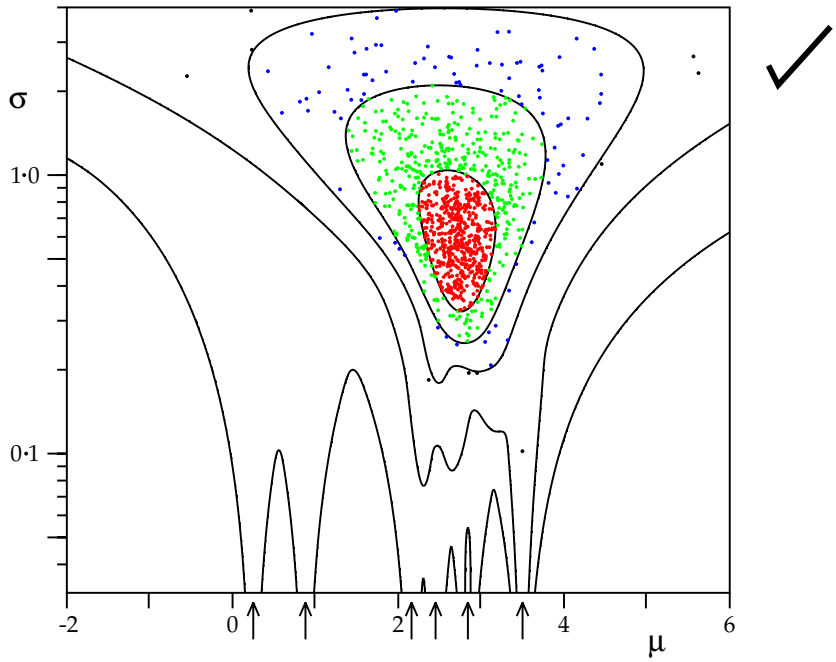


This shows no bias, and the expected amount $\pm\sqrt{H/N}$ of uncertainty ($H = 1.2216\,d$).

② **Mixture example, testing shape.**

This example has several local maxima in the $\theta = (\mu, \sigma)$ plane describing the unknown parameters of a Gaussian. Six of these are arrowed on the plot abscissa. C&R claim these small but (for small σ) unbounded peaks are "attractors for nested sampling", which exhibit "a fatal attraction" for nested sampling evaluation. That too is wrong.

Nested sampling has no way of detecting the shape of L as a function of μ and σ, so cannot respond to its corners and quirks, and doesn't. My correct version of C&R's Figure 7 (re-scaled to make the prior flat) is below. Likelihood contours enclose quantiles of true posterior at 0.0001%, 0.01%, 1%, 10%, 50%, from an exact calculation. One thousand posterior samples from nested sampling are ordered by likelihood with the lowest ten (0–1%) in black, the next ninety (1–10%) in blue, four hundred (10–50%) in green, and the top five hundred (50–100%) in red.

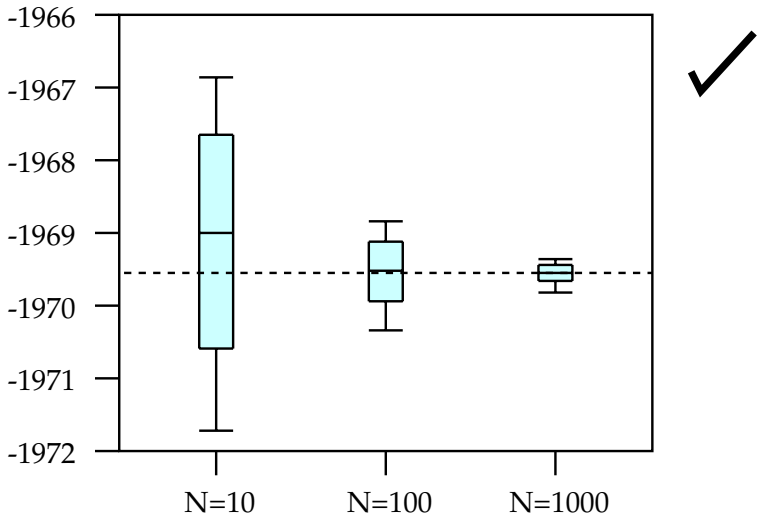


Nested sampling's points closely follow the correct quantile bands, with no false attraction. The agreement appears better than $\pm\sqrt{n}$ because the 1000 samples are at uniform quantiles of a bigger set, having too many points to plot.

③ **Probit example, testing convergence.**

This example has a product-of-probits likelihood function, using data from arsenic levels in wells in Bangladesh. Using ensembles of various sizes, ($N = 10, 100, 1000$ points θ), C&R observe a "small" bias — actually it's a very significant factor of three in Z. Once more, this is wrong.

A systematic bias would contradict the proof of convergence as $N \to \infty$, and it doesn't happen. My correct version of C&R's Figure 8 is below. It shows box-and-whisker plots at 10%, 25%, 50%, 75%, 90% quantiles of $\log Z$ for 100 runs for each number N. The truth $\log Z = -1969.552$ is shown dashed.



This shows no bias, and the expected amount $\pm\sqrt{H/N}$ of uncertainty ($H = 34.208$).

Disadvantages of nested sampling

I have found none.

Conclusion

The theory's fine, tests are fine, and a wider future beckons.

References

John Skilling (2006), *Nested sampling for general Bayesian computation*,
J. Bayesian Analysis **1**, 833–860.

John Skilling (2007), *Nested sampling for Bayesian computations*,
Bayesian Statistics **8** (Valencia conference), 491–507.

John Skilling (2008), *Nested sampling's convergence*,
submitted to Biometrika.

Sivia & Skilling (2006), *Data Analysis, a Bayesian Tutorial*,
(Oxford Univ. Press), chapter 9.

Murray, MacKay, Ghahramani & Skilling, (2006), *Nested sampling for Potts models*,
Advances in Neural Information Processing Systems **18**.

Mukherjee, Parkinson & Liddle (2006), *A nested sampling algorithm for cosmological model selection*
Astrophys. J. Letters **638**, L51– L54.

Nicolas Chopin and Christian Robert (2007),
Contemplating Evidence, properties, extensions of, and alternatives to nested sampling,
http://www.crest.fr/pageperso/Nicolas.Chopin/Nicolas.Chopin.htm (preprint link).