J. Appl. Prob. **51,** 92–105 (2014)
Printed in England

RANDOM CONSTRUCTION OF INTERPOLATING SETS FOR HIGH-DIMENSIONAL INTEGRATION

MARK HUBER,* *Claremont McKenna College*
SARAH SCHOTT,** *Duke University*

Abstract

Computing the value of a high-dimensional integral can often be reduced to the problem of finding the ratio between the measures of two sets. Monte Carlo methods are often used to approximate this ratio, but often one set will be exponentially larger than the other, which leads to an exponentially large variance. A standard method of dealing with this problem is to interpolate between the sets with a sequence of nested sets where neighboring sets have relative measures bounded above by a constant. Choosing such a well-balanced sequence can rarely be done without extensive study of a problem. Here a new approach that automatically obtains such sets is presented. These well-balanced sets allow for faster approximation algorithms for integrals and sums using fewer samples, and better tempering and annealing Markov chains for generating random samples. Applications, such as finding the partition function of the Ising model and normalizing constants for posterior distributions in Bayesian methods, are discussed.

Keywords: Integration; Monte Carlo method; cooling schedule; self-reducible

2010 Mathematics Subject Classification: Primary 60K35
Secondary 11K45; 65D30

1. Introduction

Monte Carlo methods for numerical integration can have enormous variance for the types of high-dimensional problems that arise in statistics and combinatorial optimization applications. Consider a state space Ω with measure μ, and $B \subset \Omega$ with finite measure. Then the problem considered here is approximating, for $B' \subset B$, the value of

$$A = \frac{\mu(B)}{\mu(B')}$$

to within a guaranteed level of relative error with a specified probability of success. This problem has a number of applications, including model selection in Bayesian statistics (A can be used to find the normalizing constant for a posterior distribution), approximation algorithms for $\#P$ complete problems, and likelihood functions for spatial statistics models (such as in [8]). Specific applications are discussed in Section 4.

The classical Monte Carlo approach is to create a random variable X such that $\mathbb{E}(X) = A$, where X has variance as small as possible. Unfortunately, it is often not possible to know the

Received 25 October 2012; revision received 24 February 2013.
* Postal address: Department of Mathematical Sciences, Claremont McKenna College, 850 Columbia Avenue, Claremont, CA 91711, USA. Email address: mhuber@cmc.edu
** Postal address: Department of Mathematics, Duke University, 117 Physics Building, Science Drive, Durham, NC 27708, USA. Email address: schott@math.duke.edu

variance of X ahead of time, and this must be estimated as well. How good the estimate of the variance is depends on even higher moments which are even more difficult to estimate, and so this approach typically cannot guarantee a bound on the error of the result.

The method presented here is entirely new. It obtains an estimate of A of the form $e^{X/k}$, where k is a known constant and X is a Poisson random variable with mean $k \ln(A)$. Because the mean and variance for a Poisson random variable are the same, we simultaneously obtain our estimate of A and knowledge of the variance of our estimate. In fact, the output from our method allows us to handle the following tasks.

- Estimate A to within a specified relative error with a specified failure probability using $O(\ln(A)^2)$ samples.
- Obtain a well-balanced sequence of nested sets useful in building annealing and tempering Markov chains that can be used to generate Monte Carlo samples.
- Develop an omnithermal approximation for partition functions arising from spatial point processes and Gibbs distributions.

The following definition makes the notion of a randomized approximation algorithm precise.

Definition 1. Let $\mathcal{I}$ be a set of problem inputs, $T: \mathcal{I} \to \mathbb{R}^+$ be the true answer to the problem, and $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then $\hat{T}: \mathcal{I} \times \Omega \times \mathbb{R}^+ \times (0, 1] \to \mathbb{R}^+$ is a $(1+\varepsilon, \delta)$-*randomized approximation algorithm* if, for all $I \in \mathcal{I}$, for all $\varepsilon > 0$, and for all $\delta \in (0, 1]$, we have

$$\mathbb{P}((1+\varepsilon)^{-1} T(\mathcal{I}) \le \hat{T}(\mathcal{I}, \omega, \varepsilon, \delta) \le (1+\varepsilon) T(\mathcal{I})) \ge 1 - \delta.$$

In other words, given a source of randomness to work with, the randomized approximation algorithm returns a result within a factor of $1+\varepsilon$ of the true answer with a probability of at least $1-\delta$. The new algorithm presented here is a $(1+\varepsilon, \delta)$-randomized approximation algorithm that, for $B' \subseteq B$, approximates $A = \mu(B)/\mu(B')$ using

$$2 \ln(4\delta^{-1}) \tilde{\varepsilon}^{-2} (1+\tilde{\varepsilon}) [\ln(A) + 1 + (\ln(A)+1)^2 (1-\tilde{\varepsilon})^{-1}] \tag{1}$$

samples on average, where $\tilde{\varepsilon} = \min\{\ln(1+\varepsilon), \frac{1}{2}\}$ (see Theorem 2, below, for the proof). Note that $\lim_{\varepsilon \to 0} \tilde{\varepsilon}/\varepsilon = 1$. This compares favorably with previous algorithms such as self-reducibility [9], which uses roughly 150 times as many samples to achieve the same result (see Section 2.1). A recent algorithm for approximating normalizing constants of Gibbs distributions by Štefankovič *et al.* [19] (here referred to as SVV) uses a number of samples that is nearly linear rather than quadratic in $\ln(\mu(B)/\mu(B'))$. However, the constant hidden by the big O notation is at least 10^{10}, making the algorithm impractical for real problems. We call the new algorithm presented here the Tootsie Pop Algorithm (TPA); there is a good reason behind the unusual name (see Section 3). It has several advantages over the self-reducibility and SVV approaches. It is easy to implement, requiring only a few lines of code. Yet the output can be analyzed precisely, giving rise to the number of samples given in (1).

The rest of the paper is organized as follows. Section 2 highlights previous work, and compares our new approach to these previous methods. Section 3 describes the TPA procedure in detail, then Section 4 shows some applications. Section 5 analyzes the expected running time of the method, and introduces a two-phase approach to TPA. Section 6 shows how to obtain an approximation that simultaneously works for all members of a continuous family of sets at once. Section 7 describes how TPA can be used to build well-balanced nested sets for tempering. Finally, Section 8 discusses further areas of exploration with TPA techniques.

https://doi.org/10.1239/jap/1395771416 Published online by Cambridge University Press

2. Previous work

The new method presented here follows a long line of work using interpolating sets. For instance, Valleau and Card [21] introduced what they called *multistage sampling* where an intermediate distribution was added to make estimation more effective, although they did not analyze precisely the behavior of the resulting algorithm. Jerrum *et al.* [9] used a similar idea of *self-reducibility*, and carefully analyzed the computational complexity of the resulting approximation method.

2.1. Self-reducibility

The self-reducible methodology of [9] can be viewed using the following framework, although originally the authors described their method in a very different fashion. Suppose that we are given two sets B' and B of finite measure such that $B' \subset B$ and $\mu(B')$ is known. Self-reducibility requires a sequence of sets that interpolate from B' up to B, i.e.

$$B' = B_\ell \subseteq B_{\ell-1} \subseteq B_{\ell-2} \subseteq \cdots \subseteq B_0 = B,$$

such that the relative measures of the sets $\mu(B_{i+1})/\mu(B_i) \geq \alpha$ for a fixed constant $\alpha \in (0, 1)$. Then an unbiased estimate $\hat{b}_i$ of $\mu(B_{i+1})/\mu(B_i)$ is obtained for each i. The product of these estimates will then be an unbiased estimator for $A = \mu(B)/\mu(B')$.

For fixed $\alpha \in (0, 1)$, it is easy to estimate $\mu(B_{i+1})/\mu(B_i)$ with small relative error simply by drawing samples from $\mu(B_i)$ and counting the percentage of samples that fall in $\mu(B_{i+1})$. The relative standard deviation of a Bernoulli random variable with parameter α is $(1-\alpha)/\alpha$, so it is important not to make α too small. On the other hand, if α is too large, then the nested sets are not shrinking much at each step, and it will require a lengthy sequence of such sets. To be precise, the number of sets ℓ must satisfy $\ell \geq \ln_\alpha(\mu(B')/\mu(B)) = \ln(\mu(B)/\mu(B'))/\ln(\alpha^{-1})$, which goes to infinity as α goes to 1. So there is an optimal value of α that balances these considerations (see the proof of Lemma 3, below). Dyer and Frieze [3] proved the following lemma, presented here in a form given in [19].

Lemma 1. ([19, Theorem 2.2].) *Let $W_1, \ldots, W_\ell$ be independent random variables with*

$$\frac{\mathbb{E}[W_i^2]}{\mathbb{E}[W_i]^2} \leq \alpha^{-1} \quad \text{for } i \in [\ell].$$

Let $\hat{W} = W_1 W_2 \cdots W_\ell$. Let S_i be the average of $16\alpha^{-1}\ell\varepsilon^{-2}$ independent random samples from W_i for $i \in [\ell]$. Let $\hat{S} = S_1 S_2 \cdots S_\ell$. Then

$$\mathbb{P}((1-\varepsilon)\mathbb{E}[\hat{W}] \leq \hat{S} \leq (1+\varepsilon)\mathbb{E}(\hat{W})) \geq \tfrac{3}{4}. \tag{2}$$

This is similar to the $(1+\varepsilon, \delta)$-randomized approximation algorithm with $\delta = \frac{1}{4}$. Since $(1-\varepsilon)\mathbb{E}[\hat{W}] \leq (1+\varepsilon)^{-1}\mathbb{E}[\hat{W}]$, the event in (2) is slightly weaker than that required by an $(1+\varepsilon, \frac{1}{4})$-randomized approximation algorithm. In other words, to use self-reducibility to construct a $(1+\varepsilon, \frac{1}{4})$ randomized approximation algorithm will require at least $16\alpha^{-1}\ell\varepsilon^{-2}$ samples.

The next lemma shows how to move from $\delta = \frac{1}{4}$ to an arbitrary $\delta > 0$.

Lemma 2. *With $\hat{S}$ as described in Lemma 1, let $k = \lceil \ln(\delta^{-1})(2)(\ln(4)-1)^{-1} + \frac{1}{2} \rceil$, let $\hat{S}_1, \ldots, \hat{S}_{2k-1}$ be $2k-1$ independent random variables each with the same distribution as $\hat{S}$, and let $\hat{S}_{(k)}$ be the median of these random variables. Then*

$$\mathbb{P}((1+\varepsilon)^{-1}\mathbb{E}[\hat{W}] \leq \hat{S}_{(k)} \leq (1+\varepsilon)\mathbb{E}[\hat{W}]) \geq 1-\delta.$$

https://doi.org/10.1239/jap/1395771416 Published online by Cambridge University Press

Proof. If at least half of $\{\hat{S}_1, \ldots, \hat{S}_{2k-1}\}$ are in (a, b), then the median will be as well. Let $N = \#\{i : \hat{S}_i \notin (a, b)\}$. Then N has a binomial distribution with parameters $2k - 1$ and $p \leq \frac{1}{4}$. From a Chernoff bound analysis (see, for instance, [13, Theorem 4.1]) we have

$$\mathbb{P}(N > 2\mathbb{E}[N]) = \mathbb{P}\left(N < \frac{(1+1)(2k-1)}{4}\right) \leq \left[\frac{\mathrm{e}}{2^2}\right]^{(2k-1)/4}.$$

To make this right-hand side less than δ, it is necessary that $k \geq \ln(\delta^{-1})(2)(\ln(4) - 1)^{-1} + \frac{1}{2}$.

The next lemma then states how well self-reducibility does in terms of the number of samples required.

Lemma 3. *With a sequence of B_i sets such that $\mu(B_{i+1})/\mu(B_i) = \alpha$, the number of random samples used in an $(1 + \varepsilon, \delta)$-randomized approximation algorithm with the self-reducible method is at least*

$$306 \ln(A)^2 \varepsilon^{-2} \ln(\delta^{-1}).$$

Proof. From Lemma 1, we know that $16\alpha^{-1}[\ln(\#B/\#B')/\ln(\alpha^{-1})]^2\varepsilon^{-2}$ samples are necessary, and then this procedure must be repeated $2\lceil \ln(\delta^{-1})(2)/(\ln(4) - 1) + \frac{1}{2}\rceil + 1 \geq 4/(\ln(4) - 1)\ln(\delta^{-1})$ times. The value for α that minimizes this expression is $1/\mathrm{e}^2$, giving a number of samples that is at least

$$\left[\frac{16\mathrm{e}^2(4)}{(\ln(4) - 1)/4}\right]\varepsilon^{-2}\ln(\delta^{-1}) > 306.048\varepsilon^{-2}\ln(\delta^{-1}).$$

In other words, even if the best conditions prevailed, and we were handed a sequence of B_i sets where the ratio from step to step was optimal or nearly optimal, a nuisance factor of over three hundred still appears in the number of samples. Of course, in practice we do not have such a sequence of sets available, and sequences that are constructed ad hoc are unlikely to be near the optimal α value. Because such sets are likely to be close to, but not match exactly, a fixed value of α, a well-balanced sequence can only require that they come close to this value.

Definition 2. For fixed constants $0 < \alpha_1 < \alpha_2 < 1$, a sequence of sets $B' = B_\ell \subseteq \cdots \subseteq B_0 = B$ where the ratios $\mu(B_{i+1})/\mu(B_i)$ fall in $[\alpha_1, \alpha_2]$ for all i is *well balanced*.

Well-balanced sequences have uses other than self-reducibility. The various methods of designing Markov chains such as simulated annealing [11], simulated tempering, and parallel tempering [5], [12], [20], all require such a sequence of well-balanced sets in order to mix rapidly (see [22] and [23]).

2.2. Gibbs distributions

Given a finite state space Ω and function $H : \Omega \to \mathbb{R}$, a *Gibbs distribution* is a family of probability distributions indexed by a parameter β, where, for $x \in \Omega$,

$$\pi(\{x\}) = \frac{\exp(-\beta H(x))}{Z(\beta)},$$

where $Z(\beta) = \sum_{y\in\Omega} \exp(-\beta H(y))$ is a normalizing constant known as the partition function. For $\beta_{B'} < \beta_B$, let $A = Z(\beta_B)/Z(\beta_{B'})$.

For the case where $H(x) \in \{0, 1, \ldots, n\}$, Štefankovič *et al.* [19] gave an algorithm (SVV) that yields an $(1 + \varepsilon, \delta)$-randomized approximation algorithm for A that used at least

$$10^{10}\varepsilon^{-2}\ln(A)(\ln(n) + \ln(\ln(A)))^5 \ln(\delta^{-1})$$

samples ([19, Corollary 1.5]).

https://doi.org/10.1239/jap/1395771416 Published online by Cambridge University Press

As shown in Section 4.1, it is possible to obtain such an approximation using TPA with the average number of samples given in (1). While TPA is $O(\ln(A)^2)$ and SVV is $O(\ln(A)\ln(\ln(A))^5)$, the respective constants of 2 and 10^{10} mean that TPA is the logical choice to use even for very large problems. Also, $\ln(A) < \ln(\ln(A))^5$ up to around $\ln(A) \approx 332\,000$.

Naturally, the best outcome would be to merge the TPA and SVV ideas: work in this direction is ongoing by the authors [6].

2.3. Nested sampling

A special case is when approximating A is equivalent to approximating the normalizing constant of a posterior distribution of a Bayesian analysis. Skilling [18] introduced *nested sampling* as a way of generating a random sequence of nested sets. Unlike with self-reducibility, there is no need to have the sequence of sets in hand ahead of time. Instead, it builds up sets from scratch at random according to a well-defined procedure.

Unfortunately, nested sampling also has several disadvantages. The first is that it loses the property of self-reducible algorithms that the variance of the output can be bounded prior to running the algorithm. Because deterministic numerical integration was used in the method, the error can be determined only up to a factor that depends upon the derivatives of a function that is difficult to compute. Therefore, nested sampling falls in the class of methods where the variance must be estimated, rather than bounded ahead of time as with self-reducibility.

Perhaps more importantly, the sets B_i will become disconnected when nested sampling is applied to a multimodal distribution. Therefore, the method is really only appropriate for the fairly restrictive case of unimodal posterior distributions.

2.4. Summary of relationship to previous work

The relationship between these ideas and TPA can be summarized as follows. Nested sampling [18] does not give a randomized approximation algorithm. Self-reducibility is the most general: it can handle the widest variety of problems. A subset of these problems can be handled by TPA. On these problems TPA should be used since it is about 150 times as fast as self-reducibility. The problem of approximation for Gibbs distributions is a special case of problems handled by TPA. For these problems SVV is theoretically faster but in practice much slower than TPA. A simple example of a problem that can be handled by TPA which is not a Gibbs distribution is the normalizing constant of a posterior distribution where the family of sets are truncated posteriors. This application is presented in Section 4.2.

3. The Tootsie Pop Algorithm

The new method presented here is called *The Tootsie Pop Algorithm* (TPA), and combines the best features of the self-reducibility and nested sampling approach. Like self-reducibility, it is very general, working over a wide variety of problems. This includes the nested sampling domain of Bayesian posterior normalization, but also includes many other problems where self-reducibility has been applied such as the Ising model. Portions of this work were presented at the Ninth Valencia International Meetings on Bayesian Statistics, and appeared in the conference proceedings of that meeting [7] with a discussion.

The name is somewhat unusual, and references an advertising campaign run for Tootsie Pop candies. A Tootsie Pop is a chocolate chewy center surrounded by a candy shell. The advertisement campaign asked 'How many licks does it take to get to the center of a Tootsie Pop?'. Our algorithm operates in a similar fashion. Our set B is slowly whittled away until the center B' is reached. The number of steps taken to move from B to B' will be Poisson with

mean $\ln(A)$, thereby allowing approximation of A. Therefore, the 'number of licks' is exactly what is needed to form our estimate!

TPA has four general ingredients.

1. A measure space $(\Omega, \mathcal{F}, \mu)$.
2. Two finite measurable sets B and B' satisfying $B' \subset B$ and $\mu(B') > 0$. The set B' is the *center* and B is the *shell*.
3. A family of nested sets $\{A(\beta)\colon \beta \in \mathbb{R} \cup \{\infty\}\}$ such that $\beta' < \beta$ implies $A(\beta') \subseteq A(\beta)$, $\mu(A(\beta))$ is a continuous function of β, and the limit of $\mu(A(\beta))$ as β goes to $-\infty$ is 0.
4. Special values β_B and $\beta_{B'}$ that satisfy $A(\beta_B) = B$ and $A(\beta_{B'}) = B'$.

With these ingredients, the TPA method is very simple to describe.

1. Start with $i = 0$ and $\beta_i = \beta_B$.
2. Draw a random variable Y from μ conditioned to lie in $A(\beta_i)$.
3. Let $\beta_{i+1} = \inf\{\beta\colon Y \in A(\beta)\}$.
4. If $Y \in B'$ stop and output i.
5. Else set i to be $i + 1$ and go back to step 2.

Another way of describing the draw in step 2 is that, for measurable D,

$$\mathbb{P}(Y \in D) = \mu(D \cap A(\beta_i))/\mu(A(\beta_i)).$$

At each step, the set $A(\beta_i)$ shrinks in measure with probability 1, and so is slowly worn away until the sample falls into the region B'.

Step 2 deserves special attention. Drawing a random sample Y from μ conditioned to lie in $A(\beta_i)$ is in general a very difficult problem. The good news is that the importance of this problem means that a vast literature for solving this problem exists. Markov chain Monte Carlo (MCMC) methods are critical to obtaining these samples, and variations on the early methods have blossomed over the last fifty years. Readers are referred to [4], [16], [17], and the references therein for more information.

Of course, any other method for turning samples into approximations either implicitly or explicitly depends on the ability to execute some variant of step 2 as well, so our algorithm is not actually demanding anything above and beyond what other algorithms in this area require. The algorithm is easily modified to handle different methods of simulating random variables. For instance, nested sampling [18] can be run so that it draws several such Y variables in parallel, and TPA can be written to do so as well.

The key fact about this process is given in the following result.

Theorem 1. *At any step of the algorithm, let*

$$E_i = \ln(\mu(A(\beta_i))) - \ln(\mu(A(\beta_{i+1}))).$$

Then the E_i are independent and identically distributed exponential random variables with mean 1.

https://doi.org/10.1239/jap/1395771416 Published online by Cambridge University Press

Proof. To simplify the notation, let $m(\beta) = \mu(A(\beta))$ and fix $\beta_i \geq \beta_{B'}$. Suppose that Y is drawn from μ conditioned to lie in $A(\beta_i)$ and $\beta_{i+1} = \inf\{\beta : Y \in A(\beta)\}$. Let $U_i = m(\beta_{i+1})/m(\beta_i)$. Since $\beta_{i+1} \leq \beta_i$, $U_i \in [0, 1]$.

We now show that U_i has a uniform distribution over $[0, 1]$. Fix $a \in (0, 1)$. Then, since m is a continuous function with $\lim_{\beta \to -\infty} m(\beta) = 0$, there exists a $b \in (-\infty, \beta_i]$ such that $m(b)/m(\beta_i) = a$. If $Y \in A(b)$ then $\beta_{i+1} \leq b$ and $U_i \leq m(b)/m(\beta_i) = a$. Hence, $\mathbb{P}(U_i \leq a) \geq \mathbb{P}(Y \in A(b)) = a$.

Let n be any positive integer, as before there is a value b_n such that $m(b_n) = a + 1/n$. If $Y \notin A(b_n)$, then $\inf\{\beta : Y \in A(\beta)\} \geq a + 1/n$ and $U_i \geq a + 1/n$. Hence, if $U_i \leq a$, then $Y \in A(b_n)$ and $\mathbb{P}(U_i \leq a) \leq \mathbb{P}(Y \in A(b_n)) = a + 1/n$. Since n was arbitrary, this means that $\mathbb{P}(U_i \leq a) \leq a$. Combining with $\mathbb{P}(U_i \leq a) \geq a$ gives that U_i is uniformly distributed on $[0, 1]$.

Finally, since U_i is uniform over $[0, 1]$, $-\ln(U_i) = \ln(\mu(A(\beta_i))) - \ln(\mu(A(\beta_{i+1})))$ is an exponential random variable with mean 1, completing the proof.

For $t(\beta) = \ln(\mu(A(\beta)))$, Theorem 1 says that the values of $t(\beta_i) - t(\beta_{i+1})$ in a run of TPA are exponential random variables of mean 1, so $\{t(\beta_1), \ldots, t(\beta_i)\}$ forms a homogeneous Poisson point process on $[t(\beta_{B'}), t(\beta_B)]$ of rate 1.

3.1. Taking advantage of Poisson point processes

Poisson point processes have several nice properties that will be of use to us. First, consider the total number of points used by a run of TPA, that is, the value of i at the end of the algorithm. Because the $t(\beta_i)$ values form a Poisson point process, the distribution of i is Poisson with mean $t(\beta_B) - t(\beta_{B'}) = \ln(\mu(B)/\mu(B')) = \ln(A)$.

Furthermore, the union of k independent Poisson point processes of rate 1 is also a Poisson point process of rate k. Therefore, after k runs of TPA, the distribution of the total number of samples used is still Poisson, but with a mean of $k \ln(A)$.

4. Applications

The following examples illustrate some of the applications of TPA.

4.1. The Ising model

The Ising model is an example of a Gibbs distribution as described in Section 2.2.

In the Ising model, begin with a graph $G = (V, E)$ with $\#V$ nodes and $\#E$ edges. Each node of a graph $G = (V, E)$ is assigned one of two values, -1 and 1. In order to use TPA with Gibbs distributions, it is easiest to first write the model so that $H(x) \geq 0$.

For the Ising model, this can be accomplished by setting $H(x) = 1 + \#E - \sum_{i<j} x(i)x(j)$. Note that adding a constant (such as $1 + \#E$) to $H(x)$ does not change the distribution.

In order to embed this problem in the framework of TPA, add one auxiliary dimension to the $\#V$ dimensional configuration x. The new state space is then

$$\Omega_{\text{aux}}(\beta) = \{(x, y) : x \in \{0, 1\}^V, y \in [0, \exp(\beta H(x))]\}.$$

Note the following observations on $\Omega_{\text{aux}}(\beta)$.

- The state space is a collection of one dimensional line segments in $\#V + 1$ dimensional space.
- The total length of the line segments in $\Omega_{\text{aux}}(\beta)$ is just $Z(\beta)$. That is to say, $\mu(\Omega_{\text{aux}}(\beta)) = Z(\beta)$ where μ is the one dimensional Lebesgue measure of the union of the line segments.

https://doi.org/10.1239/jap/1395771416 Published online by Cambridge University Press

- Let $\beta' < \beta$. Then, since $H(x) > 0$, $\Omega_{\text{aux}}(\beta') \subset \Omega_{\text{aux}}(\beta)$. Moreover, $Z(\beta)$ is a continuous function that goes to 0 as $\beta \to -\infty$. Therefore, Condition 2 of the TPA ingredients (see Section 3) is satisfied.
- For $\beta = 0$, $y \in [0, 1]$ for all $x \in \{0, 1\}$. That means $Z(0) = 2^V$.
- Let $\beta > 0$. Then $\Omega_{\text{aux}}(\beta)$ is the shell, and $\Omega_{\text{aux}}(0)$ is the center.

With this in mind, the TPA algorithm works as follows.

1. Start with $i = 0$ and $\beta_0 = \beta$.
2. Draw a random sample X from π_{β_i}, then draw Y (given X) uniformly from $[0, \exp(\beta_i H(X))]$.
3. Let $\beta_{i+1} = \ln(Y)/H(X)$.
4. If $\beta_{i+1} \leq 0$ stop and output i.
5. Else set i to be $i + 1$ and go back to step 2.

The expression in step 3 uses the fact that $\ln(Y)/H(X) = \inf\{b\colon \exp(bH(X)) = Y\}$. One run of TPA will require on average $1 + \ln(Z(\beta)/Z(0)) = 1 + \ln(Z(\beta)) - \#V \ln(2)$ samples from the model with different values of β.

This simple method of adding a single auxiliary variable allows TPA to be used on a variety of discrete distributions by changing the measure to one that varies continuously in the index.

4.2. Posterior distributions

In Bayesian analysis, often it is necessary to find the normalizing constant of a posterior distribution. This is known as the *evidence* for a model, and can be written

$$Z = \int_{x \in \Omega} f(x)\, \mathrm{d}x,$$

where $f(x)$ is a nonnegative density (the product of the prior density and the likelihood of the data) and $\Omega \subseteq \mathbb{R}^n$.

For a point $c \in \Omega$ and $\varepsilon > 0$, let $B_\varepsilon^1(c)$ be the points within L_1 distance ε of c. Suppose that, for a particular c and ε, $B_\varepsilon^1(c) \subset \Omega$ and there is a known M such that $\frac{1}{2}M \leq f(x) \leq M$ for all $x \in B_\varepsilon^1(c)$.

Then to estimate $Z(\varepsilon) = \int_{x \in B_\varepsilon^1(c)} f(x)\, \mathrm{d}x$, draw N independent and identically distributed samples $X_1, \ldots, X_N$ uniformly from $B_\varepsilon(c)$, and let the estimate be

$$\hat{Z}(\varepsilon) = (2\varepsilon)^{-n} \sum_i \frac{f(X_i)}{N}.$$

Then $\hat{Z}(\varepsilon)$ is an unbiased estimate for $Z(\varepsilon)$ with standard deviation bounded above by $Z(\varepsilon)/\sqrt{N}$.

Now we connect this problem to TPA. The family of sets is $\{A(\beta) = B_\beta^1(c) \cap \Omega\}$, and the measure is $\mu(A(\beta)) = \int_{x \in A(\beta)} f(x)\, \mathrm{d}x$. The shell will be $A(\infty)$ (so $Z = \mu(A(\infty))$) and the center $A(\varepsilon)$ (with measure $Z(\varepsilon)$). TPA can then be used to estimate $Z/Z(\varepsilon)$, and the estimate of $Z(\varepsilon)$ can then finish the job.

https://doi.org/10.1239/jap/1395771416 Published online by Cambridge University Press

5. Running time of TPA

Suppose that TPA is run k times, and the k values of the i variable at the end of each run are summed together. Call this sum N. Then N has a Poisson distribution with mean $k\ln(\mu(B)/\mu(B'))$. This makes N/k an unbiased estimate of $\ln(\mu(B)/\mu(B'))$. The variance of N/k is $\ln(\mu(B)/\mu(B'))/k$.

Let W be a normal random variable of mean 0 and variance 1, and W_α be the inverse cumulative distribution function of W so that $\mathbb{P}(W \le W_\alpha) = \alpha$. Then the normal approximation to the Poisson gives

$$\left[\frac{N}{k} - W_{\alpha/2}\sqrt{\frac{N}{k}}, \frac{N}{k} + W_{\alpha/2}\sqrt{\frac{N}{k}}\right]$$

as an approximately $1-\alpha$ level confidence interval for $\ln(\mu(B)/\mu(B'))$. Exponentiating then gives the $1-\alpha$ level for $\mu(B)/\mu(B')$.

For a specific output, it is also possible to build an exact confidence interval for $\mu(B)/\mu(B')$ since the distribution of the output is known exactly.

Similarly, it is easy to perform a Bayesian analysis and find a credible interval given a prior on $\ln(\mu(B)/\mu(B'))$.

But we are interested in more than confidence intervals! Consider how to build a $(1+\varepsilon, \delta)$-*randomized approximation scheme* whose output $\hat{A}$ satisfies

$$\mathbb{P}((1+\varepsilon)^{-1}A \le \hat{A} \le (1+\varepsilon)A) > 1-\delta.$$

From Section 3, the output of one run of TPA is a Poisson random variable with mean $\ln(A)$. It is well known that the sum of independent Poisson random variables is another Poisson random variable. Hence, after running TPA k times, the result is a Poisson random variable with mean $k\ln(A)$. The following lemma gives a bound on the tails of the Poisson distribution.

Lemma 4. *Let $a > 0$ and N be a Poisson random variable with mean $k\ln(A)$. Then*

$$\mathbb{P}\left(\left|\frac{N}{k} - \ln(A)\right| \ge a\right) \le 2\exp\left(-\frac{ka^2}{2[\ln(A)+a]}\right).$$

(This result follows from Lemma 5, below.)

To obtain our $(1+\varepsilon, \delta)$-randomized approximation algorithm, we would like to make $a = \ln(1+\varepsilon)$, and set k so that $2\exp(-ka^2/(2[\ln(A)+a])) \le \delta$. In other words, we wish to set k to be something like $2\ln(1+\varepsilon)^{-2}\ln(A)(1+\ln(1+\varepsilon)\ln(2\delta^{-1}))$. But $\ln(A)$ is unknown at the start of the algorithm!

There are many ways around this difficulty, here we use a two-phase method. First obtain a rough estimate of $\ln(A)$, then refine this estimate to the level demanded by ε.

Phase I. Let $\tilde{\varepsilon} = \min\{\ln(1+\varepsilon), \frac{1}{2}\}$ and $k_1 = 2\ln(4\delta^{-1})\varepsilon^{-2}(1+\tilde{\varepsilon})$. Then let N_1 be the sum of the outputs from k_1 runs of TPA.

Phase II. Set $k_2 = (N_1 + k_1)(1-\tilde{\varepsilon})^{-1}$. Let N_2 be the sum of the outputs from k_2 runs of TPA. The final estimate is $\hat{A} = \exp(N_2/k_2)$.

Note that $\tilde{\varepsilon} = \min\{\ln(1+\varepsilon), \frac{1}{2}\}$ is equivalent to requiring $\varepsilon \le \exp(\frac{1}{2}) - 1 \approx 0.648\,7$. This is necessary to ensure that $(1-\tilde{\varepsilon})^{-1}$ is bounded above.

After Phase I has been run, N_1/k_1 estimates $\ln(A)$ to within an additive error of $\tilde{\varepsilon}\ln(A)$ with a probability of at least $1-\delta/2$. Phase II then uses the Phase I estimate of $\ln(A)$ to obtain the better estimate N_2/k_1 of $\ln(A)$ to within an additive error of $\tilde{\varepsilon}$.

https://doi.org/10.1239/jap/1395771416 Published online by Cambridge University Press

Theorem 2. *The output $\hat{A}$ of the above procedure is a $(1+\varepsilon, \delta)$-randomized approximation scheme for $A = \mu(B)/\mu(B')$. The running time is random, with the expected number of samples bounded above by*

$$2\ln(4\delta^{-1})\tilde{\varepsilon}^{-2}(1+\tilde{\varepsilon})[\ln(A)+1+(\ln(A)+1)^2(1-\tilde{\varepsilon})^{-1}].$$

Proof. Call Phase I a success if N_1/k_1 is within a distance of $\tilde{\varepsilon}(\ln(A)+1)$ of $\ln(A)$. From Lemma 4, with $a = \tilde{\varepsilon}\ln(A)$, we have

$$\begin{aligned}\mathbb{P}\left(\left|\frac{N_1}{k_1} - \ln(A)\right| \geq \tilde{\varepsilon}(\ln(A)+1)\right) &\leq 2\exp\left(-\frac{k_1(\ln(A)+1)^2\tilde{\varepsilon}^2}{2(\ln(A)+\tilde{\varepsilon}(\ln(A)+1))}\right)\\ &\leq 2\exp\left(-\frac{k_1\tilde{\varepsilon}^2(\ln(A)+1)}{2(1+\tilde{\varepsilon})}\right)\\ &\leq \frac{\delta}{2}.\end{aligned}$$

Therefore, the probability that Phase I is a failure is at most $\delta/2$.

When Phase I is a success,

$$N_1 \geq k_1(\ln(A) - \tilde{\varepsilon}(\ln(A)+1)),$$

so, for $k_2 = (N_1+k_1)(1-\tilde{\varepsilon})^{-1}$, we have $k_2 \geq [\ln(A)+1]k_1$. Substituting this into Lemma 4, with $\alpha = \tilde{\varepsilon}$, yields

$$\mathbb{P}\left(\left|\frac{N_2}{k_2} - \ln(A)\right| \geq \tilde{\varepsilon}\right) \leq 2\exp\left(-\frac{[\ln(A)+1]2\tilde{\varepsilon}^{-2}(1+\tilde{\varepsilon})\ln(4\delta^{-1})\tilde{\varepsilon}^2}{2[\ln(A)+\tilde{\varepsilon}]}\right).$$

Since $(\ln(A)+1)(1+\tilde{\varepsilon}) > \ln(A)+\tilde{\varepsilon}$, the probability is at most $2\exp(-\ln(4\delta^{-1})) = \delta/2$.

That means the chance of failure in either phase is at most $\delta/2+\delta/2 = \delta$, so altogether $|(N_2/k_2) - \ln(A)| \leq \tilde{\varepsilon}$ with a probability of at least $1-\delta$. Since $\hat{A} = \exp(N_2/k_2)$,

$$\left\{-\tilde{\varepsilon} \leq \frac{N_2}{k_2} - \ln(A) \leq \tilde{\varepsilon}\right\} \quad \Longleftrightarrow \quad \{(1+\varepsilon)^{-1}A \leq \hat{A} \leq (1+\varepsilon)A\},$$

and we have our $(1+\varepsilon, \delta)$-randomized approximation algorithm.

The expected number of samples needed in Phase I is $k_1[\ln(A)+1]$, while the expected number needed in Phase II is

$$\mathbb{E}[N_1+k_1](1-\tilde{\varepsilon})^{-1}[\ln(A)+1] = (k_1\ln(A)+k_1)(1-\tilde{\varepsilon}^{-1})^{-1}[\ln(A)+1].$$

Summing the Phase I and II average samples gives a total bounded above by

$$2\ln(4\delta^{-1})\tilde{\varepsilon}^{-2}(1+\tilde{\varepsilon})[\ln(A)+1+(\ln(A)+1)^2(1-\tilde{\varepsilon})^{-1}].$$

6. Omnithermal approximation

Suppose that, instead of just a single value of interest $\mu(B)/\mu(B')$, it is necessary to obtain an approximation of $\mu(A(\beta))/\mu(B')$ that is valid for all values $\beta \in [\beta_{B'}, \beta_B]$ simultaneously. We will call this an *omnithermal approximation*. These problems appear in what are called

https://doi.org/10.1239/jap/1395771416 Published online by Cambridge University Press

doubly intractable posterior distributions arising in Bayesian analyses involving spatial point processes. They are usually dealt with indirectly using Markov chain Monte Carlo with auxiliary variables [14], but an omnithermal approximation allows for a more direct approach.

In Section 3 it was shown that the $t(A(\beta_i))$ values form a Poisson point process P on the interval $[0, \ln(A)]$. To move from P to a Poisson process, set

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

As t advances 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 with rate k.

Given $N_P(t)$, approximate $\mu(B)/\mu(A(\beta))$ by $\exp(N_P(\beta_B - \beta)/k)$. When $\beta = \beta_{B'}$, this is just the approximation given earlier, so this generalizes the description of TPA from before.

The key fact is that $N_P(t) - kt$ is a right-continuous martingale. To bound the error in $\exp(N_P(t)/k)$, it is necessary to bound the probability that $N_P(t) - kt$ has drifted too far away from 0 for $t \in [0, \ln(A)]$.

Lemma 5. *Let $\tilde{\varepsilon} > 0$. Then, for $N_P(\cdot)$, a rate k Poisson process on $[0, \ln(A)]$, we have*

$$\mathbb{P}\left(\sup_{t\in[0,\ln(A)]} \left|\frac{N_P(t)}{k} - t\right| \geq \tilde{\varepsilon}\right) \leq 2\exp\left(-\frac{k\tilde{\varepsilon}^2}{2[\ln(A) + \tilde{\varepsilon}]}\right).$$

Proof. The approach will be similar to finding a Chernoff bound [2]. Since $\exp(\alpha x)$ is convex for any positive constant α, and $N_P(t)$ is a right-continuous martingale, $\exp(\alpha N_P(t))$ is a right-continuous submartingale.

Let A_U denote the event that $(N_P(t)/k) - t > \varepsilon$ for some $t \in [0, \ln(A)]$. Then, for all $\alpha > 0$,

$$\mathbb{P}(A_U) = \mathbb{P}\left(\sup_{t\in[0,\ln(A)]} \exp(\alpha N_P(t)) \geq \exp(\alpha k t + \alpha k \varepsilon)\right).$$

It follows from basic Markov-type inequalities on right-continuous submartingales [10, p. 13] that this probability can be upper bounded for all $\alpha > 0$, i.e.

$$\mathbb{P}(A_U) \leq \mathbb{E}\left(\frac{\exp(\alpha N_P(\ln(A)))}{\exp(\alpha k \ln(A) + \alpha k\tilde{\varepsilon})}\right).$$

Using the moment generating function for a Poisson with parameter $k\ln(A)$, we obtain

$$\mathbb{E}[\exp(\alpha N_P(\ln(A)))] = \exp(k\ln(A)(\exp(\alpha) - 1)),$$

which means

$$\mathbb{P}(A_U) \leq \exp(\ln(A)(e^{\alpha} - 1 - \alpha) - \alpha\tilde{\varepsilon})^k.$$

This is true for any $\alpha > 0$, so we choose $\alpha = \ln(1 + \tilde{\varepsilon}/\ln(A))$ which minimizes the right-hand side. After simplifying,

$$\mathbb{P}(A_U) \leq \exp\left(\tilde{\varepsilon} + (\ln(A) + \tilde{\varepsilon})\ln\left(1 - \frac{\tilde{\varepsilon}}{\ln(A) + \tilde{\varepsilon}}\right)\right)^k.$$

Recalling that, for $\gamma \geq 0$, $\ln(1 - \gamma) \leq -\gamma - \gamma^2/2$ allows us to further bound the right-hand side as follows:

$$\mathbb{P}(A_U) \leq \exp\left(\frac{-k\tilde{\varepsilon}^2}{2(\ln(A) + \tilde{\varepsilon})^2}\right).$$

https://doi.org/10.1239/jap/1395771416 Published online by Cambridge University Press

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

$$\mathbb{P}\left(\sup_{t\in[0,\ln(A)]} \frac{N_P(\alpha)}{k} - t \le \tilde{\varepsilon}\right) \le \exp\left(\frac{-k\tilde{\varepsilon}^2}{2\ln(A)}\right).$$

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

Theorem 3. *The two-phase TPA algorithm generates a* $(1+\varepsilon, \delta)$*-omnithermal approximation using a number of samples (on average) bounded by*

$$2\ln(4\delta^{-1})\tilde{\varepsilon}^{-2}(1+\tilde{\varepsilon})[\ln(A) + 1 + (\ln(A)+1)^2(1-\tilde{\varepsilon})^{-1}].$$

Proof. The proof is the same as for Theorem 2, using Lemma 5 instead of Lemma 4.

6.1. Example: omnithermal approximation for the Ising model

Consider the following model. The value of β is drawn from a prior density $f_{\text{prior}}(\cdot)$ on $[0, \infty)$, and then the data (conditioned on β) is drawn from the Ising model. This was used by Besag [1] as a model for agriculture wherein soil quality of adjacent plots was more likely to be similar.

Given the data X, the Bayesian posterior density for β is

$$f_{\beta,\text{post}}(b) \propto f_{\beta,\text{prior}}(b)\frac{\exp(-bH(X))}{Z(b)}. \tag{3}$$

The *evidence* for the model is the integral of the right-hand side of (3) as b runs from 0 to ∞. This is only a one-dimensional integration, and so should be straightforward from a numerical perspective, except that $Z(b)$ is unknown.

Here is where the omnithermal approximation comes in: it gives an approximation for $Z(b)$ that is valid for *all* values of b at once. Any numerical integration technique can be used, and the final value for the evidence (not including error arising from the numerical method) will be within a factor of $1+\varepsilon$ of the true answer.

Figure 1 presents two omnithermal approximations for $\log Z_\beta$ generated using this method on a 50×50 square lattice. Part (a) is the result of a single run of TPA from $\beta = 0.5$ down

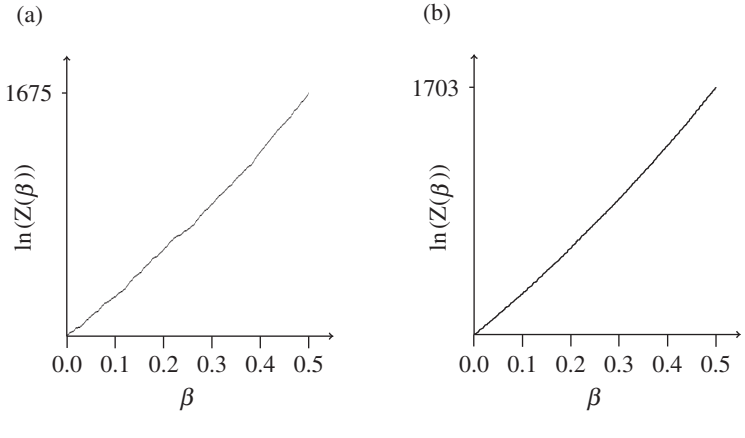

FIGURE 1: Omnithermal approximations for the partition function of the Ising model on a 50×50 square lattice. Part (a) is the result of one run of TPA; part (b) is the result of 16 runs.

https://doi.org/10.1239/jap/1395771416 Published online by Cambridge University Press

to $\beta = 0$. At each β value returned by TPA, the approximation drops by 1. Since the graph drops 1 675 times, the best estimate for $z(0.5)$ is 1 675. Part (b) is the result of runs of TPA, and gives the slightly better estimate of $z(0.5) \approx 1\,703$. Part (b) is smoother, and contains 27 254 β values compared to only 1 675 β values for the one run. The samples used within TPA came from coupling from the past [15] applied to the random scan Gibbs chain.

Note that methods such as nested sampling [18] only produce an estimate of the partition function at a single value of β rather than for β in an interval as with TPA.

7. Well-balanced nested sets

Once a user has the ability to obtain an omnithermal approximation to the $\mu(A(\beta))$ curve, finding a well-balanced set of β_i, as in Definition 2, is easy.

Fix $0 < \alpha_1 < \alpha_2 < 1$ as in Definition 2. Then let $\varepsilon = \alpha_2^{1/4}\alpha_1^{-1/4} - 1$ so that $(1+\varepsilon)^2 = \alpha_2^{1/2}\alpha_1^{-1/2}$. Now find an omnithermal approximation for $\mu(A(\beta))$ over $[\beta_{B'}, \beta_B]$. Call the approximation $\hat{\mu}(A(\beta))$. With a probability of at least $1 - \delta$, for all $\beta \in [\beta_{B'}, \beta_B]$, we have $\mu(A(\beta))$ within a factor of $1 + \varepsilon$ of $\hat{\mu}(A(\beta))$.

Now find values $\beta_{B'} = \beta_\ell < \beta_{\ell-1} < \cdots < \beta_1 < \beta_0 = \beta_B$ such that

$$\frac{\hat{\mu}(A(\beta_i))}{\hat{\mu}(A(\beta_{i-1}))} = (\alpha_2\alpha_1)^{1/2},$$

for all $i \in \{1, \ldots, \ell\}$.

Since the $\hat{\mu}$ function is within a factor of $1 + \varepsilon$ of the μ function, for any $i \in \{1, \ldots, \ell\}$, we have

$$(1+\varepsilon)^{-2}\alpha_2^{1/2}\alpha_1^{1/2} \le \frac{\mu(A(\beta_i))}{\mu(A(\beta_{i-1}))} \le (1+\varepsilon)^2\alpha_2^{1/2}\alpha_1^{1/2},$$

and from our choice of ε,

$$\alpha_1 \le \frac{\mu(A(\beta_i))}{\mu(A(\beta_{i-1}))} \le \alpha_2,$$

making the schedule well balanced.

8. Conclusions and further work

The strength of TPA is the generality of the procedure, but that same generality means that it is possible to do better in restricted circumstances. For instance, when $f(x)$ falls into the class of Gibbs distributions, Štefankovič *et al.* [19] were able to give an $\tilde{O}(\ln(Z))$ algorithm for approximating Z, but the high constants involved in their algorithm make it solely of theoretical interest. (Here the $\tilde{O}$ notation hides factors of $\ln(\ln(Z))$.) TPA can be used in conjunction with their algorithm [6] to build an $O(\ln(Z)\ln(\ln(Z)))$ algorithm, and work continues to bring this running time down to $O(\ln(Z))$.

Acknowledgements

An earlier version of this work was presented at the 2010 ISBA World Meeting. The authors wish to thank the discussants to the proceedings paper [7], Gareth Roberts, Christian Robert, Nicholas Chopin, Ian Murray, and John Skilling, for valuable comments and suggestions.

Both authors were supported in this work by NSF CAREER grant DMS-05-48153. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

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

[2] Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23,** 493–507.

[3] Dyer, M. and Frieze, A. (1991). Computing the volume of convex bodies: a case where randomness provably helps. In *Probabilistic Combinatorics and Its Applications* (Proc. Symp. Appl. Math. **44**), American Mathematical Society, Providence, RI, pp. 123–169.

[4] Fishman, G. S. (1994). Choosing sample path length and number of sample paths when starting in steady state. *Operat. Res. Lett.* **16,** 209–219.

[5] Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 156–163.

[6] Huber, M. L. (2012). Approximating algorithms for the normalizing constant of Gibbs distributions. Preprint. Available at http:/uk.arxiv.org/abs/1206.2689.

[7] Huber, M. and Schott, S. (2011). Using TPA for Bayesian inference. In *Bayesian Statistics 9* (Proc. 9th Valencia Internat. Meeting), Oxford University Press, pp. 257–282.

[8] Huber, M. L. and Wolpert, R. L. (2009). Likelihood-based inference for Matérn type-III repulsive point processes. *Adv. Appl. Prob.* **41,** 958–977.

[9] Jerrum, M. R., Valiant, L. G. and Vazirani, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.* **43,** 169–188.

[10] Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*, 2nd edn. Springer, New York.

[11] Kirkpatrick, S., Gelatt, C. D., Jr. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220,** 671–680.

[12] Marinari, E. and Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* **19,** 451–458.

[13] Motwani, R. and Raghavan, P. (1995). *Randomized Algorithms*. Cambridge University Press.

[14] Murray, I., Ghahramani, Z. and MacKay, D. J. C. (2006). MCMC for doubly-intractable distributions. In *Proc. 22nd Annual Conf. Uncertainty Artificial Intelligence*, AUAI Press, pp. 359–366.

[15] Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9,** 223–252.

[16] Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd edn. Springer, New York.

[17] Shonkwiler, R. W. and Mendivil, F. (2009). *Expolorations in Monte Carlo Methods*. Springer, New York.

[18] Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis* **1,** 833–859.

[19] Štefankovič, D., Vempala, S. and Vigoda, E. (2009). Adaptive simulated annealing: a near-optimal connection between sampling and counting. *J. ACM* **56,** 36pp.

[20] Swendsen, R. H. and Wang, J.-S. (1986). Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.* **57,** 2607–2609.

[21] Valleau, J. P. and Card, D. N. (1972). Monte Carlo estimation of the free energy by multistage sampling. *J. Chem. Phys.* **57,** 5457–5462.

[22] Woodward, D. B., Schmidler, S. C. and Huber, M. (2009). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Ann. Appl. Prob.* **19,** 617–640.

[23] Woodward, D. B., Schmidler, S. C. and Huber, M. (2009). Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electron. J. Prob.* **14,** 780–804.