BAYESIAN STATISTICS 9,
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Using TPA for Bayesian inference

Mark Huber & Sarah Schott
Claremont McKenna College, USA *Duke University, USA*
mhuber@cmc.edu schott@math.duke.edu

Summary

Finding the integrated likelihood of a model given the data requires the integration of a nonnegative function over the parameter space. Classical Monte Carlo methods for numerical integration require a bound or estimate of the variance in order to determine the quality of the output. The method called the product estimator does not require knowledge of the variance in order to produce a result of guaranteed quality, but requires a cooling schedule that must have certain strict properties. Finding a cooling schedule can be difficult, and finding an optimal cooling schedule is usually computationally out of reach. TPA is a method that solves this difficulty, creating an optimal cooling schedule automatically as it is run. This method has its own set of requirements; here it is shown how to meet these requirements for problems arising in Bayesian inference. This gives guaranteed accuracy for integrated likelihoods and posterior means of nonnegative parameters.

Keywords and Phrases: Adaptive Monte Carlo; Variance free approximation.

1. INTRODUCTION

Traditional Monte Carlo methods for numerical integration rely on estimates to determine the variance of the output. There exist methods, however, that provide guarantees on performance without the need to either calculate or estimate a variance.

TPA is one such method for approximating the integral of nonnegative functions over high dimensional spaces. Use of the method requires several precise ingredients, and the purpose of this work is to show how to obtain those ingredients for Bayesian applications.

Consider the problem of finding the integrated likelihood (also known as the evidence, marginal likelihood, or normalizing constant) for a model. For data y parameterized by the random variable θ in parameter space Ω_θ with prior measure μ_{prior} and likelihood function $L(\theta \mid y)$, the integrated likelihood is

$$Z = \mathrm{E}_{\mu_{\text{prior}}}[L(\theta \mid y)] = \int_{b \in \Omega_\theta} L(b \mid y)\, d\mu_{\text{prior}}.$$

First author supported by NSF CAREER grant DMS-05-48153.

Suppose the posterior measure of θ given data y is denoted μ_{post}. Since the Radon-Nikodym derivative $[d\mu_{\text{post}}(b)/d\mu_{\text{prior}}(b)] = L(b \mid y)/Z$, the integrated likelihood Z is also known as the *normalizing constant.* Also, Z appears in the Bayes factor for model selection, so another term for Z is the *evidence* for a model.

For a d dimensional parameter $\theta = (\theta_1, \ldots, \theta_d)$, finding the posterior mean θ_i leads to a second integration, namely:

$$\mathrm{E}_{\mu_{\text{post}}}[\theta_i \mid y] = \mathrm{E}_{\mu_{\text{prior}}}[\theta_i L(\theta \mid y)]/\mathrm{E}_{\mu_{\text{prior}}}[L(\theta \mid y)].$$

In both cases the problem reduces to integrating a density against a prior, although for the posterior mean case it is necessary to break the integral into two pieces: one where $\theta_i \geq 0$, and another where $\theta_i < 0$ in order to evaluate them using TPA.

The rest of the paper is organized as follows. The next section introduces the TPA method, and then shows two ways in which it can be applied to finding the integrated likelihood. This is followed by an artificial multimodal example, and then by another example, the hierarchical beta-binomial model. The next section then builds on the basic TPA algorithm to give an *omnithermal* estimate. This type of estimate is especially useful in spatial settings, and its application is illustrated here using the Ising model. The next section discusses the effects of imperfect samples, followed by a discussion of the use of Rao-Blackwellization with TPA. A comparison to the nested sampling method follows, and the final section discusses fully Bayesian approximation algorithms.

2. USING TPA FOR BAYESIAN PROBLEMS

The TPA method has four general ingredients:

(a) A measure space $(\Omega, \mathcal{F}, \mu)$.

(b) Two finite measurable sets B and B' satisfying $B' \subset B$. The set B' is the *center* and B is the *shell.*

(c) A family of nested sets $\{A(\beta) : \beta \in \mathbb{R}\}$ such that $\beta < \beta'$ implies $A(\beta) \subseteq A(\beta')$, $\mu(A(\beta))$ is a continuous function of β, and $\lim_{\beta \to -\infty} \mu(A(\beta)) = 0$.

(d) Special values β_B and $\beta_{B'}$ that satisfy $A(\beta_B) = B$ and $A(\beta_{B'}) = B'$.

Let $p = \mu(B')/\mu(B)$. Our goal is to create an approximation algorithm with output $\hat{p}$ such that for inputs $\epsilon > 0$ and $\delta \in [0, 1]$:

$$\Pr((1+\epsilon)^{-1} \leq \hat{p}/p \leq 1+\epsilon) > 1-\delta. \tag{1}$$

So our goal is to do more than just bound the variance of our estimates, but to also put bounds on the tails as well.

Traditional acceptance/rejection draws multiple times from $\mu(B)$, finds the sample percentage of times the resulting sample falls in B', and uses that for $\hat{p}$. With this approach, the expected number of variates generated before a single sample falls in B' is $1/p$. This method requires $\Theta(p^{-1}\epsilon^{-2}\ln(\delta^{-1}))$ samples to meet our (ϵ, δ) requirement. The $\epsilon^{-2}\ln(1/\delta)$ factor comes from standard Monte Carlo analyses, but improvement can be made in the p^{-1} factor.

TPA operates by moving inward from B to B' using a sequence of samples. Begin with $\beta = \beta_B$, so that $A(\beta) = B$. The first sample X is a draw from $\mu(A(\beta))$.

Next find the smallest value of β' such that $X \in \mu(A(\beta'))$ (condition (c) guarantees the existence of such a β'.) The set $A(\beta')$ becomes our new space, and the next sample drawn comes from $A(\beta')$. This in turn yields a new value of β and so on, repeating until the sample lands in the center, $B' = A(\beta_{B'})$. The number of samples needed to reach the center will form the basis of our approximation method.

To determine the distribution of the number of samples needed to reach the center, first note that $\mu(A(\beta'))/\mu(A(\beta))$ is a uniform random variable over $[0, 1]$. To see this, suppose $X \sim \mu(A(\beta))$, $\beta' = \max\{b : X \in A(b)\}$. The essential idea is that for any $a \in (0, 1)$, the random variate X has probability a of falling into a region $A(\beta')$ such that $\mu(A(\beta'))/\mu(A(\beta)) = a$. This argument is made precise in the following theorem.

Theorem 1 *Given ingredients (a) through (d) above and β such that $\mu(A(\beta)) < \infty$, let $X \sim \mu(A(\beta))$, $\beta' = \inf\{b : X \in A(b)\}$, and $U = \mu(A(\beta'))/\mu(A(\beta))$. Then $U \sim \mathrm{Un}([0, 1])$.*

Proof. Fix β and let $a \in [0, 1)$. Then since $\mu(A(b))$ is a continuous function in b where $\lim_{b\to-\infty} \mu(A(b)) = 0$, there must exist a $b \in (-\infty, \beta]$ such that $\mu(A(b))/\mu(A(\beta)) = a$. Call this value β_a.

Let $0 < \epsilon < 1-a$. Then there is also a value $\beta_{a+\epsilon}$ such that $\mu(A(\beta_{a+\epsilon}))/\mu(A(\beta)) = a+\epsilon$.

Now consider $X \sim \mu(A(\beta))$, set $\beta' = \inf\{b : X \in A(b)\}$, and let $U = \mu(A(\beta'))/\mu(A(\beta)$. Then $X \in A(\beta_a) \Rightarrow U \leq a$, so $\Pr(U \leq a) \geq \Pr(X \in A(\beta_a)) = a$.

On the other hand,

$$X \notin A(\beta_{a+\epsilon}) \Rightarrow \beta' \geq \beta_{a+\epsilon} \Rightarrow \mu(A(\beta'))/\mu(A(\beta)) \geq a + \epsilon \Rightarrow U \geq a + \epsilon.$$

The contrapositive of the above statement says $U < a + \epsilon \Rightarrow X \in A(\beta_{a+\epsilon})$. So viewed as a statement about probabilities (combining with the previous inequality)

$$a \leq \Pr(U \leq a) \leq \Pr(U < a + \epsilon) \leq a + \epsilon,$$

and since ϵ was an arbitrary number in $(0, 1 - a)$, $\Pr(U \leq a) = a$.

Hence $\Pr(U \leq a) = a$ for all $a \in [0, 1)$, and $U \sim \mathrm{Un}([0, 1])$. □

If this procedure is repeated k times, a sequence of β values are generated, say $\beta = \beta_0, \beta_1, \beta_2, \ldots, \beta_k$, where each of the ratios $\mu(A(\beta_i))/\mu(A(\beta_{i+1}))$ is uniform over $[0, 1]$. In other words,

$$\frac{\mu(A(\beta_k))}{\mu(A(\beta_0))} \sim U_1 U_2 \cdots U_k, \text{ where } U_i \overset{\text{iid}}{\sim} \mathrm{Un}([0, 1]).$$

Now if $U \sim \mathrm{Un}([0, 1])$, then $-\ln U \sim \mathrm{Ex}(1)$. So consider the points

$$P_k := -\ln\left(\frac{\mu(A(\beta_k))}{\mu(A(\beta_0))}\right) \sim E_1 + E_2 + \cdots E_k, \text{ where } E_i \overset{\text{iid}}{\sim} \mathrm{Ex}(1).$$

Then the points $\{P_i\}$ form a one dimensional Poisson point process with rate 1.

Suppose the process continues until $\beta' \leq \beta_{B'}$, that is, until the X variate drawn lands in the center B'. Then the number of samples drawn before the center is reached will have a Poisson distribution with parameter $\ln(\mu(B))/\ln(\mu(B'))$.

Algorithm 2.1 TPA$(r, \beta_B, \beta_{B'})$

Input: Number of runs r, initial index β_B, final index $\beta_{B'}$
Output: $\hat{p}$ (estimate of $\mu(B')/\mu(B)$)

```
k ← 0
for i from 1 to r do
    β ← β_B, k ← k − 1
    while β > β_{B'} do
        k ← k + 1, X ← μ(A(β)), β ← inf{β' ∈ [β_{B'}, β_B] : X ∈ A(β')}
    end while
end for
p̂ ← exp(k/r)
```

Recall that the union of r Poisson point processes of rate 1 is a new Poisson point process with rate r. So repeat the procedure r times and let k be the sum of the number of samples needed to reach the center in each run. Then $k \sim \mathrm{Po}(r \ln(\mu(B))/\ln(\mu(B')))$. The approximation to $\mu(B)/\mu(B')$ is $\exp(k/r)$. This is encoded in Algorithm 2.1.

To determine the value of r needed to obtain an (ϵ, δ) approximation, it is necessary to bound the tails of a Poisson distribution. In Section 7 this is accomplished using Chernoff bounds, where it is shown that for $r = 2(\ln p)^2(3\epsilon^{-1} + \epsilon^{-2})\ln(4\delta^{-1})$, TPA is an (ϵ, δ) approximation algorithm. Since in typical applications, p is exponentially small in the dimension of the problem, having the dependence on p be a polynomial in $\ln p^{-1}$ is necessary to be efficient. Of course, in practice, $\ln p^{-1}$ is not known ahead of time, so TPA can be run as a two phase procedure. In the first phase set $r = \ln 2\delta^{-1}$, so that TPA estimates $\ln p^{-1}$ within a factor of $1 + 3\sqrt{\ln p^{-1}}$ with probability at least $1 - \delta/2$. In the second phase, this initial estimate is used to determine the value of r to find the final estimate $\hat{p}$ that is accurate to a factor of $1 + \epsilon$ with probability at least $1 - \delta/2$. The union bound then states that both phases were correct with probability at least $1 - \delta$.

Two methods of setting up (a), (b), (c), and (d) will be considered here: parameter truncation and likelihood truncation.

2.1. *Parameter truncation*

For ingredient (a), the parameter space usually is a subset of $\mathbb{R}^d$ equipped with the Borel sets. The measure μ will be

$$\mu(A) = \int_{b \in A} L(b \,|\, y)\, d\mu_{\text{prior}} = \mathrm{E}_{\mu_{\text{prior}}}[L(\theta \,|\, y)\mathbf{1}(\theta \in A)]$$

In parameter truncation, the family of nested sets is formed by restricting (truncating) parameter space. A simple example of such a family is

$$A(M) = \Omega_\theta \cap \{\theta : \|\theta - c\| \le M\}, \tag{2}$$

where c is a fixed point in parameter space. When $M = \infty$ this is just the original space (so $\beta_B = \infty$), and as M decreases the restriction narrows the space down.

The norm should be chosen to make the resulting sets as easy as possible to sample from. As long as the prior measure is continuous with respect to Lebesgue measure, the measure $\mu(A(M))$ will be continuous in M.

When M is very small, it is usually possible to bound the likelihood above and below, as it will be very close to $L(c \mid y)$. Then this $A(M)$ becomes B', and $\mu(A(\beta_{B'})) \approx \mu_{\text{prior}}(A(\beta_{B'}))L(c \mid y)$. This procedure is illustrated on examples in Section 3 and in Section 4.

2.2. *Likelihood truncation*

When a slice sampler Markov chain is being used to generate the samples, a more natural approach to creating the family of sets is to truncate the likelihood rather than the parameter.

Begin by noting that

$$Z = \int_{b \in \Omega_\theta} L(b \mid y)\, d\mu_{\text{prior}} = \int_{b \in \Omega_\theta} \int_0^{L(b \mid y)} 1\, dw\, d\mu_{\text{prior}},$$

where dw is just Lebesgue measure. In other words, $Z = \mu(\{(t_1, t_2) \in \Omega_\theta \times [0, \infty) : 0 \le t_2 \le L(t_1 \mid y)\})$. Here $\mu = \mu_{\text{prior}} \times m$ and m is Lebesgue measure. This μ is the measure over $\Omega \times [0, \infty)$ required by ingredient (a).

An auxiliary variable M can be introduced to this setup to create a series of nested sets as follows

$$A(M) = \{(t_1, t_2) \in \Omega_\theta \times [0, \infty) : 0 \le t_2 \le \min\{L(t_1 \mid y), M\}. \tag{3}$$

Then $\mu(A(\infty)) = Z$, and $A(\infty)$ will be the shell B in ingredient (b).

The value of $\mu(A(M))$ will vary continuously from 0 up to Z as M runs from 0 to ∞. So this provides our family of nested sets for ingredient (c).

Finding the center B' to go along with the shell is more tricky. Since the goal is to estimate $p = \mu(B')/\mu(B)$, setting the center to be $A(0)$ with measure 0 is not an option. Instead, the center needs to be a value M_{center} that is larger than 0, but for which $\mu(A(M_{\text{center}}))$ is easy to approximate (say by $\hat{c}$.) Then use $\hat{p}$ to approximate p, and use $(\hat{c}/\hat{p})$ as an approximation for $\mu(A(\infty)) = Z$.

The solution is to draw a set of samples from the prior distribution, and calculate the likelihood for each sample. The sample median of these likelihoods becomes the temperature for the center, M_{center}. For any $\delta > 0$, draw enough samples so that the probability that the sample median is actually below the 0.4 quantile is at most $\delta/2$. From Hoeffding's inequality (Hoeffding, 1963), $50 \ln(2/\delta)$ samples suffice.

Now for a random variable X drawn from the prior distribution,

$$\mathrm{E}[\min\{L(X \mid y), M_{\text{center}}\}] = \int_{\Omega_\theta} \min\{L(X \mid y), M_{\text{center}}\}\, d\mu_{\text{prior}},$$

or just $\mu(A(1))$. And since M_{center} was chosen so that $\Pr(L(X \mid y) \ge M_{\text{center}}) \ge .4$,

$$0.4 M_{\text{center}} \le \mathrm{E}[\min\{L(X \mid y), M_{\text{center}}\}] \le M_{\text{center}}.$$

This means that (by another application of Hoffding's inequality) it is possible to estimate $\mathrm{E}[\min\{L(X \mid y), M_{\text{center}}\}]$ within a factor of $1 + \epsilon$ with probability at least

$\delta/2$ by taking the sample mean of $.3\epsilon^{-2}\ln(\delta/2)$ draws. Hence from the union bound, the final estimate of $\mu(A(M_{\text{center}}))$ is an (ϵ, δ) approximation.

For actually generating samples from the family of truncated likelihoods, the slice sampler (see Robert and Casella, 2004, pp. 320–333 for a description) is just as easy to implement for sampling from $\min\{L(\theta \mid y), M\}$ as for $L(\theta \mid y)$, and as M shrinks should actually mix faster as local modes are truncated away.

2.3. *The name TPA*

This idea of sampling from nested sets appears also in the nested sampling algorithm of Skilling (Skilling, 2006), so a new name was needed for our method. We choose the rather whimsical name of Tootsie Pop Algorithm. A Tootsie Pop is a hard candy shell that encloses a chocolate chewy center. By licking the shell away, the chewy chocolate center is eventually revealed. In TPA, counting how long it takes to chip away the shell and reach the center is the essential statistic that allows us to approximate the ratio of the measure between the shell and center.

3. EXAMPLE: A MULTIMODEL LIKELIHOOD

This section illustrates the general theory with a specific multimodal example that was examined in on p. 854 of Skilling (2006), where it was acknowledged to be a difficult case for nested sampling. The prior for the parameter θ is uniform over $[-1/2, 1/2]^d$, and the likelihood for θ is

$$L(\theta) = 100\prod_{i=1}^{d}\frac{1}{\sqrt{2\pi}u}\exp\left(-\frac{(\theta_i - 0.2)^2}{2u^2}\right) + \prod_{i=1}^{d}\frac{1}{\sqrt{2\pi}v}\exp\left(-\frac{\theta_i^2}{2v^2}\right). \tag{4}$$

That is, the likelihood consists of a Gaussian spike centered at $(0.2, 0.2, \ldots, 0.2)$ mixed with a much smaller spike centered at $(0, 0, \ldots, 0)$. When $u = .01$ and $v = .02$, the chance of a draw from the d dimensional prior landing anywhere near one of the two modes is vanishingly small. This is typical in these types of problems: the likelihood is typically far more concentrated than the prior distribution.

It is important to note that TPA is not a solution to the problem of how to generate samples from a multimodal likelihood. It does, however, have the positive property that as the algorithm progresses, the sampling problem does not usually become any more difficult. In both parameter and likelihood truncation, the multimodality disappears as the algorithm progresses.

3.1. *Parameter truncation for the multimodal example*

To specify the truncation given by (2), it is necessary to specify the norm and the center point c. A natural choice of c is a mode or the center of parameter space, although any point in parameter space could be used. In this case the origin is both a mode and the center of parameter space, and a simple norm is the L_∞ norm that takes the maximum among the components of the parameter.

Set $\beta_{B'} = .0001$, so $B' = \{\theta \in \mathbb{R}^{20} : |\theta_i| \leq .0001 \text{ for all } i\}$. For (4), when $\|\theta\|_\infty < .0001$ the likelihood lies within .999 and 1.001 of $L((0, \ldots, 0))$, which equals $(2\pi v^2)^{-10}$ to at least 20 significant digits. Since the prior is uniform over $[-1/2, 1/2]^{20}$, the prior measure of B' is just $.0002^{20}$.

Hence $\mu(B')$ is within 1.001 of $.0002^{20}$, and all TPA needs to do is approximate $Z/\mu(B')$. Most things about this example can be calculated exactly. In particular $\ln(Z/\mu(B')) \approx 115.0993$.

The algorithm for generating samples and running TPA was coded in R. The code is available on the first author's website, or by request. After 10^5 runs of TPA, the estimate of $\ln(Z/\mu(B'))$ was 115.10321, so the number of samples generated during the course of the algorithm was 11510321, or about 10^7. This means the final approximation was within a factor of (1.004)(1.001) of the true integrated likelihood of 101. The first factor of 1.004 arises from Monte Carlo error and the second factor of 1.001 from the approximation to the integral for $M = \beta_{B'} = .0001$.

As expected from the theory of TPA, the number of samples used in each run followed a Poisson distribution, as can be seen in Figure 1. The bars are the empirical distribution of the runs, and the line is the density of a Poisson with the analytically determined mean.

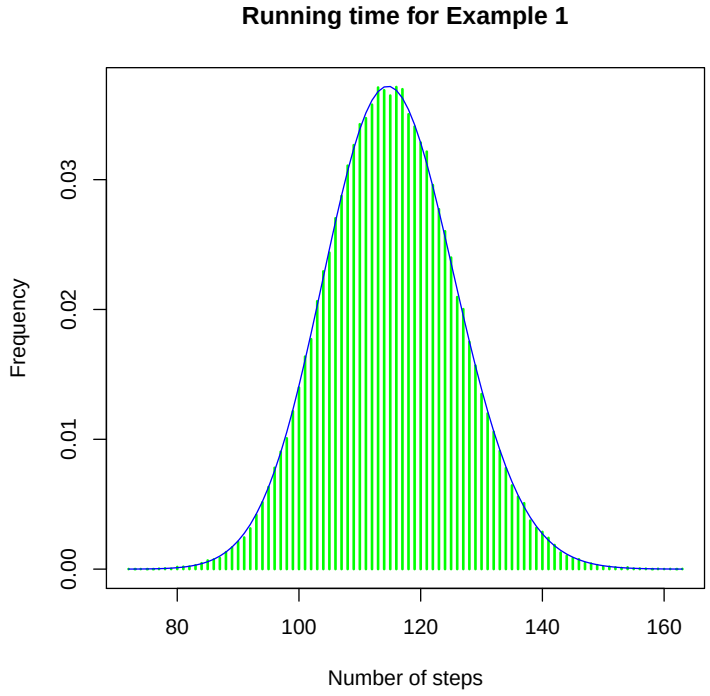

Figure 1: *TPA using parameter truncation for multimodal example with* 10^5 *runs.*

4. EXAMPLE: THE BETA-BINOMIAL MODEL

Raftery et al. (2006) considered finding the integrated likelihood for a beta-binomial hierarchical model for free throw data from the NBA. The counts y_i are modeled as binomially distributed with known number of trials n_i and unknown $p_i \overset{\text{iid}}{\sim} \text{Be}(a, b)$,

where a and b are independent hyperparameters with an Ex(1) prior shifted by 1.

The data used here consists of the number of free throws attempted and made during the 2008-2009 season. In this season, 429 NBA players attempted at least one free throw. Once the hyperparameters a and b are known the rest of the parameters can be integrated out, therefore, it is possible to find the exact answer numerically to test the accuracy of TPA.

As with the previous example, the algorithm was implemented in R. The true log integrated likelihood (to three decimal places) is -1577.250. After 10^5 runs of TPA, the average number of samples per draw was 30.71754, that is, 3071754 total samples from the posterior truncated at various values were generated. The resulting estimate of -1577.256 for the loglikelihood is well within the standard deviation of 0.017 predicted by theory.

5. APPROXIMATE SAMPLING

In many situations it is not possible to obtain exactly random samples from $\mu(A(\beta))$. Instead some approximate method such as Markov chain Monte Carlo will be used. The effect will be to heterogeneously stretch out or compress the Poisson process generated by TPA. As long as the same method is used at each step for creating samples, this will at least be a consistent effect. If more than one method for generating approximate samples is used, and if one or more of these methods stretch the state space they are unlikely to do so in the same fashion. Therefore, a simple diagnostic to test the effect of approximate sampling is to run the procedure with two unrelated Markov chains, and compare the results. Since the Monte Carlo error can be bounded precisely when using TPA, any remaining difference in the results can be correctly attributed to at least one of the Markov chains being used.

As usual, this method can show that the Markov chains are not mixing well, but in order to guarantee that the quality of the result an exact or perfect simulation method must be used.

6. POSTERIOR MEANS

The examples considered so far involved finding the integrated likelihood. However, the same methods can be applied to finding the posterior mean of a distribution. To find the mean of θ_i, instead of integrating against μ_{prior}, simply integrate against the measure with density $d\mu_{\text{mean}} = \theta_i d\mu_{\text{prior}}$. This will keep the integrand nonnegative as long as $\theta_i \geq 0$. In this case, it is possible (as with the integrated likelihood) to find the posterior mean without any need to consider the posterior variance.

7. OMNITHERMAL APPROXIMATION

We shall call an approximation of $\mu(A(\beta))/\mu(A(\beta_B))$ that is valid for all $\beta \in [\beta_{B'}, \beta_B]$ simultaneously an *omnithermal approximation*. The "thermal" portion of the name comes from the fact that in many models of interest (such as the Ising model), the parameter β is known as the inverse temperature. Therefore, an omnithermal approximation is one that is valid for all temperatures simultaneously.

Recall in Section 2 it was shown that the β values generated by r runs of TPA (not including the initial β value of each run) formed a one dimensional Poisson point process with rate r in logspace. Let P denote this set of β values. These points can be used to derive an omnithermal approximation. To go from a Poisson

point process to a Poisson process, set

$$N_P(t) = \#\{b \in P : b \geq \beta_B - t\}.$$

Then as t runs from 0 to $\beta_B - \beta_{B'}$, $N_P(t)$ increases by 1 whenever it hits a β value. By the theory of Poisson point processes, this happens at intervals that will be independent exponential random variables of rate r.

Given $N_P(t)$, approximate $\mu(B)/\mu(A(\beta))$ by $\exp(N_P(\beta_B - \beta)/r)$. When $\beta = \beta_{B'}$, this is just our usual approximation, and so this is a generalization of the basic TPA procedure.

Note $\mathrm{E}[N_P(t)] = rt$, and $N_P(t) - rt$ is a right continuous martingale. To bound the error in $\exp(N_P(t)/r)$, it is necessary to bound the probability that $N_P(t) - rt$ has drifted too far away from 0.

Theorem 2 *Let $\epsilon > 0$. Then for $N_P(\cdot)$ a rate r Poisson process on $[0, T]$, where $\epsilon/T \leq 2.3$:*

$$\Pr\left(\sup_{t\in[0,T]} |(N_P(t)/r) - t| \geq \epsilon\right) \leq 2\exp\left(-\frac{r\epsilon^2}{2T}\left(1 - \frac{\epsilon}{T}\right)\right).$$

Proof. The approach will be similar to finding a Chernoff (1952) Bound. Since $\exp(\alpha x)$ is convex for any positive constant α, and $N_P(t)$ is right continuous, $\exp(\alpha N_P(t))$ is a right continuous submartingale.

Let A_U denote the event that $(N_P(t)/r) - t > \epsilon$ for some $t \in [0, T]$. Then for all $\alpha > 0$:

$$\Pr(A_U) = \Pr\left(\sup_{t\in[0,T]} \exp(\alpha N_P(t)) \geq \exp(\alpha rt + \alpha r\epsilon)\right).$$

It follows from basic Markov-type inequalities on right continuous submartingales (p.13 of Karatzas and Shreve, 1991) that this probability can be upper bounded as

$$\Pr(A_U) \leq \mathrm{E}(\alpha \exp(N_P(T))/\exp(\alpha rT + \alpha r\epsilon)).$$

Using the moment generating function for a Poisson with parameter rT:

$$\mathrm{E}[\exp(\alpha N_P(T)) = \exp(rT(\exp(\alpha) - 1)),$$

which means

$$\Pr(A_U) \leq \exp(T(e^\alpha - 1 - \alpha) + \alpha\epsilon)^r.$$

A Taylor series expansion shows that $e^\alpha - 1 - \alpha \leq (\alpha^2/2)(1+\alpha)$ as long as $\alpha \in [0, 2.31858...]$. Set $\alpha = \epsilon/T$. Simplifying the resulting upper bound yields

$$\Pr(A_U) \leq \exp\left(-\frac{r\epsilon^2}{2T}\left(1 - \frac{\epsilon}{T}\right)\right).$$

The other tail can be dealt with in a similar fashion, yielding a bound

$$\Pr(\sup_{t\in[0,T]} [N_P(\alpha)/r] - t < \epsilon) \leq \exp\left(-\frac{r\epsilon^2}{2T}\right).$$

The union bound on the two tails then yields the theorem. □

Since the Poisson process operates in logspace, $T = \ln(\mu(B)/\mu(B'))$.

Corollary 2.1 *For $\epsilon \in (0, 0.3)$, $\delta \in (0, 1)$, and $\ln(\mu(B)/\mu(B')) > 1$, after*

$$r = 2(\ln(\mu(B)/\mu(B'))(3\epsilon^{-1} + \epsilon^{-2}) \ln(2/\delta)$$

runs of TPA, the points obtained can be used to build an (ϵ, δ) omnithermal approximation. That is,

$$\Pr((\forall \beta \in [\beta_{B'}, \beta_B])((1+\epsilon)^{-1} \leq \exp(N_P(\beta_B - \beta)/r)/[\mu(B)/\mu(A(\beta))] \leq 1+\epsilon)) < \delta.$$

Proof. In order for the final result to be within a multiplicative factor of $1+\epsilon$, in logspace the approximation must be accurate to an additive term of $\ln(1+\epsilon)$. Let $T = \ln(\mu(B)/\mu(B'))$, so $r = 2T(3\epsilon^{-1} + \epsilon^{-2}) \ln(2/\delta)$. To prove the corollary from the theorem, it suffices to show that $2\exp(-2T(3\epsilon^{-1} + \epsilon^{-2} \ln(2/\delta)[\ln(1+\epsilon)]^2(1-\epsilon/T)/(2T)) < \delta$. After canceling the factors of T, and noting that when $T > 1$, $1 - \epsilon/T < 1 - \epsilon$, it suffices to show that $(3\epsilon^{-1} + \epsilon^{-2})(1-\epsilon)[\ln(1+\epsilon)]^2 > 1$. This can be shown for $\epsilon \in (0, .3)$ by a Taylor series expansion. □

8. EXAMPLE: OMNITHERMAL APPROXIMATION FOR THE ISING MODEL

The Ising model falls into the broad class of automodels, spatial models where the distribution of a site conditioned on its neighbors comes from the same family (see Besag, 1974). For Ising, each node of a graph $G = (V, E)$ is assigned either 0 or 1 (hence it is an autobernoulli model.) In the simplest form of the model, the weight of a configuration $x \in \{0, 1\}^V$ is

$$\pi_{\text{Ising}}(x) = \frac{1}{Z_\beta} \exp(2\beta H(x)), \text{ where } H(x) = \sum_{\{i,j\} \in E} \mathbf{1}(x(i) = x(j)).$$

With models of this type, the function Z_β cannot be explicitly calculated for most graphs. At first glance, the problem appears simple: with a prior on the one dimensional parameter β, a one dimensional numerical integration should be easy. However, because the posterior density includes a Z_β^{-1} factor, in order to find the posterior, it is necessary to find Z_β.

Note that any (ϵ, δ) omnithermal approximation of Z_β will yield an (ϵ, δ) approximation for the integrated likelihood. Finding the posterior mean requires two integrals involving Z_β, and so the approximation for the posterior mean will be accurate to within a factor of $(1+\epsilon)^2$ with probability at least $1 - \delta$.

To obtain such an approximation, it is necessary to put the Ising model within the context of TPA. This is accomplished by introducing an auxiliary random variable Y, such that for $X \sim \pi_{\text{Ising}}$, $[Y|X] \sim \text{Un}([0, \exp(2\beta H(X))])$. This makes

$$Z_\beta = \mu(A(\beta)), \text{ where } A(\beta) = \{(x, y) : x \in \{0, 1\}^V, y \in [0, \exp(2\beta H(x))]\},$$

where μ is the direct product of counting measure on $\{0, 1\}^V$ and Lebesgue measure on $[0, \infty)$.

TPA operates as follows: Start with $\beta \leftarrow \beta_B$. Draw $X \leftarrow \pi_\beta$ and then $Y \leftarrow \text{Un}([0, \exp(2\beta H(X))])$. [Then the next value of β will be the value of β' such that $Y = \exp(2\beta H(X))$, so that $(X, Y) \in A(\beta')$ but not in any smaller set.] If $H(X) > 0$ set $\beta \leftarrow [\ln Y]/[2H(X)]$, else set $\beta \leftarrow 0$. Repeat until $\beta \leq 0$.

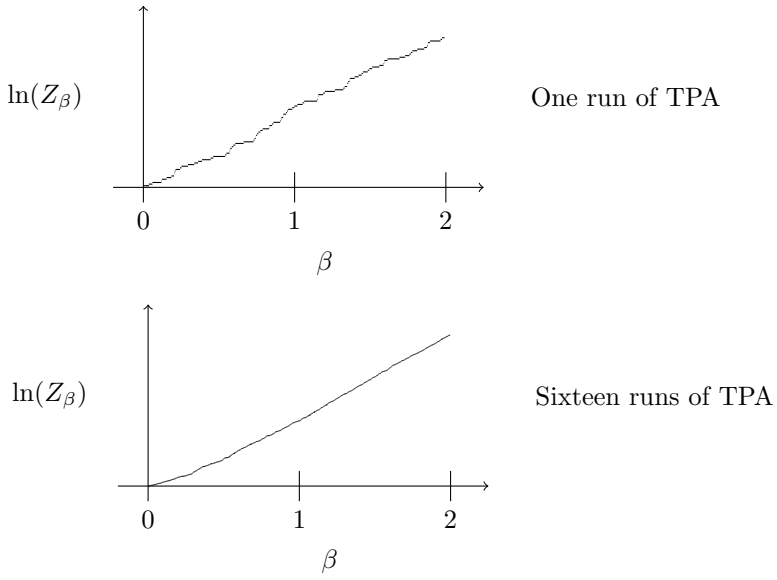

Figure 2: *Omnithermal approximations for the Ising model on a* 4×4 *lattice*

Figure 2 presents two omnithermal approximations for $\log Z_\beta$ generated using this method on a small 4×4 square lattice. The top graph is the result of a single run of TPA from $\beta = 2$ down to $\beta = 0$. At each β value returned by TPA, the approximation drops by 1. The bottom graph is the result of $\lceil \ln(4 \cdot 10^6) \rceil = 16$ runs of TPA. This run told us that $Z_2 \leq 217$ with confidence $1 - 10^{-6}/2$. Therefore, using $\epsilon = .1$, and $\delta = 10^6/2$ in Theorem 2 shows that $r = 330000$ samples suffice for a $(0.1, 10^{-6})$ omnithermal approximation.

9. DETERMINATION OF A COOLING SCHEDULE

The omnithermal approximation can then be used to build a nicely balanced deterministic cooling schedule. Consider the approximation of $\ln(Z_\beta)$ of the previous section, and let M denote the maximum value of $\ln(Z_\beta)$ over the region of interest. Then for i from 1 to d, let

$$\beta_d := \sup\{b : N_P(\beta_B - b) \geq M(i/d)\}.$$

This is illustrated in Figure 3, where logspace for the Ising model on a 4×4 lattice is partitioned into three equal parts, leading to a cooling schedule of length 4.

In general, partitioning logspace into d pieces yields a deterministic cooling schedule of length $d + 1$: $\beta_{B'} = \beta_d < \beta_{d-1} < \cdots < \beta_0 = \beta_B$. It is nicely balanced in the sense that for all i, $\ln(\mu(A(\beta_i))/\mu(A(\beta_{i+1}))) \approx \ln(\mu(B))/d$. In other words, the ratios $\mu(A(\beta_i))/\mu(A(\beta_{i+1}))$ are all roughly equal.

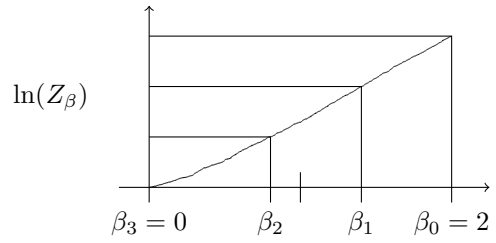

Figure 3: *Finding a cooling schedule from an omnithermal approximation*

Having such a schedule is important for many reasons, here we discuss two. First, this can be used to construct a new Markov chain, perhaps faster than the original that created the samples, using parallel tempering or some other technique based on a cooling schedule. If d is chosen to be near $\ln(\mu(B))$, then the schedule will be very well balanced in the sense that ratio of the measure of successive levels will be close to e^{-1} for all levels.

If the Markov chain is then modified by a multiplicative factor of e raised to the level, then the measure of successive levels will be roughly the same. Such chains with levels weighted towards equality have been shown to be fast in practice (see, for instance, Wang and Landau, 2001).

The second reason is that this allows for a product estimator approximate to be built. The product estimator goes back at least to self-reducibility algorithms of Jerrum et al. (1986), and operates by estimating $\mu(A(\beta_{i+1})/\mu(A(\beta_i))$ and then forming the estimator for $\mu(A(\beta_B))/\mu(A(\beta_{B'}))$ by taking the product of the estimates for the individual levels.

The advantage of TPA over the product estimator was the ability to analyze the tails of the distribution of the output without the need to have a balanced cooling schedule. However, once TPA creates such a cooling schedule, the product estimator can be used.

If bounding the probability mass in the tails is less important than restricting the standard deviation of the estimate, the product estimator can be preferable in some situations. This is because the product estimator can be partially derandomized, what is often referred to as *Rao-Blackwellization* of the procedure.

Consider the truncated likelihood approach of Section 2. Before derandomization, to estimate $\mu(A(T_i))/\mu(A(T_{i+1}))$, several samples would be drawn from $\mu(A(T_i))$ as a two stage process. In the first stage, draw θ, which has density proportional to $\min\{L(\theta \,|\, y), T_i\}$ with respect to the prior. In the second stage, draw auxiliary variable W that is uniform on $[0, \min\{L(\theta \,|\, y), T_i\}]$. Then count the percentage of the time the auxillary variable falls below $\min\{L(\theta \,|\, y), T_{i+1}\}$.

To Rao-Blackwellize this procedure, do not draw the final auxiliary variable. Instead record the probability the final auxiliary variable falls below $\min\{L(\theta \,|\, y), T_{i+1}\}$. That is, begin as before by drawing θ from $\min\{L(\cdot \,|\, y), T_i\}\, d\mu_{prior}(\cdot)$. Then let $f(\theta) = \min\{L(\theta \,|\, y), T_i\} / \min\{\mathrm{L}(\theta \,|\, y, T_{i+1}\}$. Then $f(\theta)$ is an unbiased estimate of the ratio of the measures of the two levels. That is, $\mathrm{E}[f(\theta)] = \mu(A(T_i))/\mu(A(T_{i+1}))$, and so using the sample mean of $f(\theta)$ over several draws gives an estimate for

$\mu(A(T_i))/\mu(A(T_{i+1}))$ that has lower variance than the original method.

The final estimate is then the product of the estimates for the ratios of each level, hence the name of the method: the product estimator.

10. COMPARISON TO NESTED SAMPLING

Since TPA involves the creation of a nested family of subsets to sample from, it naturally brings to mind the idea of Skilling (2006) known as nested sampling. There are some key differences, however.

- The nested sets in nested sampling are formed by considering the sets $\{w : L(w \mid y) > k\}$ for increasing values of k. As seen in Subsection 2.2, the nested sets used for TPA can be formed by considering $\{w : L(w \mid y) \leq T\}$ for some constant T. So if the likelihood is multimodel, by moving downward the extra modes are removed, making the problem easier as TPA progresses.

- In nested sampling, the accuracy of the final result depends on being able to sample near the maximum of the likelihood, hence the problem is typically as difficult as finding the posterior mode.

- However, it should be noted that the same method used in 2.2 to find the center for TPA with truncated likelihoods can also be used to find a suitable truncation value for the likelihood. By definition, the maximum of the truncated likelihood is known, so the error term arising from the unknown maximum can be eliminated in nested sampling.

- Nested sampling is a hybrid of a Monte Carlo and classical one dimensional numerical integration. This often reduces the error in practice, but theoretically introduces terms into the error bound that are usually unknown (related to the derivatives of unknown functions.) This means that the output can only be analyzed asymptotically. For TPA, it is possible to completely determine the distribution of the output, even for small problems.

11. FULLY BAYESIAN APPROXIMATION

The standard theoretical computer science definition of an (ϵ, δ) approximation algorithm is that the output must be within a factor of $1 + \epsilon$ of the true answer with probability at least $1 - \delta$. This is equivalent to saying that for output $\hat{p}$, $[(1 + \epsilon)^{-1}\hat{p}, (1 + \epsilon)\hat{p}]$ forms a $1 - \delta$ confidence interval for p. For TPA, $\hat{p}$ is the exponentiated maximum likelihood estimator for $\ln p$.

However, extra knowledge of the normalizing constant Z (and hence p) could come from something as simple as known bounds on Z in terms of dimension. Because the output distribution of TPA (and hence the likelihood) given $\ln Z$ can be written down explicitly, it is possible to conduct a fully Bayesian analysis of $\ln Z$ given the data generated by TPA. If extra information about $\ln Z$ is available, this can then be utilized to improve the estimate.

REFERENCES

J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc. B*, 36:192–236, 1974.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. of Math. Stat.*, 23:493–509, 1952.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* , 58(301):13–30, 1963.

M. Jerrum, L. Valiant, and V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43:169–188, 1986.

I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus (2nd Ed.)*. Springer, 1991.

Adrian E. Raftery, Michael A. Newton, Jaya M. Satagopan, and Pavel N. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. Technical Report 6, Memorial Sloan-Kettering Cancer Center, 2006. URL `http://www.bepress.com/mskccbiostat/paper6`.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods (2nd ed.)*. Springer, 2004.

J. Skilling. Nested Sampling for general Bayesian computation. *Bayesian Analysis* , 1(4):833–860, 2006.

F. Wang and D.P. Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Phys. Rev. E*, 64, 2001.