# Nested Sampling with Constrained Hamiltonian Monte Carlo

Michael Betancourt

*Massachusetts Institute of Technology, Cambridge, MA 02139*

**Abstract.** Nested sampling is a powerful approach to Bayesian inference ultimately limited by the computationally demanding task of sampling from a heavily constrained probability distribution. An effective algorithm in its own right, Hamiltonian Monte Carlo is readily adapted to efficiently sample from any smooth, constrained distribution. Utilizing this constrained Hamiltonian Monte Carlo, I introduce a general implementation of the nested sampling algorithm.

**Keywords:** Bayesian Inference, Nested Sampling, Hamiltonian Monte Carlo
**PACS:** 02.50.Tt, 02.70.Uu

## BAYESIAN INFERENCE

Bayesian inference is a diverse and robust analysis methodology [1, 2] based on Bayes' Theorem,

$$p(\alpha|\mathscr{D},H) = \frac{p(\mathscr{D}|\alpha,H)\,p(\alpha|H)}{p(\mathscr{D}|H)} \equiv \frac{\mathscr{L}(\alpha)\,\pi(\alpha)}{Z},$$

where information about the parameters $\alpha$ is extracted from the data $\mathscr{D}$. All model assumptions are captured by the conditioning hypothesis $H$.

While Bayes' Theorem is simple enough to formulate, in practice the individual components are often sufficiently complex that analytic manipulation is not feasible and one must resort to approximation. One of the more successful approximation techniques, Markov Chain Monte Carlo (MCMC) produces samples directly from the posterior distribution that are often sufficient to characterize even high dimensional distributions. The one manifest limitation of MCMC, however, is the inability to directly calculate the evidence $Z$.

Nested sampling [3] is an alternative to sampling from the posterior that instead emphasizes the calculation of the evidence.

## NESTED SAMPLING

Consider the support of the likelihood above a given bound $L$,

$$\tilde{\alpha} = \{\alpha|\mathscr{L}(\alpha) > L\},$$

and the associated prior mass across that support,

$$x(L) = \int_{\tilde{\alpha}} \mathrm{d}^m\alpha\,\pi(\alpha).$$

The differential $\mathrm{d}x$ gives the prior mass associated with the likelihood $L = \mathscr{L}(\alpha)$,

$$\mathrm{d}x(L) = \mathrm{d}\int_{\tilde{\alpha}} \mathrm{d}^m\alpha\,\pi(\alpha) = \int_{\partial\tilde{\alpha}} \mathrm{d}^m\alpha\,\pi(\alpha)$$

where $\partial\tilde{\alpha}$ is the $m-1$ dimensional boundary of constant likelihood,

$$\partial\tilde{\alpha} = \{\alpha|\mathscr{L}(\alpha) = L\}.$$

Introducing the coordinate $\alpha_\perp$ perpendicular to the likelihood constraint boundary and the $m-1$ coordinates $\alpha_\parallel$ parallel to the constraint, the integral over $\partial\tilde{\alpha}$ simply marginalizes $\alpha_\parallel$ and the differential becomes

$$\begin{aligned}
\mathrm{d}x(L) &= \int_{\partial\tilde{\alpha}} \mathrm{d}\alpha_\perp \mathrm{d}^{m-1}\alpha_\parallel\,\pi(\alpha) \\
\mathrm{d}x(L) &= \mathrm{d}\alpha_\perp \int_{\partial\tilde{\alpha}} \mathrm{d}^{m-1}\alpha_\parallel\,\pi(\alpha) \\
\mathrm{d}x(L) &= \mathrm{d}\alpha_\perp\,\pi(\alpha_\perp).
\end{aligned}$$

Returning to the evidence,

$$\begin{aligned}
Z &= \int \mathrm{d}^m\alpha\,\mathscr{L}(\alpha)\,\pi(\alpha) \\
Z &= \int \mathrm{d}\alpha_\perp \mathrm{d}^{m-1}\alpha_\parallel\,\mathscr{L}(\alpha)\,\pi(\alpha).
\end{aligned}$$

By construction the likelihood is invariant to changes in $\alpha_\parallel$ and the integral simplifies to

$$\begin{aligned}
Z &= \int \mathrm{d}\alpha_\perp \mathscr{L}(\alpha_\perp) \int \mathrm{d}^{m-1}\alpha_\parallel\,\pi(\alpha) \\
Z &= \int \mathrm{d}\alpha_\perp \mathscr{L}(\alpha_\perp)\,\pi(\alpha_\perp) \\
Z &= \int \mathrm{d}x\,L(x)
\end{aligned}$$

where $L(x) = \mathscr{L}(\alpha_\perp(x))$ is the likelihood bound resulting in the prior mass $x$.

This clever change of variables has reduced the $m$ dimensional integration over the parameters $\alpha$ to a one dimensional integral over the bounded support of $x$. Although this simplified integral is easier to calculate in theory, it is fundamentally limited by the need to compute $L(x)$.

Numerical integration, however, needs only a set of points $(x_k, L_k)$ and not $L(x)$ explicitly. Sidestepping $L(x)$, consider instead the problem of generating the set $(x_k, L_k)$ directly.

In particular, consider a stochastic approach beginning with $n$ samples drawn from $\pi(\alpha)$. The sample with the smallest likelihood, $\mathscr{L}_{\min}$, bounds the largest $x$ but otherwise nothing can be said of the exact value, $x_{\max}$, without an explicit, and painful, calculation from the original definition.

The cumulative probability of $x_{\max}$, however, is simply the probability of $x_{\max}$ exceeding the $x$ of each sample,

$$\begin{aligned}
P(x_{\max}) &= P(x_1 \leq x_{\max}) \cdots P(x_n \leq x_{\max}) \\
P(x_{\max}) &= \int_0^{x_{\max}} \mathrm{d}x\, \pi(x) \cdots \int_0^{x_{\max}} \mathrm{d}x\, \pi(x) \\
P(x_{\max}) &= \left( \int_0^{x_{\max}} \mathrm{d}x\, \pi(x) \right)^n
\end{aligned}$$

where $\pi(x)$ is uniformly distributed:

$$\begin{aligned}
\pi(x) &= \int_{\partial\tilde{\alpha}} \mathrm{d}^{m-1}\alpha_{\parallel}\, \pi(\alpha(x)) \left| \frac{\mathrm{d}\alpha}{\mathrm{d}x} \right| \\
\pi(x) &= \int_{\partial\tilde{\alpha}} \mathrm{d}^{m-1}\alpha_{\parallel}\, \pi(\alpha(x)) \frac{1}{\pi(\alpha_{\perp}(x))} \\
\pi(x) &= \pi(\alpha_{\perp}(x)) \frac{1}{\pi(\alpha_{\perp}(x))} \\
\pi(x) &= \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases} .
\end{aligned}$$

Simplifying, the cumulative probability of the largest sample reduces to

$$P(x_{\max}) = \left( \int_0^{x_{\max}} \mathrm{d}x\, \pi(x) \right)^n = \left( \int_0^{x_{\max}} \mathrm{d}x \right)^n = x_{\max}^n$$

with the corresponding probability distribution

$$p(x_{\max}) = \frac{\mathrm{d}P(x_{\max})}{\mathrm{d}x_{\max}} = n x_{\max}^{n-1}.$$

Estimating $x_{\max}$ from the probability distribution $p(x_{\max})$ immediately yields a pair

$$(x_1 = x_{\max}, L_1 = \mathscr{L}_{\min}).$$

A second pair follows by drawing from the constrained prior

$$\tilde{\pi}(\alpha) \propto \begin{cases} \pi(\alpha), & \mathscr{L}(\alpha) > \mathscr{L}_1 \\ 0, & \text{otherwise} \end{cases}$$

or, in terms of $x$,

$$,\tilde{\pi}(x) = \begin{cases} 1/x_1, & 0 \leq x \leq x_1 \\ 0, & \text{otherwise} \end{cases} .$$

$n$ samples from this constrained prior yield a new minimum $L_2$ with $x_2$ distributed as

$$p(x_2|x_1) = \frac{n}{x_1} \left( \frac{x_2}{x_1} \right)^{n-1}$$

Making another point estimate gives $(x_2, L_2)$.

Generalizing, the $n$ samples at each iteration are drawn from a uniform prior restricted by the previous iteration,

$$\tilde{\pi}(x) = \begin{cases} 1/x_{k-1}, & 0 \leq x \leq x_{k-1} \\ 0, & \text{otherwise} \end{cases}.$$

The distribution of the largest sample, $x_k$, follows as before,

$$p(x_k|x_{k-1}) = \frac{n}{x_{k-1}} \left( \frac{x_k}{x_{k-1}} \right)^{n-1}.$$

Note that this implies that the shrinkage at each iteration, $t_k = x_k/x_{k-1}$, is identically and independently distributed as

$$p(t_k) = p(t) = n t_k^{n-1}.$$

Moreover, a point estimate for $x_k$ can be written entirely in terms of point estimates for the $t_k$,

$$x_k = \frac{x_k}{x_{k-1}} \cdot \frac{x_{k-1}}{x_{k-2}} \ldots \frac{x_1}{x_0} \cdot x_0 = t_k \cdot t_{k-1} \ldots t_2 \cdot x_0 = \left( \prod_{i=1}^{k} t_i \right) x_0.$$

More appropriate to the large range common to many problems, $\log x_k$ becomes

$$\log x_k = \log \left( \prod_{i=1}^{k} t_i \right) x_0 = \sum_{i=1}^{k} \log t_i + \log x_0,$$

where the logarithmic shrinkage is distributed as

$$p(\log t) = n e^{n \log t}$$

with the mean and standard deviation

$$\log t = -\frac{1}{n} \pm \frac{1}{n}.$$

Taking the mean as the point estimate for each $\log t_i$ finally gives

$$\log \frac{x_k}{x_0} = -\frac{k}{n} \pm \frac{\sqrt{k}}{n}.$$

Parameterizing $x_k$ in terms of the shrinkage proves immediately advantageous – because the $\log t_i$ are independent, the errors in the point estimates tend to cancel and the estimate for the $x_k$ grow increasingly more accurate with $k$.

At each iteration, then, a pair $(x_k, L_k)$ is given by the point estimate for $x_k$ and the smallest likelihood of the $n$ drawn samples.

A proper implementation of nested sampling begins with the initial point $(x_0 = 1, L_0 = 0)$. At each iteration, $n$ samples are drawn from the constrained prior

$$\tilde{\pi}(\alpha) \propto \begin{cases} \pi(\alpha), & \mathscr{L}(\alpha) > \mathscr{L}_{k-1} \\ 0, & \text{otherwise} \end{cases}$$

and the sample with the smallest likelihood provides a "nested" sample with $L_k = \mathscr{L}(\alpha_k)$ and $\log x_k = -\frac{k}{n}$. $\mathscr{L}(\alpha_k)$ defines a new constrained prior for the following iteration. Note that the remaining samples from the given iteration will already satisfy this new likelihood constraint and qualify as $n-1$ of the samples necessary for the next iteration – only one new sample will actually need to be generated.

As the algorithm iterates, regions of higher likelihood are reached until the nested samples begin to converge to the maximum likelihood. Determining this convergence is tricky, but heuristics have been developed that are quite successful for well behaved likelihoods [3, 4].

Once the iterations have terminated, the evidence is numerically integrated using the nested samples. The simplest approach is a first order numerical quadrature:

$$Z \approx \sum_k (x_{k-1} - x_k) L_k$$

Errors from the numerical integration are dominated by the errors from the use of point estimates and, consequently, higher order quadrature offers little improvement beyond the first order approximation.

The remaining obstacle to a fully realized algorithm is the matter of sampling from the prior given the likelihood constraint $\mathscr{L} > \mathscr{L}_{\min}$. Sampling from constrained distributions is a notoriously difficult problem but a slight extension of Hamiltonian Monte Carlo offers samples directly from the constrained prior and provides an immediate implementation of nested sampling.

## CONSTRAINED HAMILTONIAN MONTE CARLO

Hamiltonian Monte Carlo [1, 5, 6] is an efficient method for generating samples from the $m$ dimensional probability distribution

$$p(\mathbf{x}) \propto \exp\left[-E(\mathbf{x})\right].$$

First, consider instead the larger distribution

$$p(\mathbf{x}, \mathbf{p}) = p(\mathbf{x})\, p(\mathbf{p})$$

where the latent variables $\mathbf{p}$ are i.i.d. standardized Gaussians

$$p(\mathbf{p}) \propto \exp\left(-\frac{1}{2}|\mathbf{p}|^2\right).$$

The joint distribution of the initial $\mathbf{x}$ and the latent $\mathbf{p}$ is then

$$p(\mathbf{x}, \mathbf{p}) \propto \exp\left(-\frac{1}{2}|\mathbf{p}|^2 - E(\mathbf{x})\right) = \exp(-H)$$

where $H \equiv \frac{1}{2}|\mathbf{p}|^2 + E(\mathbf{x})$ takes the form of the Hamiltonian of classical mechanics.

Applying Hamilton's equations

$$\frac{d\mathbf{x}}{dt} = \frac{\partial H}{\partial \mathbf{p}} = \mathbf{p}$$

$$\frac{d\mathbf{p}}{dt} = -\frac{\partial H}{\partial \mathbf{x}} = -\nabla E(\mathbf{x})$$

to a given sample $\{\mathbf{x}, \mathbf{p}\}$ produces a new sample $\{\mathbf{x}', \mathbf{p}'\}$. Note that the properties of Hamiltonian dynamics, in particular Liouville's Theorem and conservation of $H$, guarantee that differential probability masses from $p(\mathbf{x}, \mathbf{p})$ are conserved by the mapping. As a result, this dynamic evolution serves as a transition matrix $T(\mathbf{x}, \mathbf{p}; \mathbf{x}', \mathbf{p}')$ with the invariant distribution $p(\mathbf{x}, \mathbf{p})$. Moreover, the time reversal symmetry of the equations ensures that the evolution satisfies detailed balance:

$$T(\mathbf{x}, \mathbf{p}; \mathbf{x}', \mathbf{p}') = T(\mathbf{x}', \mathbf{p}'; \mathbf{x}, \mathbf{p}).$$

Because $H$ is conserved, however, the transitions are not ergodic and the samples do not span the full support of $p(\mathbf{x}, \mathbf{p})$. Ergodicity is introduced by adding a Gibbs sampling step for the $\mathbf{p}$. Because the $\mathbf{x}$ and $\mathbf{p}$ are independent, sampling from the conditional distribution for $\mathbf{p}$ is particularly easy

$$p(\mathbf{p}|\mathbf{x}) = p(\mathbf{p}) = \prod_{i=1}^{m} \mathcal{N}(0,1).$$

The algorithm proceeds by alternating between dynamical evolution and Gibbs sampling and the resulting samples $\{\mathbf{x}_k, \mathbf{p}_k\}$ form a proper Markov chain.

In practice the necessary integration of Hamilton's equations cannot be performed analytically and one must resort to numerical approximations. Unfortunately, any discrete approximation will lack the symmetry necessary for both Liouville's Theorem and energy conservation to hold, and the exact invariant distribution will no longer be $p(\mathbf{x}, \mathbf{p})$. This can be overcome by treating the evolved sample as a Metropolis proposal, accepting proposed samples with probability

$$P(\text{accept}) = \min(1, \exp(-\Delta H)).$$

Further implementation details, particularly insight on the choice of step size and total number of steps, can be found in [6].

Now consider the constrained distribution

$$\tilde{p}(\mathbf{x}) \propto \begin{cases} p(\mathbf{x}), & C(\mathbf{x}) \geq 0 \\ 0, & \text{else} \end{cases}.$$

Sampling from $\tilde{p}(\mathbf{x})$ is challenging. The simplest approach is to sample from $p(\mathbf{x})$ and discard those not satisfying the constraint. For most nontrivial constraints, however, this approach is extremely inefficient as the majority of the computational effort is spent generating samples that will be immediately discarded.

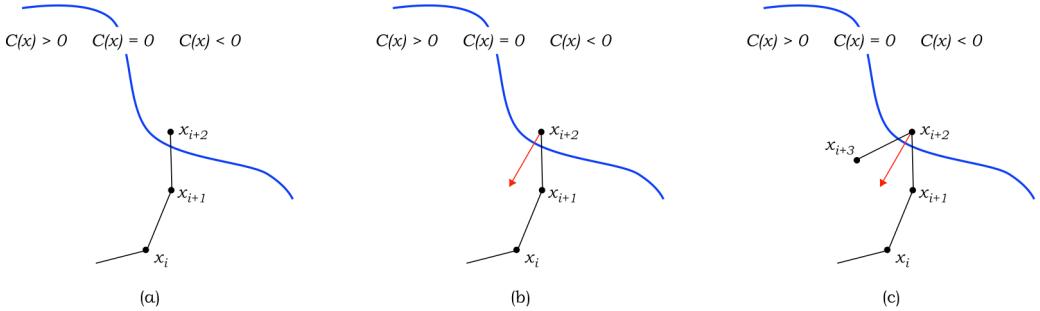


**FIGURE 1.** Cartoon of a particle bouncing off the constraint boundary $C(x) = 0$. (a) At step $i+2$ the particle violates the constraint, at which point (b) the normal at $x_{i+2}$ is computed and the momenta reflected in lieu of the normal leapfrog update. (c) The next spatial update is no longer in violation of the constraint.

From the Hamiltonian perspective, the constraint becomes an infinite barrier

$$\tilde{E}(\mathbf{x}) = \begin{cases} E(\mathbf{x}), & C(\mathbf{x}) \geq 0 \\ \infty, & \text{else} \end{cases}.$$

Incorporating infinite barriers directly into Hamilton's equations is problematic, but physical intuition provides an alternative approach. Particles incident on an infinite barrier bounce, the momenta perpendicular to the barrier perfectly reflecting:

$$\mathbf{p}' = \mathbf{p}_T - \mathbf{p}_N = \mathbf{p} - 2(\mathbf{p} \cdot \hat{\mathbf{n}})\hat{\mathbf{n}}.$$

Instead of dealing with infinite gradients, then, one can replace the momenta updates with reflections when the equations integrate beyond the support of $\tilde{p}(\mathbf{x})$.

Discrete updates proceed as follows. After each spatial update the constraint is checked and if violated then the normal $\hat{\mathbf{n}}$ is computed at the new point and the ensuing momentum update is replaced by reflection (Fig 1). Note that the spatial update cannot be reversed, nor can an interpolation to the constraint boundary be made, without spoiling the time-reversal symmetry of the evolution.

For smooth constraints $C(\mathbf{x}) \geq 0$ the normal is given immediately by

$$\hat{\mathbf{n}} = \nabla C(\mathbf{x}) / |\nabla C(\mathbf{x})|.$$

The normal for many discontinuous constraints, which are particularly useful for sampling distributions with limited support without resorting to computationally expensive exponential reparameterizations, can be determined by the geometry of the problem.

Finally, if the evolution ends in the middle of a bounce, with the proposed sample laying just outside of the support of $\tilde{p}(\mathbf{x})$, it is immediately rejected as the acceptance probability is zero,

$$P(\text{accept}) = \exp(-\Delta H) = \exp(-\infty) = 0.$$

Given a seed satisfying the constraint, the resultant Markov chain bounces around $\tilde{p}(\mathbf{x})$ and avoids the inadmissible regions almost entirely. Computational resources are spent on the generation of relevant samples and the sampling proceeds efficiently no matter the scale of the constraint.

## Application to Nested Sampling

Constrained Hamiltonian Monte Carlo (CHMC) naturally complements nested sampling by taking

$$p(\mathbf{x}) \to \pi(\alpha)$$
$$C(\mathbf{x}) \to \mathscr{L}(\alpha) - L.$$

The CHMC samples are then exactly the samples from the constrained prior necessary for the generation of the nested samples. A careful extension of the constraint also allows for the addition of a limited support constraint, making efficient nested sampling with, for example, gamma and beta priors immediately realizable.

Initially, the $n$ independent samples are generated from $n$ Markov chains seeded at random across the full support of $\pi(\alpha)$. After each iteration of the algorithm, the Markov chain generating the nested sample is discarded and a new chain is seeded with one of the remaining chains. Note that this new seed is guaranteed to satisfy the likelihood constraint and the resultant CHMC will have no problems bouncing around the constrained distribution to produce the new sample needed for the following iteration.

A suite of C++ classes implementing nested sampling with CHMC is available for general use at `http://web.mit.edu/~betan/www/code.html` . The accompanying documentation provides comprehensive details of the implementation.

## CONCLUSIONS

Constrained Hamiltonian Monte Carlo is a natural addition to nested sampling, the combined implementation allowing efficient and powerful inference for any problem with a smooth likelihood.

## ACKNOWLEDGEMENTS

I thank Tim Barnes, Chris Jones, John Rutherford, Joe Seele, and Leo Stein for insightful discussion and comments.

## REFERENCES

1. MacKay, D. J. C. (2003) *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, New York
2. Jaynes, E. T. (2003) *Probability Theory: The Logic of Science*, Cambridge University Press, New York
3. Skilling, J. (2004) Nested Sampling. In *Maximum Entropy and Bayesian methods in science and engineering* (ed. G. Erickson, J. T. Rychert, C. R. Smith). *AIP Conf. Proc.,* **735**: 395-405.
4. Sivia, D. S. with Skilling, J. (2006) *Data Analysis*. Oxford, New York
5. Bishop, C.M. (2007) *Pattern Classification and Machine Learning*. Springer, New York
6. Neal, R. M. *MCMC using Hamiltonian dynamics*, `http://www.cs.utoronto.ca/~radford/ham-mcmc.abstract.html`, March 5, 2010.