Using Thermodynamic Integration to Calculate the Posterior Probability in Bayesian Model Selection Problems

Paul M. Goggans* and Ying Chi*

*University of Mississippi
Department of Electrical Engineering
University, MS 38677

Abstract. This paper gives an algorithm for calculating posterior probabilities using thermodynamic integration. The thermodynamic integration calculations are accomplished by annealing an ensemble of Markov chains with an adaptive schedule. The algorithm includes a method for determining "good" starting positions for the chains at each new value of the annealing parameter.

INTRODUCTION

At The 22nd Annual Conference on Bayesian Methods and Maximum Entropy in Science and Engineering, John Skilling gave a presentation suggestively titled "How I Do It" in which he presented a method for making Bayesian model selection calculations [1]. Skilling's method uses thermodynamic integration to calculate the posterior probabilities. The thermodynamic integration calculations are accomplished by annealing an ensemble of Markov chains. In this paper we present some details of Skilling's method as we have worked them out.

MODEL SELECTION

Given the set of propositions $M_k =$ "The measured data are from model k, $g_k(t,\mathbf{x}_k)$," $k = 1,\ldots,K$ the goal of model selection is to determine which model has the highest credibility in light of the observed data and known information. The propositions are assumed to be mutually exclusive but not necessarily complete.

Model selection is performed by comparing the posterior probabilities for the K models. Using Bayes rule, an expression for the posterior probability for model k can be written as

$$p(M_k|DI) \propto p^\star(M_k|I)p^\star(D|M_kI). \tag{1}$$

In (1) the symbol D is a proposition that denotes the data produced in an observation while the symbol I is a proposition denoting the prior information. Here the superscript on $p^\star$ denotes an unnormalized probability or *pdf*. In (1) the term $p^\star(M_k|I)$ is a prior probability and the term $p^\star(D|M_kI)$ is the likelihood. The prior must be assigned. Taking

the log of (1) yields

$$\ln p(M_k|DI) = \ln p^\star(D|M_kI) + \ln p^\star(M_k|I) + \text{Constant}. \tag{2}$$

For an incomplete set of models it is convenient to compare the models by computing the natural log of the posterior odds ratio for each model:

$$\ln(\text{odds ratio}) = \ln\left\{\frac{p(M_k|DI)}{\min[p(M_k|DI)]}\right\}. \tag{3}$$

The value of the log odds ratio for model k does not dependent on the constant in (2) and so can be calculated using $\ln p^\star(D|M_kI)$ and $\ln p^\star(M_k|I)$. Because the prior $p(M_k|I)$ is assigned, calculation of the odds ratios reduces to calculating values that are proportional to the likelihoods $p(D|M_kI)$ or, equivalently, calculating $\ln p^\star(D|M_kI)$.

Using Bayes' Rule the posterior *pdf* for the parameters of model k can be written as:

$$p(\mathbf{x}_k|DM_kI) = \frac{p(\mathbf{x}_k|M_kI)\,p^\star(D|M_k\mathbf{x}_kI)}{p^\star(D|M_kI)}, \tag{4}$$

where the normalizing constant

$$p^\star(D|M_kI) = \int d\mathbf{x}_k\, p(\mathbf{x}_k|M_kI)\,p^\star(D|M_k\mathbf{x}_kI) \tag{5}$$

is proportional to the desired likelihoods. While (5) is a formal expression for the likelihoods, it is usually not useful for actually calculating the likelihoods. Multidimensional integrals over the model parameters must be evaluated in order to calculate the likelihoods using (5). Because $p^\star(D|M_k\mathbf{x}_kI)$ can only be evaluated for specific values of $\mathbf{x}_k$ these integrals cannot be evaluated analytically. In general, numerical evaluation of multidimensional integrals using quadrature is difficult and becomes more difficult as the dimensionality of the integral increases. The dimensionality of the integral required to evaluate (5) can be high since it is equal to the number of parameters in the model under consideration. In addition, the integrand often has one or more very large and very narrow peaks so that a few small regions of the parameter space contribute most of the integral's value. Also, for some problems the dynamic range of $p^\star(D|M_k\mathbf{x}_kI)$ is sufficiently large so that only the log of $p^\star(D|M_k\mathbf{x}_kI)$ can be expressed as a floating-point number. In this case the likelihoods can not be evaluated using (5).

THERMODYNAMIC INTEGRATION

Thermodynamic integration is an indirect method for calculating $\ln p^\star(D|M_kI)$ that avoids the difficulties associated with the direct evaluation of (5). Thermodynamic integration comes originally from statistical thermodynamics but is derived here mathematically following [2].

Derivation of the method begins by introducing an annealing parameter β into (4) and defining

$$p(\mathbf{x}|MD\beta I) \triangleq \frac{p(\mathbf{x}|MI)\{p^\star(D|M\mathbf{x}I)\}^\beta}{p^\star(D|M\beta I)} \quad \text{for} \quad 0 \le \beta \le 1 \tag{6}$$

where

$$p^\star(D|M\beta I) = \int d\mathbf{x}\, p(\mathbf{x}|MI)\{p^\star(D|M\mathbf{x}I)\}^\beta. \tag{7}$$

In the expressions above, the subscript k has been dropped from M and $\mathbf{x}$ to simplify the notation. For $\beta = 1$,

$$p^\star(D|M\beta I)\Big|_{\beta=1} = p^\star(D|MI) \tag{8}$$

is the desired likelihood. For $\beta = 0$

$$p^\star(D|M\beta I)\Big|_{\beta=0} = \int d\mathbf{x}\, p(\mathbf{x}|MI) = 1 \tag{9}$$

because a normalized prior for the model parameters must be assigned.

Using the chain rule,

$$\frac{d}{d\beta}\ln p^\star(D|M\beta I) = \frac{1}{p^\star(D|M\beta I)}\frac{d}{d\beta}p^\star(D|M\beta I). \tag{10}$$

Substituting (7) into the right-hand side of (10), taking the derivative and simplifying the result yields

$$\frac{d}{d\beta}\ln p^\star(D|M\beta I) = -\int d\mathbf{x}\, E_L(\mathbf{x})\, p(\mathbf{x}|MD\beta I), \tag{11}$$

where

$$E_L(\mathbf{x}) \stackrel{\triangle}{=} -\ln p^\star(D|M\mathbf{x}I). \tag{12}$$

The integral in (11) can be written as the expected value of the energy $E_L(\mathbf{x})$ so that

$$\frac{d}{d\beta}\ln p^\star(D|M\beta I) = -\langle E_L(\mathbf{x})\rangle_\beta. \tag{13}$$

Integrating the equation above with respect to β from 0 to 1 yields the desired expression for the log likelihood:

$$\ln p^\star(D|MI) = -\int_0^1 d\beta\, \langle E_L(\mathbf{x})\rangle_\beta. \tag{14}$$

CALCULATING THE LOG LIKELIHOOD

The method for calculating $\ln p^\star(D|MI)$ using (14) depends on the ability of the Markov chain Monte Carlo method (MCMC) to easily approximate the expected value of functions. For each value of β, the integrand of (14) is given by the expression

$$\langle E_L(\mathbf{x})\rangle_\beta = \int d\mathbf{x}\, E_L(\mathbf{x})\, p(\mathbf{x}|DM\beta I). \tag{15}$$

To approximate $\langle E_L(\mathbf{x})\rangle_\beta$, a sample of $\mathbf{x}$ is drawn from each of an ensemble of J Markov chains[1]. Replacing $p(\mathbf{x}|DM\beta I)$ in (15) with its Monte Carlo approximation,

$$p(\mathbf{x}|DM\beta I) \approx \frac{1}{J}\sum_{j=1}^{J}\delta(\mathbf{x}-\mathbf{x}_j), \tag{16}$$

yields

$$\langle E_L(\mathbf{x})\rangle_\beta \approx \frac{1}{J}\sum_{j=1}^{J} E_L(\mathbf{x}_j). \tag{17}$$

With $\langle E_L(\mathbf{x})\rangle_\beta$ calculated for discrete values of β for $0 \le \beta_i \le 1$, the one dimensional integral in (14) can be determined using any appropriate quadrature rule. For example, using the Trapezoidal rule, the integral can be approximated by

$$\int_0^1 d\beta\, \langle E_L(\mathbf{x})\rangle_\beta \approx \sum_{i=1}^{i=I-1} \frac{[\langle E_L(\mathbf{x})\rangle_{\beta_{i+1}} - \langle E_L(\mathbf{x})\rangle_{\beta_i}]\Delta\beta_i}{2}, \tag{18}$$

where $\beta_1 = 0$, $\beta_I = 1$, and $\Delta\beta_i = \beta_{i+1} - \beta_i > 0$ for $i = 1,\ldots,I-1$. In the evaluation of (17) for use in (18), the idea is to use the final samples drawn from the ensemble of J Markov chains at β_i to determine both β_{i+1} and the starting positions for the chains at β_{i+1}. In [1] Skilling presented practical methods for accomplishing both of these tasks.

Putting aside for the moment the problem of determining β_{i+1}, we assume that β_{i+1} has been determined and focus on determining starting positions for the chains at β_{i+1} from the ending positions of the chains at β_i. It is convenient to assume that the model has been reparametrized so that $p(\mathbf{x}|MI)$ is uniform on the unit hypercube[2]. In this case, no starting chain positions are needed for $\beta = 0$ ($i = 1$) since the J samples of $\mathbf{x}$ can be drawn directly using a uniform random number generator. For the assumed uniform prior,

$$p^\star(\mathbf{x}|MD\beta I) = \exp(-\beta E_L(\mathbf{x})). \tag{19}$$

Importance sampling with resampling [5, 6] is used to determine starting positions for the chains at β_{i+1} from the ending positions of the chains at β_i. At β_{i+1}, the importance weights are calculated for the ending $\mathbf{x}_j$ at β_i using the expression

$$w_j = \frac{p^\star(\mathbf{x}_j|MD\beta_{i+1}I)}{p^\star(\mathbf{x}_j|MD\beta_i I)} = \exp(-\Delta\beta_i E_L(\mathbf{x}_j)). \tag{20}$$

---

[1] Note that the $\mathbf{x}_j$ can be drawn from $p(\mathbf{x}|DM\beta I)$ with MCMC using $\ln p^\star(D|M\mathbf{x}I)$ (and the assigned prior *pdf* for $\mathbf{x}$) without needing to know the value of the normalizing constant in the denominator of (6). This is important because $\ln p^\star(D|M\mathbf{x}I)$ can be calculated directly solving the dynamic range problem and the normalizing constant is the quantity we wish to determine.

[2] In the Markov chain Monte Carlo method, this reparametrization is necessary if the Hilbert curve is to be used in a binary slice sampling algorithm [3, 4]. With the Hilbert curve a single integer can represent two or more real parameters so that multi-dimensional slice sampling is reduced to one-dimensional slice sampling. Use of the Hilbert curve with binary slice sampling avoids the problems often encountered in setting the adjustable parameters of a multivariate MCMC method

The weights are then normalized so that

$$W_j = J\frac{w_j}{\sum_{j=1}^{J} w_j}. \tag{21}$$

Using the normalized weights to form a Monte Carlo approximation gives

$$p(\mathbf{x}|DM\beta_{i+1}I) \approx \frac{1}{J}\sum_{j=1}^{J} W_j \delta(\mathbf{x}-\mathbf{x}_j). \tag{22}$$

Resampling (22) according to the normalized importance weights so that the new weights are non-negative integers yields a Monte Carlo approximation from which the chain starting positions can be determined;

$$p(\mathbf{x}|DM\beta_{i+1}I) \approx \frac{1}{J}\sum_{j=1}^{J} N_j \delta(\mathbf{x}-\mathbf{x}_j), \tag{23}$$

where $\langle N_j \rangle = W_i$ and $\sum_{j=1}^{J} N_j = J$. Because N_j can be zero some samples can be deleted and because N_j can be greater than one, some samples can be repeated. The values of N_j are chosen so that for any bounded function $f(\mathbf{x})$

$$\langle f(\mathbf{x}) \rangle \approx \frac{1}{J}\sum_{j=1}^{J} N_j f(\mathbf{x}_j) \tag{24}$$

and so that the expectation of the right-hand side of (24) converges (in some sense, see [5]) to $\langle f(\mathbf{x}) \rangle$ as $J \to \infty$. Because of this, the set of samples with N_j copies of $\mathbf{x}_j$ for $j = 1 \ldots J$ can be thought of as representative of $p(\mathbf{x}|DM\beta_{i+1}I)$ and so used as the starting positions for the chains at β_{i+1}. These starting positions are clearly not independent so a sufficient number of chain steps must be taken so that the chain positions are reasonably independent before the final chain positions are used to approximate the value of $\langle E_L(\mathbf{x}) \rangle_{\beta+1}$.

In Skilling's method for resampling the chains are first sorted according to their weights so that $j = 1$ corresponds to the chain with the minimum weight and $j = J$ corresponds to the chain with the maximum weight. After sorting, the integer valued weights are determined for $j = 1, \ldots, J$ using the expression

$$N_j = \sum_{k=0}^{J-1} \left[\mathrm{U}\left(u + k - \sum_{i=1}^{j-1} W_i \right) - \mathrm{U}\left(u + k - \sum_{i=1}^{j} W_i \right) \right] \tag{25}$$

where $u \sim \text{uniform}(0,1)$ and the unit step function

$$\mathrm{U}(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{for } x \leq 0. \end{cases}$$

Figure 1 illustrates the determination of N_j for an example with five chains. The figure is a vertically stacked bar chart where the width of each bar is the normalized weight

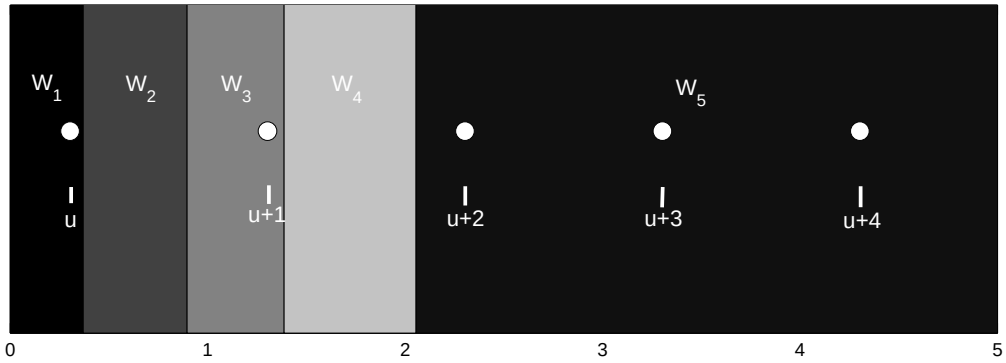

FIGURE 1. Vertically stacked bar chart of normalized importance weights.

for each chain. The bars are stacked in order of increasing width because this ensures that similarly-weighted chains are treated similarly[3]. The points $u+k$ for $k=0,\ldots,4$ are also plotted on the bar chart. The value of N_j is equal to the number of points that fall within the bar of width W_j. For the example in Figure 1,

$$p(\mathbf{x}|DM\beta_{i+1}I) \approx \frac{1}{5}\Big(\delta(\mathbf{x}-\mathbf{x}_1)+\delta(\mathbf{x}-\mathbf{x}_3)+3\delta(\mathbf{x}-\mathbf{x}_5)\Big). \tag{26}$$

Because of the use of resampling some of the chains in the ensemble can begin at the same starting positions. If too many of the chains begin at the same starting positions then the MCMC will have to be run for many chain steps to achieve reasonably independent samples. Because of this, it is prudent to choose $\Delta\beta_i$ so that most of the N_j are equal to one. In particular, it is important to avoid discarding most of the samples in favor of a few samples with the highest weights. Choosing $\Delta\beta_i$ to achieve a fixed ratio of the maximum weight to the minimum weight accomplishes these goals. Using (19) it is straightforward to show that

$$\Delta\beta_i = \frac{\ln\left\{\frac{\max(w_j)}{\min(w_j)}\right\}}{\max[E_L(\mathbf{x}_j)]-\min[E_L(\mathbf{x}_j)]}. \tag{27}$$

The value of $\max(w_j)/\min(w_j)$ should be slightly greater that one but not too much greater. Use of (27) results in adaptive annealing that decreases $\Delta\beta_i$ (slows cooling) when $\max[E_L(\mathbf{x}_j)]-\min[E_L(\mathbf{x}_j)]$ increases indicating that the MCMC is having difficulty.

Figure 2 illustrates the adaptive annealing that occurs in a problem proposed by Cornelius Lanczos [7]. In this problem the data are obtained by evaluating $0.0951\exp(-t)+0.8607\exp(-3t)+1.5576\exp(-5t)$ at $t=0.05m$ for $0\le m\le 23$ and then rounding the

[3] For example, if a group of 10 chains each have weight $W=2.4$, then exactly 24 starting positions will be taken from these chains. In contrast, random ordering of the chains would result in a Poisson distribution of mean 24 for the number of starting positions taken from the group.

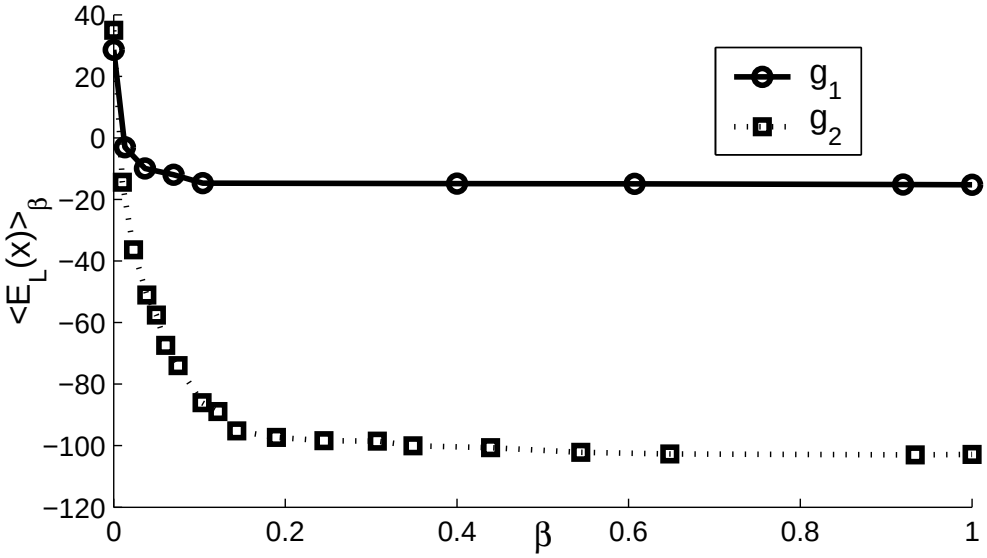

FIGURE 2. Expected value of the energy for two different models.

result to two decimal places. In Figure 2, $\langle E_L(\mathbf{x})\rangle_{\beta_i}$ is plotted for the following models: $g_1(t) = 3x_1 \exp[-3x_2 t]$ and $g_2(t) = 3x_1 \exp[-3x_2 t] + 3x_3 \exp[-3(x_2 + x_4)t]$. For the purpose of making a clear illustration, $\max(w_j)/\min(w_j) = 2$ and an ensemble of 10 chains were used.

SUMMARY

The following pseudo-code summarizes the calculation of $\ln p^\star(D|MI)$ using the *selective annealing* algorithm presented in the previous section:

$\beta_1 = 0$
Loop on the β index $i = 1, 2, \ldots$
  If $\beta_i = 0$ {
    Draw x_j for $j = 1, \ldots, J$ from the unit hypercube }
  If $\beta_i \neq 0$ {
    Draw x_j for $j = 1, \ldots, J$ from $p(\mathbf{x}|DM\beta_i I)$ using MCMC and the starting
      positions determined at β_{i-1} }
  Calculate $\langle E_L(\mathbf{x})\rangle_{\beta_i}$ using (17)
  If $\beta_i = 1$ {
    $I = i$
    Break out of the β index loop }
  Calculate $\Delta\beta_i$ using (27)
  $\beta_{i+1} = \beta_i + \Delta\beta_i$
  If $\beta_{i+1} > 1$ {
    $\beta_{i+1} = 1$
    $\Delta\beta_i = 1 - \beta_i$ }

Calculate w_j for $j = 1, \ldots, J$ using (20)
Calculate W_j for $j = 1, \ldots, J$ using (21)
Sort the chains according to W_j
Calculate N_j for $j = 1, \ldots, J$ using (25)
Determine the chain starting positions for β_{i+1} using N_j and x_j
End of β index loop
Calculate $\ln p^*(D|MI)$ using (18)

ACKNOWLEDGMENTS

This work is based upon and motivated by a presentation given by John Skilling at The 22nd Annual Conference on Bayesian Methods and Maximum Entropy in Science and Engineering. We are grateful to Dr. Skilling for providing us with a copy of his transparencies and to Dr. C. Ray Smith for reading a draft of this paper.

This material is based in part upon work supported by the US Army Research Office under contract number DAA19-99-1-0108.

REFERENCES

1. Skilling, J., "How I Do It," at *The 22nd Annual Conference on Bayesian Methods and Maximum Entropy in Science and Engineering*, 2002, Presentation Only.
2. Neal, R. M., Probabilistic inference using Markov chain Monte Carlo methods, Technical Report CRG-TR-93-1, University of Toronto (1993).
3. Skilling, J., and MacKay, D. J. C., *Annals of Statistics*, **31**, 753–755 (2003), discussion of *Slice Sampling* by Radford M. Neal.
4. Mackay, D. J. C., *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.
5. Doucet, A., Freitas, N. D., and Gordon, N. J., "An Introduction to Sequential Monte Carlo Methods," in *Sequential Monte Carlo Methods in Practice*, edited by A. Doucet, N. D. Freitas, and N. J. Gordon, Springer-Verlag, 2001.
6. Liu, J. S., Chen, R., and Logvinenko, T., "A Theoretical Framework for Sequential Importance Sampling with Resampling," in *Sequential Monte Carlo Methods in Practice*, edited by A. Doucet, N. D. Freitas, and N. J. Gordon, Springer-Verlag, 2001.
7. Lanczos, C., *Applied Analysis*, Prentice Hall, 1954, reprinted by Dover Books.