A review of nested sampling

Abstract

Accurately estimating the evidence is the main obstacle in Bayesian model selection. However, standard approaches to Bayesian inference, such as MCMC, are exclusively tailored to posterior sampling and do not provide satisfactory estimates of the evidence. To address this issue, nested sampling frames the evidence as a one-dimensional integral via the tail formula for expectations, which is approximated by quadrature methods. As a byproduct, weighted samples from the posterior are obtained.

1. Introduction

Let π and $\mathcal{L}$ be the prior and likelihood functions defined on a parameter space Θ. The evidence is defined as

$$Z = \int_\theta \pi(\theta)\mathcal{L}(\theta)\,\mathrm{d}\theta = \mathbb{E}_{\theta\sim\pi}[\mathcal{L}(\theta)]. \tag{1}$$

Define $\varphi(\lambda) = \Pr[\mathcal{L} \geq \lambda]$. Then,

$$Z = \int_0^\infty \varphi(\lambda)\,\mathrm{d}\lambda = \int_0^1 \lambda(\varphi)\,\mathrm{d}\varphi. \tag{2}$$

Equation 2 transforms a potentially high-dimensional problem into a one-dimensional problem, which can be solved via quadrature methods. The challenge now is to estimate $\lambda(\varphi)$.

2. Algorithm

Skilling (2006) [1] proposed an algorithm that concomitantly yields quadrature points and estimates of $\lambda(\varphi)$ to approximate Equation 2. The algorithm works as follows. First, we sample N points from the prior. Then, we select the point with smallest likelihood, θ_1, and the corresponding likelihood, $\mathcal{L}(\theta_1) = L_1$. Then, we discard θ_1 and sample from the constrained prior $\pi(\theta)1_{\mathcal{L}(\theta)>L_1}$. We proceed in this fashion until we have a collection $(L_1, ..., L_n)$ of likelihoods, which are embedded into the quadrature formula with placeholder points $x_i = \exp(-\frac{i}{n})$.

3. Explanation

We follow Betancourt's approach [2] to better understand this algorithm. Firstly, define $\tilde{\alpha} = \{\alpha : \mathcal{L}(\alpha) > L\}$. The prior mass associated to this quantity is denoted by

$$x(L) = \int_{\tilde{\alpha}}^\infty \mathrm{d}^m\alpha\pi(\alpha). \tag{3}$$

Intuitively, the differential $\mathrm{d}x(L)$ may be computed as $\int_{\partial\tilde{\alpha}}^\infty \mathrm{d}^m\alpha\pi(\alpha)$. By letting $\alpha_\perp$ and $\alpha_\parallel$ the coordinates perpendicular and parallel to the surface $\partial\alpha = \{\alpha : \mathcal{L}(\alpha) = L\}$, we observe that

$$\mathrm{d}x(L) = \mathrm{d}\alpha_\perp\pi(\alpha_\perp), \tag{4}$$

i.e., we're simply marginalizing over $\alpha_\parallel$. Under these conditions and noticing that changes in $\alpha_\parallel$ do not affect $\mathcal{L}(\alpha)$, the evidence may be computed as

$$\begin{aligned} Z &= \int \mathrm{d}\alpha_\perp \mathrm{d}^{m-1}\alpha_\parallel \mathcal{L}(\alpha_\perp)\pi(\alpha) \\ &= \int \mathrm{d}\alpha_\perp \mathcal{L}(\alpha_\perp) \int \mathrm{d}^{m-1}\alpha_\parallel \pi(\alpha) \\ &= \int \mathrm{d}\alpha_\perp \mathcal{L}(\alpha_\perp)\pi(\alpha_\perp), \end{aligned} \tag{5}$$

which is a one-dimensional integral. By letting $L(x)$ represent the likelihood associated to the prior mass x, the prior integral may be written as

$$Z = \int \mathrm{d}x L(x). \tag{6}$$

Intuitively, this computation is grounded on Adam's law and the conditioning of $\mathcal{L}(\alpha_\perp)$ on $\mathcal{L}(\alpha_\perp) = L$.

To find a collection of points (x_k, L_k) for numerical integration, we first note that, when the α are sampled from π, x are uniformly distributed,

$$\begin{aligned}\pi(x) &= \int_{\partial\tilde{\alpha}} d^{m-1}\alpha_{\parallel}\pi(\alpha(x))\left|\frac{\mathrm{d}\alpha}{\mathrm{d}x}\right| \\ &= \int_{\partial\tilde{\alpha}} d^{m-1}\alpha_{\parallel}\pi(\alpha(x))\left|\frac{1}{\pi(\alpha_{\perp})}\right| \\ &= \left|\frac{1}{\pi(\alpha_{\perp})}\right| \int_{\partial\tilde{\alpha}} d^{m-1}\alpha_{\parallel}\pi(\alpha(x)) \\ &= \left|\frac{1}{\pi(\alpha_{\perp}(x))}\right| \pi(\alpha_{\perp}(x)) = 1\end{aligned} \tag{7}$$

when $x \in (0,1)$ (we used the change of variables' formula, the fact that $\mathrm{d}\alpha_{\parallel}$ does not depend on x and $\mathrm{d}x = \mathrm{d}\alpha_{\perp}\pi(\alpha_{\perp})$). Then, we notice that our best estimate for the largest value $x_{\max}$ is the sample associated with the minimum likelihood. Since $\pi(x)$ is uniform, $x_{\max}$ has distribution $p(x_{\max}) = nx_{\max}^{n-1}$. Given this first sample, which we call (x_1, L_1), the following sample will be distributed as $\pi(x) = \frac{1}{x_1}$ when $0 \leq x \leq x_1$. By following this algorith, we note that the shrinkage operators, $t_i = \frac{x_i}{x_{i-1}}$ are independently and identically distributed with $p(t_k) = nt_k^{n-1}$. In particular, one may readily notice that

$$\log x_k = \log x_o + \sum \log t_k, \tag{8}$$

and hence that the expected value of $\log x_k$ is $-\frac{i}{n}$ (recall that n is the number of samples). This provides the value for the placement points.

Overall, the algorithm seems to work well in practice; however, a deeper understanding of its convergence rates to the evidence are mostly lacking from the literature. The method is widely adopted in the literature of natural sciences, but is often neglected by standard computational statistics textbooks. Our approach here was quite heuristic.

Betancourt [2] implemented a constrained HMC algorithm to sample from the likelihood-constrained prior distribution. [3], on the other hand, showed that the HMC scheme may rely on reverse-mode autodiff for evaluateing the gradients and developed an adaptive step-size when simulating the Hamiltonian dynamics. Interestingly, the resulting samples were used to train a GFlowNets with both forward (from a fixed initial state) and backward (from the sampled trajectories) sampling, and the results were quite impressive.

Bibliography

[1] J. Skilling, "Nested sampling for general Bayesian computation," Bayesian Analysis, vol. 1, no. 4, pp. 833–859, 2006, doi: 10.1214/06-BA127.

[2] M. Betancourt, A. Mohammad-Djafari, J.-F. Bercher, and P. Bessiére, "Nested Sampling with Constrained Hamiltonian Monte Carlo," in AIP Conference Proceedings, AIP, 2011. doi: 10.1063/1.3573613.

[3] P. Lemos, N. Malkin, W. Handley, Y. Bengio, Y. Hezaveh, and L. Perreault-Levasseur, "Improving Gradient-guided Nested Sampling for Posterior Inference." [Online]. Available: https://arxiv.org/abs/2312.03911