The Annals of Applied Probability
2018, Vol. 28, No. 2, 1099–1135
https://doi.org/10.1214/17-AAP1326

THE SAMPLE SIZE REQUIRED IN IMPORTANCE SAMPLING

BY SOURAV CHATTERJEE[1] AND PERSI DIACONIS[2]

Stanford University

The goal of importance sampling is to estimate the expected value of a given function with respect to a probability measure ν using a random sample of size n drawn from a different probability measure μ. If the two measures μ and ν are nearly singular with respect to each other, which is often the case in practice, the sample size required for accurate estimation is large. In this article, it is shown that in a fairly general setting, a sample of size approximately $\exp(D(\nu \parallel \mu))$ is necessary and sufficient for accurate estimation by importance sampling, where $D(\nu \parallel \mu)$ is the Kullback–Leibler divergence of μ from ν. In particular, the required sample size exhibits a kind of cut-off in the logarithmic scale. The theory is applied to obtain a general formula for the sample size required in importance sampling for one-parameter exponential families (Gibbs measures).

1. Theory. Let μ and ν be two probability measures on a set $\mathcal{X}$ equipped with some sigma-algebra. Suppose that ν is absolutely continuous with respect to μ. Let ρ be the probability density of ν with respect to μ. Let $X_1, X_2, \ldots$ be a sequence of $\mathcal{X}$-valued random variables with law μ. Let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function. Suppose that our goal is to evaluate the integral

$$I(f) := \int_{\mathcal{X}} f(y)\, d\nu(y).$$

The *importance sampling estimate* of this quantity based on the sample $X_1, \ldots, X_n$ is given by

$$I_n(f) := \frac{1}{n} \sum_{i=1}^{n} f(X_i)\rho(X_i).$$

Sometimes, when the probability density ρ is known only up to a normalizing constant, that is, $\rho(x) = C\tau(x)$ where τ is explicit but C is hard to calculate, and the following alternative estimate is used:

$$J_n(f) := \frac{\sum_{i=1}^{n} f(X_i)\tau(X_i)}{\sum_{i=1}^{n} \tau(X_i)}. \tag{1.1}$$

Received November 2015; revised June 2017.
[1]Supported in part by NSF Grant DMS-1441513.
[2]Supported in part by NSF Grant DMS-1208775.
MSC2010 subject classifications. 65C05, 65C60, 60F05, 82B80.
Key words and phrases. Importance sampling, Monte Carlo methods, Gibbs measure, phase transition.

It is easy to see that

$$\mathbb{E}\big(I_n(f)\big) = \int_{\mathcal{X}} f(x)\rho(x)\,d\mu(x) = \int_{\mathcal{X}} f(y)\,d\nu(y).$$

Therefore, the expected value of $I_n(f)$ is the quantity $I(f)$ that we are trying to estimate. However, $I_n(f)$ may have large fluctuations. The two main problems in importance sampling are: (a) given μ, ν and f, to determine the sample size required for getting a reliable estimate, and (b) given ν and f, to find a sampling measure μ that minimizes the required sample size among a given class of measures. We address the first problem in this paper.

A straightforward approach for computing an upper bound on the required sample size is to compute the variance of $I_n(f)$. Indeed, this is easy to compute:

$$\begin{aligned}\operatorname{Var}\big(I_n(f)\big) &= \frac{1}{n}\left(\int_{\mathcal{X}} f(x)^2\rho(x)^2\,d\mu(x) - I(f)^2\right)\\ &= \frac{1}{n}\left(\int_{\mathcal{X}} f(y)^2\rho(y)\,d\nu(y) - I(f)^2\right).\end{aligned} \tag{1.2}$$

The formula for the variance can be used, at least in theory, to calculate a sample size that is sufficient for guaranteeing any desired degree of accuracy for the importance sampling estimate. In practice, however, this number is often much larger than what is actually required for good performance.

Sometimes the variance formula (1.2) is estimated using the simulated data $X_1, \dots, X_n$. This estimate is known as the empirical variance. There is an inherent unreliability in using the empirical variance to determine convergence of importance sampling. We will elaborate on this in Section 2.

We begin by stating our main theorems. Proofs are collected together in Section 4. A literature review on importance sampling is given at the end of this Introduction.

There are three main results in this article. The first theorem, stated below, says that under a certain condition that often holds in practice, the sample size n required for $|I_n(f) - I(f)|$ to be close to zero with high probability is roughly $\exp(D(\nu \parallel \mu))$ where $D(\nu \parallel \mu)$ is the Kullback–Leibler divergence of μ from ν. More precisely, it says that if s is the typical order of fluctuations of $\log \rho(Y)$ around its expected value, then a sample of size $\exp(D(\nu \parallel \mu) + O(s))$ is sufficient and a sample of size $\exp(D(\nu \parallel \mu) - O(s))$ is necessary for $|I_n(f) - I(f)|$ to be close to zero with high probability. The necessity is proved by considering the worst possible f, which as it turns out, is the function that is identically equal to 1.

An immediate concern that the reader may have is that $|I_n(f) - I(f)| \approx 0$ may not always be the desired criterion for convergence. If $I(f)$ is very small, then one may want to have $I_n(f)/I(f) \approx 1$ instead. A necessary and sufficient condition for this, when f is the indicator of a rare event, is given in Theorem 1.3 later in this section.

THEOREM 1.1. *Let $\mathcal{X}$, μ, ν, ρ, f, $I(f)$ and $I_n(f)$ be as above. Let Y be an $\mathcal{X}$-valued random variable with law ν. Let $L = D(\nu \parallel \mu)$ be the Kullback–Leibler divergence of μ from ν, that is,*

$$L = D(\nu \parallel \mu) = \int_{\mathcal{X}} \rho(x) \log \rho(x)\, d\mu(x) = \int_{\mathcal{X}} \log \rho(y)\, d\nu(y) = \mathbb{E}\big(\log \rho(Y)\big).$$

Let $\|f\|_{L^2(\nu)} := (\mathbb{E}(f(Y)^2))^{1/2}$. If $n = \exp(L + t)$ for some $t \geq 0$, then

$$\mathbb{E}\big|I_n(f) - I(f)\big| \leq \|f\|_{L^2(\nu)} \big(e^{-t/4} + 2\sqrt{\mathbb{P}\big(\log \rho(Y) > L + t/2\big)}\big).$$

Conversely, let 1 denote the function from $\mathcal{X}$ into $\mathbb{R}$ that is identically equal to 1. If $n = \exp(L - t)$ for some $t \geq 0$, then for any $\delta \in (0, 1)$,

$$\mathbb{P}\big(I_n(1) \geq 1 - \delta\big) \leq e^{-t/2} + \frac{\mathbb{P}(\log \rho(Y) \leq L - t/2)}{1 - \delta}.$$

Note that Theorem 1.1 does not just give the sample size required to ensure that $I_n(f)$ is close to $I(f)$ in the L^1 sense; the second part of the theorem implies that if we are below the sample size prescribed by Theorem 1.1, then for $f \equiv 1$, there is a substantial chance that $I_n(f)$ is actually *not close* to $I(f)$. Such lower bounds cannot be given merely by moment estimates. For example, lower bounds on moments like $\mathbb{E}|I_n(f) - I(f)|$ and $\mathrm{Var}(I_n(f))$ imply nothing; $I_n(f)$ may be close to $I(f)$ with high probability and yet $\mathbb{E}|I_n(f) - I(f)|$ and $\mathrm{Var}(I_n(f))$ may be large. The second part of Theorem 1.1 gives an actual lower bound on the sample size required to ensure that $I_n(f)$ is close to $I(f)$ with high probability, and the first part shows that this lower bound matches a corresponding upper bound. It is interesting that the sample size required for small L^1 error turns out to be the actual correct sample size for good performance.

As shown later in this section, it is fairly common that $\log \rho(Y)$ is concentrated around its expected value in large systems. In this situation, a sample of size roughly $\exp(D(\nu \parallel \mu))$ is both necessary and sufficient.

The second main result of this article, stated below, gives the analogous result for the estimate $J_n(f)$. The conclusion is essentially the same.

THEOREM 1.2. *Let all notation be as in Theorem 1.1 and let $J_n(f)$ be the estimate defined in (1.1). Suppose that $n = \exp(L + t)$ for some $t \geq 0$. Let*

$$\varepsilon := \big(e^{-t/4} + 2\sqrt{\mathbb{P}\big(\log \rho(Y) > L + t/2\big)}\big)^{1/2}.$$

Then

$$\mathbb{P}\bigg(\big|J_n(f) - I(f)\big| \geq \frac{2\|f\|_{L^2(\nu)}\varepsilon}{1 - \varepsilon}\bigg) \leq 2\varepsilon.$$

Conversely, suppose that $n = \exp(L - t)$ for some $t \geq 0$. Let $f(x)$ denote the function from $\mathcal{X}$ into $\mathbb{R}$ that is equal to 1 when $\log \rho(x) \leq L - t/2$ and 0 otherwise. Then $I(f) = \mathbb{P}(\log \rho(Y) \leq L - t/2)$ but $\mathbb{P}(J_n(f) \neq 1) \leq e^{-t/2}$.

Sometimes importance sampling is used to estimate the probabilities of rare events under the target measure ν. Typically, the quantity of interest is $\nu(A)$, where A is a rare event under ν but is not a rare event under μ. The method of estimation is the same as before, that is, let $1_A(x)$ be the function that is 1 if $x \in A$ and 0 otherwise, and let $I_n(1_A)$ be the importance sampling estimate of $\nu(A)$. The difference with the previous setting is that when estimating $\nu(A)$, we are not satisfied if $|I_n(1_A) - \nu(A)|$ is small because $\nu(A)$ itself is a small number. Rather, it is satisfactory if the ratio $I_n(1_A)/\nu(A)$ is close to 1. It turns out that the sample size that is necessary and sufficient for this purpose is not $\exp(D(\nu \parallel \mu))$, but $\exp(D(\nu_A \parallel \mu))$, where ν_A is the probability measure ν conditioned on the event A. This is quantified by the following theorem, which is the third main result of this paper.

THEOREM 1.3. *Let all notation be as in Theorem 1.1. Let A be any event such that $\nu(A) > 0$ and let 1_A be the indicator function of A, defined above. Let ν_A be the measure ν conditioned on the event A, that is, for any event B,*

$$\nu_A(B) := \frac{\nu(A \cap B)}{\nu(A)}.$$

Let $\rho_A(x) := \rho(x)1_A(x)/\nu(A)$ be the probability density function of ν_A with respect to μ. Let $L_A := D(\nu_A \parallel \mu)$. If $n = \exp(L_A + t)$ for some $t \geq 0$, then

$$\mathbb{E}\left|\frac{I_n(1_A)}{\nu(A)} - 1\right| \leq e^{-t/4} + 2\sqrt{\mathbb{P}\big(\log \rho_A(Y) > L_A + t/2 \mid Y \in A\big)}.$$

Conversely, suppose that $n = \exp(L_A - t)$ for some $t \geq 0$. Then for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(\frac{I_n(1_A)}{\nu(A)} \geq 1 - \delta\right) \leq e^{-t/2} + \frac{\mathbb{P}(\log \rho_A(Y) \leq L_A - t/2 \mid Y \in A)}{1 - \delta}.$$

We would like to remark here that the upper bounds in Theorems 1.1, 1.2 and 1.3 may not be tight. The only purpose of these theorems is to give matching upper and lower bounds on the sample size required for good performance of importance sampling. No attempt was made to get optimal error bounds, especially of the type that is relevant to practitioners.

Another remark is that in practice, μ is chosen depending on ν, to minimize the required sample size. One potential use for our theorems is that they may be used to choose μ by minimizing the Kullback–Leibler divergence of μ from ν among some class of candidate measures. This point is elaborated in the literature review at the end of this section.

Sometimes, however, μ is chosen depending on both ν and f. Since Theorems 1.1 and 1.2 give bounds that depend only on the $L^2(\nu)$ norm of f, they will not be useful for choosing μ using fine properties of f. This is particularly problematic if f is something like the indicator of a rare event. This issue is partially

addressed in Theorem 1.3, where $f = 1_A$ for some rare event A, and the required sample size depends on μ, ν and the event A. Therefore, Theorem 1.3 can be used for choosing μ depending on properties of both ν and f.

Let us now investigate the implications of our theorems in a few simple examples. More complex examples are given in later sections.

EXAMPLE 1.4 (Binomial distributions). Let $\mu = \text{Binomial}(N, p)$ and $\nu = \text{Binomial}(N, r)$, where $r > p$. Then

$$\log \rho(x) = x \log \frac{r}{p} + (N - x) \log \frac{1-r}{1-p}.$$

Let $Y \sim \nu$. Then $L = \mathbb{E}(\log \rho(Y)) = NH(r, p)$, where

$$H(r, p) = r \log \frac{r}{p} + (1 - r) \log \frac{1-r}{1-p}.$$

Moreover, the standard deviation of $\log \rho(Y)$ is of order $\sqrt{N}$. Thus, the required sample size is $\exp(NH(r, p) + O(\sqrt{N}))$. On the other hand, a simple calculation shows that if variance is used to determine sample size, the required size would be $\exp(NV(r, p))$, where

$$V(r, p) = \log\left(\frac{r^2}{p} + \frac{(1-r)^2}{1-p}\right).$$

By Jensen's inequality, $V(r, p) \geq H(r, p)$. Figure 1 shows that graph of $H(r, p)$ versus the graph of $V(r, p)$, as r varies and p is fixed at $1/2$. This elementary ex-

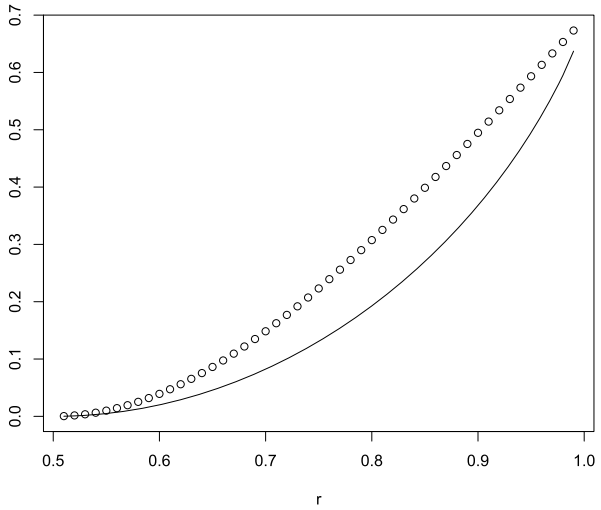

FIG. 1. *Let H and V be as in Example 1.4. The dotted line represents $V(r, p)$ and the solid line represents $H(r, p)$. Here, $p = 0.5$ and r goes from 0.5 to 1 on the x-axis.*

ample demonstrates how using the variance can lead to unnecessarily large sample sizes.

EXAMPLE 1.5 (Directed paths). Let $\mathcal{X}$ be the set of all monotone paths from $(0,0)$ to (n,n) in the two-dimensional lattice. Here, paths are only allowed to go up and to the right. The target measure is the uniform distribution on all such paths. Clearly, $|\mathcal{X}| = \binom{2n}{n}$. The sampling measure μ in this example constructs a random path γ as follows (this is known as sequential importance sampling): Choose one of the two directions "up" or "right" with probability $1/2$ until the walk hits the top or right side of the $n \times n$ "box," when the remainder of the walk is forced. If $T(\gamma)$ is the first time the path hits the top or right side, then

$$\mu(\gamma) = 2^{-T(\gamma)}.$$

Both the uniform distribution $\nu(\gamma) = 1/\binom{2n}{n}$ and $\mu(\gamma)$ have the property that, conditional on $T(\gamma) = j$, the paths are uniformly distributed. Thus, distributional questions are determined by the distribution of $T(\gamma)$.

The following proposition from Bassetti and Diaconis [4] shows that under the sampling distribution μ, $T(\gamma)$ is usually about $O(\sqrt{n})$ from the maximum possible $2n-1$, but under the uniform distribution ν, $T(\gamma)$ is usually about $O(1)$ away from $2n-1$.

PROPOSITION 1.6. *With the notation above*:

(a) *Under the importance sampling distribution* μ,

$$\mu\{T(\gamma) = j\} = 2^{1-j}\binom{j-1}{n-1}, \qquad n \le j \le 2n-1.$$

(b) *For* n *large and fixed positive* x,

$$\mu\left\{\frac{2n-1-T(\gamma)}{\sqrt{n}} \le x\right\} \sim \frac{1}{\pi}\int_0^x e^{-y^2/4}\,dy.$$

(c) *Under the uniform distribution* ν,

$$\nu\{T(\gamma) = j\} = \frac{2\binom{j-1}{n-1}}{\binom{2n}{n}}, \qquad n \le j \le 2n-1.$$

Further $\mathbb{E}_\nu(T(\gamma)) = (2 - \frac{2}{n+1})n$.

(d) *For* n *large and any fixed* k,

$$\nu\{T(\gamma) = 2n-1-k\} \sim \frac{1}{2^{k+1}}, \qquad 0 \le k < \infty.$$

The quantity L of Theorem 1.1 is determined from $\rho(\gamma) = \nu(\gamma)/\mu(\gamma)$ as

$$\begin{aligned} L &= \sum_{\gamma} \nu(\gamma) \log \frac{\nu(\gamma)}{\mu(\gamma)} \\ &= -\log \binom{2n}{n} + \frac{\log 2}{\binom{2n}{n}} \sum_{\gamma} T(\gamma) \\ &= -\log \binom{2n}{n} + \left(2 - \frac{2}{n+1}\right) n \log 2 \\ &= \log \sqrt{\pi n} - 2 \log 2 + O\left(\frac{1}{n}\right). \end{aligned}$$

Thus, $e^L \sim \sqrt{\pi n}/4$, and moreover, $\log \rho(\gamma)$ has fluctuations of order 1 around its mean. Thus, a sample size of order $\sqrt{n}$ is necessary and sufficient for accuracy of importance sampling in this example. The sufficiency was already observed using variance computations in Bassetti and Diaconis [4]; the necessity is a new result. Similar computations can be carried out for paths allowed to go left or right or up (staying self avoiding) using results of Bousquet-Mélou [12].

EXAMPLE 1.7 (Estimating the probability of a rare event). As an example for Theorem 1.3, fix N and $p > 1/2$ and let $A = \{j : Np \le j \le N\}$. Take ν to be the Binomial$(N, 1/2)$ distribution. Let $b(A; N, 1/2)$ be the probability of A under ν. Estimating $b(A; N, 1/2)$ by simple sampling from ν would be a crazy task; for example, when $N = 100$ and $p = 0.9$, $b(A; 100, 1/2) \approx 0.676049 \times 10^{-45}$, which means that we would need roughly 10^{45} samples to directly estimate this probability. A standard importance sampling approach (Siegmund [54]) is to sample $X_1, X_2, \ldots, X_n$ from $\mu = \text{Binomial}(N, \theta)$ for some θ and use

$$I_n(A) = \frac{1}{n} \sum_{i=1}^{n} \frac{\nu(X_i)}{\mu(X_i)} 1_A(X_i).$$

Theorem 1.3 shows that this will be accurate in ratio for n of order e^{L_A}. The following proposition shows that when μ is Binomial(N, θ), $\theta = p$ minimizes e^{L_A}, agreeing with the variance minimization in Siegmund [54]. When $N = 100$ and $p = 0.9$, $e^{L_A} \approx 1.723 \times 10^{28}$ (still an impossible sample size, but much smaller than 10^{45}).

PROPOSITION 1.8. *Fix N and $p > 1/2$ such that Np is an integer. Let μ be the* Binomial(N, θ) *distribution, ν be the* Binomial$(N, 1/2)$ *distribution and $A = \{j : Np \le j \le N\}$. Then the quantity L_A of Theorem 1.3 is asymptotically minimized when $\theta = p$, and with this choice of θ, L_A is asymptotic to $-2N \log(p^p(1-p)^{1-p})$ as $N \to \infty$.*

Review of the literature. Our interest in this topic started with a question from our colleague Don Knuth in [37]. He used sequential importance sampling to generate random self-avoiding paths starting at $(0, 0)$ and ending at (N, N) in a two-dimensional $N \times N$ grid. For $N = 10$, he calculated the number of paths (about 1.6×10^{24}), the average path length (92 ± 5) and the proportion of paths passing through $(5, 5)$ $(81\% \pm 10\%)$. He noticed huge fluctuations along the way and wanted to know about the accuracy of his estimates. In the follow-up work Knuth [38], exact computation showed surprising accuracy for his example. Basseti and Diaconis [4] and Bousquet-Mélou [12] studied toy versions of Knuth's problem where exact calculations can be done; they confirm the extreme variability and make the accuracy observed mysterious.

In our work, the choice of the proposal measure μ is considered fixed. A good deal of the art of successful implementation of importance sampling consists in a careful choice of μ, adapted to the problem under study. This is often done to minimize the variance of the resulting estimate. Our work, especially the main result of Section 2, suggests that the variance is a poor measure of accuracy for these long tailed problems. Thus, there is work to be done, exploring ways of adapting the many good ideas below, based on the variance, to minimizing the Kullback–Leibler divergence.

Any book on simulation will treat importance sampling. We recommend Hammersley and Handscomb [29], Srinivasan [55], Cappé, Moulines and Rydén [13] and Liu [40]. To begin our review of the research literature, a classical choice of the sampling measure μ for estimating $I(f) = \int f \, d\nu$ is to take $d\mu(x)$ proportional to $|f(x)| \, d\nu(x)$ (Kahn and Marshall [33]). Hesterberg [30] suggests using a mixture of measures for μ with one component proportional to $|f(x)|$ near its maximum. This is closely related to the widely used method of umbrella sampling (Torrie and Valleau [57]; nicely developed in Madras [42]). Owen and Zhou [48] combine Hesterberg's idea with control variates to give an attractive, practical approach. In later work, Owen and Zhou [47] suggest an adaptive version, attempting to improve the proposed μ using previous sampling. This is based on the empirical variance which means that our laments in Section 2 apply.

The idea of using L^1 distance to measure performance of importance sampling has appeared in a few prior instances. Two notable examples are Owen [49] and Owen [50], where L^1 error was used to compare the Monte Carlo and quasi-Monte Carlo approaches to estimating singular integrands via importance sampling.

Importance sampling is often used to do rare event simulation. Then it is natural to tilt the sampling distribution μ toward the region of interest. Siegmund [54] gives an asymptotically principled approach to doing this, which has given rise to much follow-up work, some of it quite deep mathematically. A unifying account of a variety of importance sampling algorithms for simulating the maxima of Gaussian fields appears in Shi, Siegmund and Yakir [53]. A host of novel ways of building importance sampling estimates for problems such as estimating the size of the union of a collection of sets when the size of each is known is in Naiman

and Wynn [46]. The work of Paul Dupuis with many coauthors is notable here. Dupuis and Wang [25] and Dupuis, Spiliopoulos and Wang [24] are representative papers with useful pointers to an extensive literature. Asmussen and Glynn [2] give a textbook account of this part of the subject.

An important part of the literature adapts importance sampling from the case of independent proposals considered here to use with a Markov chain generating proposals. Madras and Piccioni [43] give a clear development as do the textbook accounts of Robert and Casella [51] or Liu [40].

An important class of techniques for building proposal distributions is known as sequential importance sampling. An early appearance of this to sampling self-avoiding paths occurs in Rosenbluth and Rosenbluth [52]. For contingency table examples, see Chen, Diaconis, Holmes and Liu [17]. For degree sequences of graphs, see Blitzstein and Diaconis [11]. For time series and a general review, see the textbook by Doucet, de Freitas and Gordon [23] or the survey of Chen and Liu [18].

A relatively recent technique choosing the proposal distribution, which has been particularly successful in the heavy-tailed setting, is a method based on Lyapunov functions developed by Blanchet and Liu [9, 10], Blanchet and Glynn [7] and Blanchet, Glynn and Leder [8].

One large related topic is the connection between importance sampling and particle filters. Roughly, when building a proposal μ sequentially, one begins with a number N of starts. As the proposals are independently built up, some weights may be much larger than others. One can generate N new proposals from the present ones (say with probability proportional to weights). This will replicate some proposals and kill of those with smaller weights. This resampling can be repeated several times. The final weighted samples are used, in the usual way, to form importance sampling estimates. This large enterprise can be surveyed in the textbooks of Del Moral [19, 20] and Doucet, de Freitas and Gordon [23]. Work of Chan and Lai [14, 15] harnesses martingale central limit theorems to get the limiting distribution of these importance sampling methods in a variety of complex stochastic models. The web page of Arnaud Doucet is extremely useful. A very clear recent paper is Del Moral, Kohn and Patras [21].

Besides the broad classifications outlined above, importance sampling has a variety of other applications that are harder to categorize. A recent example is the paper by Efron [26] that suggests the use of importance sampling for generating from Bayesian posterior distributions. In this context, an interesting note is that simulating from a Bayesian posterior by rejection sampling was investigated by Freer, Mansinghka and Roy [27], who found a connection with the Kullback–Leibler divergence that bears some similarities with the results of this paper.

Two other recent papers have similarities with our work. One is that of Hult and Nyquist [32], who analyze the performance of importance sampling in the estimation of probabilities of rare events using large deviation techniques. The Kullback–Leibler divergence arises naturally in this work, due to its appearance in large

deviation rate functions. The other is a paper of Agapiou, Papaspiliopoulos, Sanz-Alonso and Stuart [1], who prove that $|I_n(f) - I(f)|$ is small if $n \geq \mathbb{E}(\rho(Y)) \geq e^L$, in the notation of our Theorem 1.1. This result is applied to a class of problems that do not overlap with our set of examples, making [1] and this paper complementary to each other.

2. Testing for convergence. The theory developed in Section 1, while theoretically interesting, is possibly not very useful from a practical point of view. Determining $D(\nu \parallel \mu)$ requires in-depth knowledge of not only the measure μ, but also the usually much more complicated measure ν. It is precisely the lack of understanding about ν that motivates importance sampling, so it seems pointless to ask a practitioner to compute the required sample size by using properties of ν.

To determine whether the importance sampling estimate has converged, a common practice is to estimate $\mathrm{Var}(I_n(f))$ by estimating the variance formula (1.2) using the data from μ. One natural estimate is

$$v_n(f) := \frac{1}{n^2} \sum_{i=1}^{n} f(X_i)^2 \rho(X_i)^2 - \frac{I_n(f)^2}{n}.$$

If this estimate is used, then importance sampling is declared to have converged if for some n, $v_n(f)$ turns out to be smaller than some pre-specified tolerance threshold ε (see Robert and Casella [51]).

The following theorem shows that using $v_n(f)$ as a diagnostic for convergence of importance sampling is problematic, because for any given tolerance level ε, there is high probability that the test declares convergence at or before a sample size that depends only on ε and not on μ, ν or f. This is absurd, since convergence may take arbitrarily long, depending on the problem.

THEOREM 2.1. *Given any $\varepsilon > 0$, there exists $n \leq \varepsilon^{-2} 2^{1+\varepsilon^{-3}}$ such that the following is true. Take any μ and ν as in Theorem 1.1, and any $f : \mathcal{X} \to \mathbb{R}$ such that $\|f\|_{L^2(\nu)} \leq 1$. Let $v_n(f)$ be defined as above. Then $\mathbb{P}(v_n(f) < \varepsilon) \geq 1 - 4\varepsilon$.*

Although the upper bound on n is very large, for example, for $\varepsilon = 0.1$ the upper bound is roughly 2.14×10^{303}, Theorem 2.1 gives a conceptual proof that using $v_n(f)$ for testing convergence of importance sampling is fundamentally flawed. As the measures μ and ν get more and more singular with respect to each other (which often happens as system size gets larger), importance sampling should take longer to converge. A test that does not respect this feature cannot be a plausible test for convergence. Incidentally, it is not clear whether the upper bound on n in Theorem 2.1 can be improved to something more reasonable.

The ineffectiveness of the variance diagnostic is not hard to demonstrate in examples. One such examples are given below.

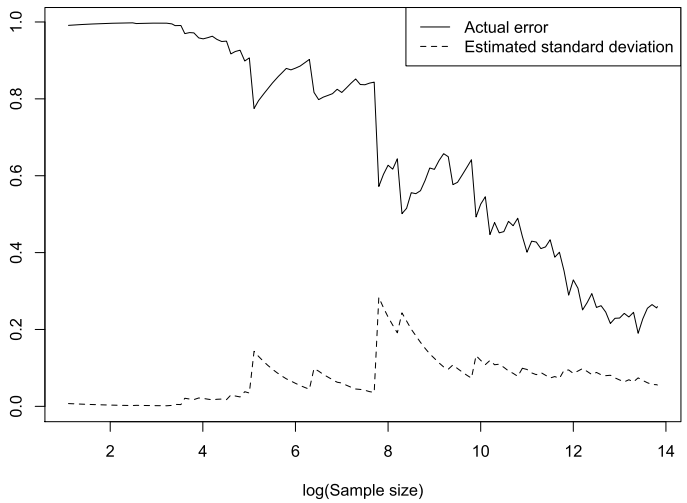

FIG. 2. *Estimated standard deviation* $\sqrt{v_n(f)}$ *of* $I_n(f)$, *as* n *ranges from* 1 *to* 10^6, *and the actual error* $|I_n(f) - I(f)|$. *Here,* $\mu = \text{Binomial}(100, 0.5)$, $\nu = \text{Binomial}(100, 0.7)$ *and* $f \equiv 1$.

EXAMPLE 2.2. In Example 1.4 with large N, $v_n(f)$ stays extremely close to zero for any realistic value of n because $\rho(X_i)$ is very close to zero with high probability. But here we know that the actual convergence takes place at a sample size that is exponentially large in N. For instance, consider $\mu = \text{Binomial}(100, 0.5)$ and $\nu = \text{Binomial}(100, 0.7)$. Let f be the function that is identically equal to 1. Figure 2 shows the plot of the estimated standard deviation $\sqrt{v_n(f)}$ against n, as n ranges from 1 to 10^6. The estimated standard deviation remains fairly small throughout. However, since we know the actual value of $I(f)$ in this case (which is 1), it is easy to compute the actual error $|I_n(f) - I(f)|$ and check that the variance diagnostic is giving a false conclusion.

There are results in the literature that claim to show that the variance estimation method gives a valid criterion for the convergence of importance sampling. However, what these results actually show is that if n is so large that the importance sampling estimates are accurate, then $v_n(f)$ is small. In other words, the smallness of $v_n(f)$ is a *necessary* condition for convergence of importance sampling, but not a *sufficient* condition. For a diagnostic criterion to be useful, it needs to be both necessary and sufficient for convergence.

In practice, $v_n(f)$ is not usually the preferred diagnostic. Various self-normalized versions of $v_n(f)$ are used. It is possible that these more complicated estimates are also problematic in the same way, but we do not have a proof. It would be interesting to prove analogs of Theorem 2.1 for self-normalized diagnostic statistics.

In view of Theorem 1.1, it is natural to consider estimates of the Kullback–Leibler divergence as possible diagnostic tools for convergence. However, an in-

spection of the proof of Theorem 2.1 indicates that such estimates are likely to suffer from similar problems. The issue is that any diagnostic criterion that is itself dependent on the accuracy of an estimate obtained by importance sampling, is unlikely to be effective as a measure of the efficacy of importance sampling.

We suggest the following alternative diagnostic that is not itself an importance sampling estimate of any quantity. As usual, let μ be the sampling measure, ν be the target measure, and $\rho = d\nu/d\mu$. Let $X_1, X_2, \ldots$ be i.i.d. random variables with law μ. Define $q_n := \mathbb{E}(Q_n)$, where

$$Q_n := \frac{\max_{1\le i\le n} \rho(X_i)}{\sum_{i=1}^n \rho(X_i)}.$$

The size of q_n is our criterion for diagnosing convergence of importance sampling. The general prescription is that if for some value of n the quantity q_n is smaller than some pre-specified threshold (say, 0.01), declare that n is large enough for importance sampling to work. Note that the random variable Q_n always lies between 0 and 1 and, therefore, $q_n \in [0, 1]$. Moreover, given any n, it is possible to estimate q_n up to any desired degree of accuracy by repeatedly simulating Q_n and taking an average, since $q_n = \mathbb{E}(Q_n)$ and Q_n always lies between 0 and 1. Lastly, note that for estimating q_n using simulations in the above manner, it suffices to know the density ρ up to an unspecified normalizing constant. Repeatedly calculating Q_n, however, may be computationally expensive if either n is too large or ρ is too complex.

Why should one expect the smallness of the quantity q_n to be a valid diagnostic criterion for convergence of importance sampling? First, let us hasten to add the caveat that one can produce examples where it does not work. One such example is the following: Take a large number N. Let μ be the uniform distribution on $\{1, 2, \ldots, N\}$. Let ν be the distribution that puts mass $1/2N$ on the points $1, 2, \ldots, N-1$, and mass $(N+1)/2N$ on the point N. Then $\rho(x) = 1/2$ for $x = 1, 2, \ldots, N-1$ and $\rho(N) = (N+1)/2$. Under the sampling measure μ, $\rho = 1/2$ with probability $1 - 1/N$. Therefore, when $1 \ll n \ll N$, the quantity q_n will be small; but convergence of importance sampling will not happen until $n \gg N$.

In spite of the above counterexample, we expect that q_n is a valid diagnostic for many natural examples. This is made precise to a certain extent in the setting of Gibbs measures by Theorem 3.5 in the next section. A general heuristic argument for the effectiveness of the q_n diagnostic, on which the proof of Theorem 3.5 is based, can be described as follows.

Suppose that $\log \rho$ is concentrated under ν, so that Theorem 1.1 applies, and the sample size required for convergence of importance sampling is roughly e^L, where $L = \mathbb{E}_\nu(\log \rho)$. Take any n below this threshold. Let $M_n := \max_{1\le i\le n} \rho(X_i)$. Since $\rho(X_1), \rho(X_2), \ldots$ are i.i.d. random variables, it is easy to see that under mild conditions, $M_n \approx a$ with high probability, where a solves

$$n\mathbb{P}\big(\rho(X_1) \ge a\big) = 1. \tag{2.1}$$

Next, let $S_n := \sum_{i=1}^n \rho(X_i)$. Since $M_n \approx a$, therefore,

$$S_n \approx \sum_{i=1}^n \rho(X_i) 1_{\{\rho(X_i)\le a\}}.$$

Therefore,

$$\mathbb{E}(S_n) \approx n\mathbb{E}\big(\rho(X_1)1_{\{\rho(X_1)\le a\}}\big) = n\mathbb{P}_\nu(\rho \le a) = n\mathbb{P}_\nu(\log\rho \le \log a). \tag{2.2}$$

Now, $\mathbb{E}_\nu(\log\rho) = L > \log a$. Thus, $\mathbb{P}_\nu(\log\rho \le \log a)$ is a large deviation probability. Therefore, under mild conditions, one may expect that

$$\mathbb{P}_\nu(\log\rho \le \log a) \approx \mathbb{P}_\nu(\log\rho \approx \log a).$$

Plugging this into (2.2), we get

$$\begin{aligned}\mathbb{E}(S_n) &\approx n\mathbb{P}_\nu(\log\rho \approx \log a)\\ &= n\mathbb{E}\big(\rho(X_1)1_{\{\rho(X_1)\approx a\}}\big)\\ &= na\mathbb{P}\big(\rho(X_1)\approx a\big) \le na\mathbb{P}\big(\rho(X_1)\ge a\big).\end{aligned}$$

Using the equation (2.1) to evaluate the last term, we get $\mathbb{E}(S_n) \lesssim a$ and, therefore, $S_n = O(a)$ by Markov's inequality. Since $M_n \approx a$, this shows that

$$q_n = \mathbb{E}\left(\frac{M_n}{S_n}\right) = \Omega(1),$$

where $\Omega(1)$ means a quantity that is uniformly bounded away from zero as $n\to\infty$. The above heuristic shows that if $n \ll e^L$ and some appropriate conditions hold, then $q_n = \Omega(1)$. In other words, smallness of q_n should be a sufficient condition for convergence of importance sampling. This sketch can be made rigorous under certain circumstances. An instance of this is illustrated by Theorem 3.5 in the next section.

The smallness of q_n is also a necessary condition for convergence of importance sampling. Unlike sufficiency, the necessity can be rigorously proved in full generality.

THEOREM 2.3. *Let all notation be as in Theorem 1.1. Let q_n be defined as above. Let $\varepsilon_n := \mathbb{E}|I_n(1) - 1|$. Then*

$$q_n \le C\max\left\{\frac{1}{n}, \frac{\log\log(1/\varepsilon_n)}{\log(1/\varepsilon_n)}\right\},$$

where C is a universal constant.

As mentioned above, this theorem shows that the smallness of q_n is a necessary condition for convergence of importance sampling (recalling that by Theorem 1.1, convergence in L^1 is equivalent to actual good performance); if ε_n is small, then

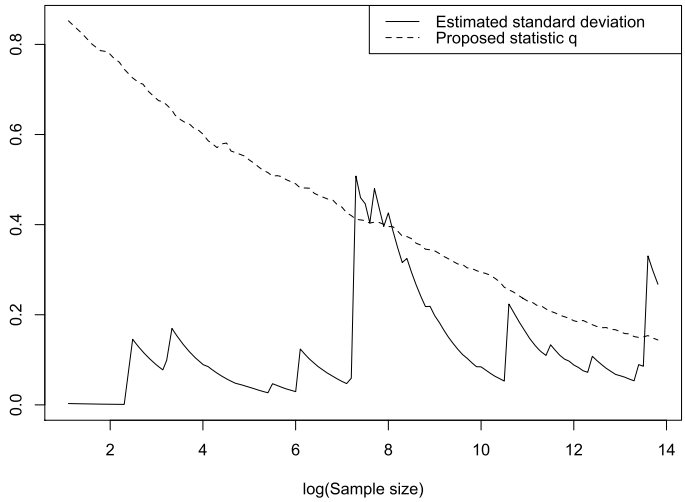

FIG. 3. *Performance of q_n in Example 2.2, plotted against the natural logarithm of the sample size.*

q_n is forced to be small. This is, however, a conceptual theorem. The bound is too poor to be applicable in practice, and the unspecified universal constant C can also be too large for the theorem to have any practical relevance.

The performance of q_n in Example 2.2 is depicted in Figure 3. The figure plots the estimated standard deviation $\sqrt{v_n(f)}$ and the statistic q_n, against $\log n$ as n ranges from 1 to 10^6. As in Figure 2, we see that the estimated standard error is generally quite misleading and unstable. On the other hand, the statistic q_n detects the nonconvergence in small samples and is very stable. The estimation of q_n was based on a sample of size 500 for each n.

Another illustration is given in Figure 4, which investigates the performance of q_n for Knuth's self-avoiding walks on a 10×10 grid, that was described in the literature review part of Section 1. The plot shows the behavior of q_n as n ranges from 1 to 10^5. We see that q_n is not too small (greater than 0.2) when $n = 10^3$, but starts getting appreciably small around $n = 10^4$. When $n = 10^5$, q_n is minuscule.

The random quantity Q_n is closely related to some existing diagnostics in the literature on sequential Monte Carlo (particle filters). It has the same form as the ∞-ESS statistic proposed by Huggins and Roy [31] in the context of sequential Monte Carlo. Here, ESS stands for "Effective Sample Size," a familiar concept in the sequential Monte Carlo literature. There is a substantial body of work on the efficacy of the effective sample size as a diagnostic tool, possibly beginning with Liu and Chen [41] and Doucet, de Freitas and Gordon [23]. See Whiteley, Lee and Heine [58] for some latest results. Huggins and Roy [31] established similar properties for the ∞-ESS. It would be interesting to see whether analogs of these results can be proved for the Q_n and q_n statistics proposed in this section.

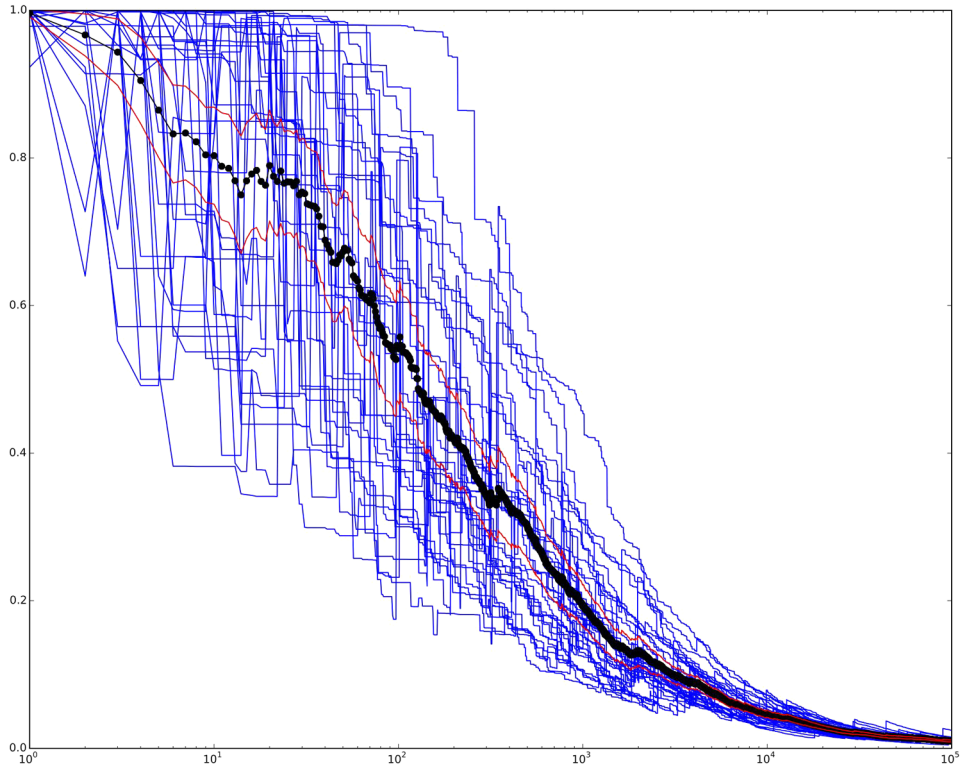

FIG. 4. *Performance of q_n for Knuth's self-avoiding walks on a 10×10 grid. The values of q_n, denoted by the thick dots, were estimated from 31 simulations of Q_n, which are depicted by the solid lines. Picture courtesy of Marc Coram.*

3. Importance sampling for exponential families (Gibbs measures). As in Section 1, let $\mathcal{X}$ be a set equipped with some sigma-algebra. Let λ be a finite measure on $\mathcal{X}$ that we shall call the "base measure." Let $H : \mathcal{X} \to \mathbb{R}$ be a measurable function, called the Hamiltonian, and let $\beta \in \mathbb{R}$ be a parameter, called the inverse temperature. The exponential family distribution (Gibbs measure) G_β on $\mathcal{X}$ defined by the sufficient statistic (Hamiltonian) H at a parameter value (inverse temperature) β is the probability measure on $\mathcal{X}$ that has probability density

$$Z(\beta)^{-1} \exp\bigl(-\beta H(x)\bigr)$$

with respect to the base measure λ, where

$$Z(\beta) = \int_{\mathcal{X}} \exp\bigl(-\beta H(x)\bigr)\, d\lambda(x)$$

is the normalizing constant, which is assumed to be finite. Let

$$F(\beta) := \log Z(\beta).$$

In physics parlance, the quantity $-F(\beta)/\beta$ is known as the free energy of the system at inverse temperature β.

Often, the normalizing constant $Z(\beta)$ is hard to calculate theoretically. Importance sampling is used to estimate $Z(\beta)$ in a variety of ways. See Gelman and Meng [28] for a useful review. Lelièvre, Rousset and Stoltz [39] show the breadth of this problem. One simple technique: Let β_0 be an inverse temperature at which we know how to generate a sample from the Gibbs measure. For example, $\beta_0 = 0$ is often a good choice, because G_0 is nothing but the base measure λ normalized to have total mass one. The goal is to estimate $Z(\beta)$ using a sample from G_{β_0}. Let $X_1, \ldots, X_n$ be an i.i.d. sample of size n from G_{β_0}. The importance sampling estimate of $Z(\beta)$ based on this sample is the following:

$$\hat{Z}_n(\beta) := \frac{Z(\beta_0)}{n} \sum_{i=1}^{n} \exp\bigl(-(\beta - \beta_0)H(X_i)\bigr).$$

It is easy to see that $\mathbb{E}(\hat{Z}_n(\beta)) = Z(\beta)$. The question is, how large does n need to be, so that the ratio $\hat{Z}_n(\beta)/Z(\beta)$ is close to 1 with high probability?

The following theorem shows that under favorable conditions, a sample of size approximately $\exp(F(\beta_0) - F(\beta) - (\beta_0 - \beta)F'(\beta))$ is necessary and sufficient. The proof, given in Section 4, is a simple consequence of Theorem 1.1 since $F(\beta_0) - F(\beta) - (\beta_0 - \beta)F'(\beta)$ is actually the Kullback–Leibler divergence of G_{β_0} from G_β. This theorem is a result for finite systems. A more general version of this result that applies in the thermodynamic limit is given later in this section.

THEOREM 3.1. *Let all notation be as above. Suppose that the Hamiltonian H satisfies the condition that for some $\beta' > |\beta|$,*

$$\int_{\mathcal{X}} \exp\bigl(\beta'|H(x)|\bigr)\, d\lambda(x) < \infty.$$

Then F is infinitely differentiable at β. Let

$$L := F(\beta_0) - F(\beta) - (\beta_0 - \beta)F'(\beta)$$

and

$$\sigma := |\beta_0 - \beta|\sqrt{F''(\beta)}.$$

If $n = \exp(L + r\sigma)$ for some $r \geq 0$, then

$$\mathbb{E}\left|\frac{\hat{Z}_n(\beta)}{Z(\beta)} - 1\right| \leq e^{-r\sigma/4} + \frac{4}{r}.$$

Conversely, if $n = \exp(L - r\sigma)$ for some $r \geq 0$, then for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(\frac{\hat{Z}_n(\beta)}{Z(\beta)} \geq 1 - \delta\right) \leq e^{-r\sigma/2} + \frac{4}{(1-\delta)r^2}.$$

It is not difficult to verify by direct calculation that F'' is always nonnegative. This implies, in particular, that F is convex. As a consequence of this feature, L and σ are also nonnegative.

In standard examples, F, F' and F'' are all of the same order of magnitude, and the magnitudes are large. Therefore, L is large and $\sigma = O(\sqrt{L})$, which implies that the required sample size is concentrated in the logarithmic scale at $\exp(L + O(\sqrt{L}))$. The situation is illustrated through the following examples.

EXAMPLE 3.2 (Independent spins). Take some $N \geq 1$ and let $\mathcal{X} = \{-1, 1\}^N$. Let λ be the counting measure on this set, and for $x = (x_1, \ldots, x_N) \in \mathcal{X}$, let

$$H(x) = -\sum_{i=1}^{N} x_i.$$

The G_β is nothing but the joint law of N i.i.d. random variables that take value 1 with probability $e^\beta/(e^\beta + e^{-\beta})$ and -1 with probability $e^{-\beta}/(e^\beta + e^{-\beta})$. A simple computation gives $Z(\beta) = 2^N(\cosh\beta)^N$. Therefore,

$$F(\beta) = N \log\cosh\beta + N\log 2.$$

Thus, for any given β_0 and β,

$$L = N \log\frac{\cosh\beta_0}{\cosh\beta} - N(\beta_0 - \beta)\tanh\beta$$

and

$$\sigma = 4\sqrt{N}|\beta_0 - \beta|\operatorname{sech}\beta.$$

Therefore, L is typically of order N and σ is typically of order $\sqrt{N}$.

EXAMPLE 3.3 (1D Ising model with periodic boundary). As in the previous example, let $\mathcal{X} = \{-1, 1\}^N$ and let λ be the counting measure on this set. For $x = (x_1, \ldots, x_N) \in \{-1, 1\}^N$, let

$$H(x) = -J\sum_{i=1}^{N} x_i x_{i+1} - h\sum_{i=1}^{N} x_i,$$

where $J \geq 0$, $h \in \mathbb{R}$, and x_{N+1} in the first sum stands for x_1. This is the Hamiltonian for the one-dimensional Ising model for a system of N spins with periodic boundary. The parameters J and h are traditionally known as the coupling constant and the strength of the external magnetic field. The partition function of this model is easily computed by the transfer matrix method (see Baxter [5]): $Z(\beta) = \operatorname{Tr}(V(\beta)^N)$, where $V(\beta)$ is the 2×2 matrix

$$\begin{bmatrix} e^{\beta(h+J)} & e^{-\beta J} \\ e^{-\beta J} & e^{-\beta(h-J)} \end{bmatrix}.$$

In other words, if $\lambda_1(\beta)$ and $\lambda_2(\beta)$ are the two eigenvalues of this matrix (arranged such that $|\lambda_1| \geq |\lambda_2|$), then

$$Z(\beta) = \lambda_1(\beta)^N + \lambda_2(\beta)^N.$$

Consequently,

$$F(\beta) = \log\bigl(\lambda_1(\beta)^N + \lambda_2(\beta)^N\bigr).$$

It is not hard to verify that

$$\lambda_1(\beta) = e^{\beta J}\cosh\beta h + \sqrt{e^{2\beta J}(\sinh\beta h)^2 + e^{-2\beta J}}$$

and

$$\lambda_2(\beta) = e^{\beta J}\cosh\beta h - \sqrt{e^{2\beta J}(\sinh\beta h)^2 + e^{-2\beta J}}.$$

Using these formulas, it is easy to write down explicit formulas for L and σ for any given β and β_0, and compute $a(\beta_0, \beta)$ and $b(\beta_0, \beta)$ such that as $N \to \infty$, $L \sim Na(\beta_0, \beta)$ and $\sigma \sim \sqrt{N}b(\beta_0, \beta)$.

Examples 3.2 and 3.3 demonstrate how Theorem 3.1 can be applied to calculate the sample size required for importance sampling in statistical mechanical models. However, these examples required exact computations in finite systems, which is rarely possible in complex models. Our next theorem deals with a generic sequence of models that converge to a limit. Exact computations are assumed to be possible only in the limit.

Let $\{\mathcal{X}_N\}_{N\geq 1}$ be a sequence of spaces equipped with sigma-algebras and finite measures $\{\lambda_N\}_{N\geq 1}$. For each N, let $H_N : \mathcal{X} \to \mathbb{R}$ be a measurable function, and for each $\beta \in \mathbb{R}$ let $G_{N,\beta}$ be the probability measure that has probability density proportional to $\exp(-\beta H_N(x))$ with respect to λ_N. Let

$$Z_N(\beta) := \int_{\mathcal{X}_N} \exp\bigl(-\beta H_N(x)\bigr)\, d\lambda_N(x)$$

be the normalizing constant of $G_{N,\beta}$, and assume that these quantities are finite. Let

$$F_N(\beta) := \log Z_N(\beta).$$

Let $\{L_N\}_{N\geq 1}$ be a sequence of numbers tending to infinity, and let

$$p(\beta) := \lim_{N\to\infty} \frac{F_N(\beta)}{L_N}$$

whenever the limit exists and is finite. For a suitable choice of L_N depending on the situation, the function $p(\beta)$ is sometimes referred to as the thermodynamic limit (or the thermodynamic free energy) of the sequence of systems described above. The thermodynamic limit is said to have a kth order phase transition at an

inverse temperature β if the first $k-1$ derivatives of p are continuous at β but the kth derivative is discontinuous at β.

Fix two inverse temperatures β_0 and β such that $\beta_0 < \beta$. The goal is to estimate $F_N(\beta)$ using importance sampling with a sample of size n from the Gibbs measure G_{N,β_0}, and determine how fast n needs to grow with N such that the ratio of this estimate and the true value tends to one as $N \to \infty$. Recall that the importance sampling estimate of $Z_N(\beta)$ is

$$\hat{Z}_{n,N}(\beta) = \frac{Z_N(\beta_0)}{n} \sum_{i=1}^{n} \exp\bigl(-(\beta-\beta_0)H_N(X_i)\bigr),$$

where $X_1, \ldots, X_n$ are i.i.d. draws from G_{N,β_0}. The following theorem identifies the sample size required for good performance of the above estimate as long as the system does not exhibit a first-order phase transition at β in the thermodynamic limit.

THEOREM 3.4. *Let all notation be as above. Let $\{L_N\}_{N\geq 1}$ be a sequence of constants such that the thermodynamic free energy p exists and is differentiable in a neighborhood of β, and exists at β_0. Assume that the derivative p' is continuous at β, and that there exists a finite constant C such that for all N and all $x \in \mathcal{X}_N$, $|H_N(x)| \leq C L_N$. Suppose that the sample size $n = n(N)$ grows with N in such a way that $L_N^{-1} \log n$ converges to a limit $b \in [0,\infty]$, and let*

$$q(\beta) := p(\beta_0) - p(\beta) - (\beta_0 - \beta)p'(\beta).$$

Then the following conclusions hold:

(i) *If $b > q(\beta)$, then $\hat{Z}_{n,N}(\beta)/Z_N(\beta) \to 1$ in probability as $N \to \infty$.*

(ii) *If $b < q(\beta)$, then $\hat{Z}_{n,N}(\beta)/Z_N(\beta) \not\to 1$ in probability as $N \to \infty$.*

(iii) *If $b = q(\beta)$ and p' is not constant in any neighborhood of β, then $L_N^{-1} \log \hat{Z}_{n,N}(\beta) \to p(\beta)$ in probability as $N \to \infty$. Note that this is a weaker version of the conclusion of part* (i).

Theorem 3.4 has potentially much wider applicability than Theorem 3.1, since thermodynamic limits are known in many important statistical mechanical systems. Classical examples from statistical physics include the 2D Ising model, the six and eight vertex models and many others (see Baxter [5] and McCoy [44]). Recently, a variety of exponential random graph models have been explicitly "solved" (see Chatterjee and Diaconis [16], Kenyon, Radin, Ren and Sadun [35], Kenyon and Yin [36] and Bhattacharya, Ganguly, Lubetzky and Zhao [6]). Similar progress has been made for nonuniform distributions on permutations (see Starr [56], Mukherjee [45] and Kenyon, Kral, Radin and Winkler [34]). All of these models provide examples for our theory.

The main strength of Theorem 3.4 is also its main weakness: While it gives a definitive answer for exactly solvable models, the theorem is not useful for systems

that are not exactly solvable in a thermodynamic limit. As discussed in Section 2, what a practitioner really wants is a diagnostic test that will confirm whether importance sampling has converged. Interestingly, it turns out that the use of the alternative diagnostic test proposed in Section 2 can be partially justified in the setting of Theorem 3.4, under one additional assumption. The extra assumption is that the system has no first-order phase transition at any point between β_0 and β, strengthening the assumption made in Theorem 3.4 that there is no first-order phase transition at β.

Take β_0 and β such that $\beta_0 < \beta$. Recall the quantities Q_n and q_n defined in Section 2. Since there are two parameters n and N involved here, we will write $q_{n,N}$ and $Q_{n,N}$ instead of q_n and Q_n. Then note that

$$Q_{n,N} = \frac{\max_{1\le i\le n} \exp(-(\beta-\beta_0)H_N(X_i))}{\sum_{i=1}^n \exp(-(\beta-\beta_0)H_N(X_i))},$$

and $q_{n,N} = \mathbb{E}(Q_{n,N})$. [Note that $q_{n,N}$ has nothing to do with $q(\beta)$.] The following theorem shows that if n is large enough (depending on N) for the importance sampling to work, then $q_{n,N}$ is exponentially small in L_N. Otherwise, it is not exponentially small.

THEOREM 3.5. *Let all notation and assumptions be as in Theorem 3.4. Additionally, assume that there is an open interval $I \supseteq [\beta_0, \beta]$ such that the thermodynamic free energy p is well defined and continuously differentiable in I, and that p' is not constant in any nonempty open subinterval of I. Then*:

(i) *If $b \le q(\beta)$, then*

$$\lim_{N\to\infty} \frac{\log q_{n,N}}{L_N} = 0.$$

Moreover, $L_N^{-1} \log Q_{n,N} \to 0$ in probability as $N \to \infty$.

(ii) *If $b > q(\beta)$, then*

$$\limsup_{N\to\infty} \frac{\log q_{n,N}}{L_N} < 0.$$

Moreover, there exists $c < 0$ such that $\mathbb{P}(L_N^{-1} \log Q_{n,N} \le c) \to 1$ as $N \to \infty$.

In particular, if n grows with N so fast that $q_{n,N}$ decays to zero like a negative power of L_N, then the estimated free energy $L_N^{-1} \log \hat{Z}_{n,N}(\beta)$ converges to the correct limit $p(\beta)$ in probability.

Incidentally, the binomial distribution, as well as more complicated systems like Knuth's self-avoiding paths, can be put into the framework of Theorem 3.5 by an appropriate choice of the Hamiltonian and the inverse temperatures β_0 and β, so that the system at inverse temperature β_0 gives the sampling distribution and the system at inverse temperature β gives the target distribution. The main theoretical question would be to prove the absence of a phase transition between β_0 and β.

4. Proofs. PROOF OF THEOREM 1.1. Suppose that $n = e^{L+t}$ and let $a := e^{L+t/2}$. Let $h(x) = f(x)$ if $\rho(x) \le a$ and 0 otherwise. Then

$$|I_n(f) - I(f)| \le |I_n(f) - I_n(h)| + |I_n(h) - I(h)| + |I(h) - I(f)|.$$

First, note that by the Cauchy–Schwarz inequality,

$$|I(h) - I(f)| \le \mathbb{E}\big(|f(Y)|; \rho(Y) > a\big) \le \|f\|_{L^2(\nu)} \sqrt{\mathbb{P}\big(\rho(Y) > a\big)}.$$

Similarly,

$$\begin{aligned}\mathbb{E}|I_n(f) - I_n(h)| &\le \mathbb{E}|\rho(X_1)f(X_1) - \rho(X_1)h(X_1)| \\ &= \mathbb{E}\big(|f(Y)|; \rho(Y) > a\big) \\ &\le \|f\|_{L^2(\nu)} \sqrt{\mathbb{P}\big(\rho(Y) > a\big)}.\end{aligned}$$

Finally, note that

$$\begin{aligned}\mathbb{E}|I_n(h) - I(h)| &\le \sqrt{\operatorname{Var}\big(I_n(h)\big)} \\ &= \sqrt{\frac{\operatorname{Var}(\rho(X_1)h(X_1))}{n}} \\ &\le \sqrt{\frac{\mathbb{E}(\rho(X_1)^2 h(X_1)^2)}{n}} \\ &\le \sqrt{\frac{a\mathbb{E}(\rho(X_1)f(X_1)^2)}{n}} \\ &= \|f\|_{L^2(\nu)} \sqrt{\frac{a}{n}}.\end{aligned}$$

Combining the upper bounds obtained above, we get the first inequality in the statement of the theorem.

Next, suppose that $n = e^{L-t}$ and let $a = e^{L-t/2}$. Markov's inequality gives

$$\mathbb{P}\big(\rho(X_1) > a\big) \le \frac{\mathbb{E}(\rho(X_1))}{a} = \frac{1}{a}. \tag{4.1}$$

Also,

$$\mathbb{E}\big(\rho(X_1); \rho(X_1) \le a\big) = \mathbb{P}\big(\rho(Y) \le a\big).$$

Thus,

$$\begin{aligned}&\mathbb{P}\big(I_n(1) \ge 1 - \delta\big) \\ &\quad\le \mathbb{P}\Big(\max_{1 \le i \le n} \rho(X_i) > a\Big) + \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \rho(X_i) 1_{\{\rho(X_i) \le a\}} \ge 1 - \delta\right)\end{aligned}$$

$$\le \sum_{i=1}^{n} \mathbb{P}\big(\rho(X_i) > a\big) + \frac{1}{1-\delta}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \rho(X_i) 1_{\{\rho(X_i)\le a\}}\right)$$
$$\le \frac{n}{a} + \frac{\mathbb{P}(\rho(Y) \le a)}{1-\delta}.$$

This completes the proof of the second inequality in the statement of the theorem. □

PROOF OF THEOREM 1.2. Suppose that $n = e^{L+t}$ and let $a = e^{L+t/2}$. Let

$$b := \sqrt{\frac{a}{n}} + 2\sqrt{\mathbb{P}\big(\rho(Y) > a\big)}.$$

Then by Theorem 1.1, for any $\varepsilon, \delta \in (0, 1)$,

$$\mathbb{P}\big(\big|I_n(1) - 1\big| \ge \varepsilon\big) \le \frac{b}{\varepsilon}$$

and

$$\mathbb{P}\big(\big|I_n(f) - I(f)\big| \ge \delta\big) \le \frac{\|f\|_{L^2(\nu)} b}{\delta}.$$

Now, if $|I_n(f) - I(f)| < \delta$ and $|I_n(1) - 1| < \varepsilon$, then

$$\begin{aligned} \big|J_n(f) - I(f)\big| &= \left|\frac{I_n(f)}{I_n(1)} - I(f)\right| \\ &\le \frac{|I_n(f) - I(f)| + |I(f)||1 - I_n(1)|}{I_n(1)} \\ &< \frac{\delta + |I(f)|\varepsilon}{1-\varepsilon}. \end{aligned}$$

Taking $\varepsilon = \sqrt{b}$ and $\delta = \|f\|_{L^2(\nu)}\varepsilon$ completes the proof of the first inequality in the statement of the theorem. Note that if ε turns out to be bigger than 1, then the bound is true anyway.

Next, suppose that $n = e^{L-t}$ and let $a = e^{L-t/2}$. Let $f(x) = 1$ if $\rho(x) \le a$ and 0 otherwise. Then $I(f) = \mathbb{P}(\rho(Y) \le a)$ and by (4.1),

$$\mathbb{P}\big(J_n(f) \ne 1\big) \le \sum_{i=1}^{n} \mathbb{P}\big(\rho(X_i) > a\big) \le \frac{n}{a}.$$

This completes the proof of the theorem. □

PROOF OF THEOREM 1.3. Let

$$K_n := \frac{I_n(1_A)}{\nu(A)} = \frac{1}{n}\sum_{i=1}^{n} \rho_A(X_i).$$

Suppose that $n = e^{L+t}$ and let $a = e^{L+t/2}$. Applying Theorem 1.1 with ρ replaced by ρ_A, this gives

$$\mathbb{E}|K_n - 1| \le \sqrt{\frac{a}{n}} + 2\sqrt{\mathbb{P}\big(\rho_A(Y) > a \mid Y \in A\big)},$$

which is the first assertion of the theorem. The second claim follows similarly. □

PROOF OF PROPOSITION 1.8. Let

$$Z = \frac{1}{2^N} \sum_{Np \le j \le N} \binom{N}{j},$$

so that

$$\nu_A(j) = \frac{\binom{N}{j}}{2^N Z} 1_A(j).$$

To explore the choice of sampling distribution let μ be the Binomial(N, θ) distribution for fixed $1/2 < \theta < 1$. Then

$$\begin{aligned} L_A = D(\nu_A \| \mu) &= \sum_j \log\big(\nu_A(j)/\mu(j)\big)\nu_A(j) \\ &= -\frac{1}{2^N Z} \sum_{Np \le j \le N} \log\big(2^N Z \theta^j (1-\theta)^{N-j}\big)\binom{N}{j} \\ &= -\log\big(2^N Z(1-\theta)^N\big) - \frac{\log(\theta/(1-\theta))}{2^N Z} \sum_{Np \le j \le N} j\binom{N}{j}. \end{aligned}$$

An identity of de Moivre (see Diaconis and Zabell [22]) shows that for any k, $0 \le k \le N$,

$$\frac{1}{2^N} \sum_{k \le j \le N} \binom{N}{j}\left(j - \frac{N}{2}\right) = \frac{k}{2} b(k; N, 1/2).$$

Thus, since Np is an integer,

$$L_A = -\log\big(2^N Z(1-\theta)^N\big) - \log\big(\theta/(1-\theta)\big)\left(\frac{Np}{2Z} b(Np; N, 1/2) + \frac{N}{2}\right).$$

To approximate Z, use an inequality of Bahadur [3], specialized here: Let

$$R = \frac{1}{2} b(Np; N, 1/2) \frac{Np+1}{Np+1-(N+1)/2}.$$

Then

$$1 \le \frac{R}{Z} \le 1 + x^{-2},$$

where

$$x = \frac{Np - N/2}{\sqrt{N/4}}.$$

For large N and p fixed, this gives

$$Z \sim b(Np; N, 1/2)\frac{p}{2p-1}.$$

Stirling's formula gives

$$2^N b(Np; N, 1/2) \sim \frac{(p^p(1-p)^{1-p})^N}{\sqrt{2\pi Np(1-p)}}.$$

Putting these approximations into L_A, we get

$$L_A \sim -N\log\bigl(p^p(1-p)^{1-p}(1-\theta)\bigl(\theta/(1-\theta)\bigr)^p\bigr).$$

The right-hand side, as a function of θ, is minimized when $\theta = p$. Plugging this in gives the claim. □

PROOF OF THEOREM 2.1. Let X be a random variable with law μ. Then note that

$$\begin{aligned} 1 = \mathbb{E}\bigl(\rho(X)\bigr) &= \int_0^\infty \mathbb{P}\bigl(\rho(X) \geq t\bigr)\,dt \\ &\geq \sum_{k=0}^{\infty} \int_{2^k}^{2^{k+1}} \mathbb{P}\bigl(\rho(X) \geq t\bigr)\,dt. \end{aligned}$$

Therefore, for any $l \geq 0$,

$$\min_{0\leq k\leq l} \int_{2^k}^{2^{k+1}} \mathbb{P}\bigl(\rho(X) \geq t\bigr)\,dt \leq \frac{1}{l+1}\sum_{k=0}^{l} \int_{2^k}^{2^{k+1}} \mathbb{P}\bigl(\rho(X) \geq t\bigr)\,dt \leq \frac{1}{l+1}.$$

Thus, there exists $k \leq l$ such that

$$\int_{2^k}^{2^{k+1}} \mathbb{P}\bigl(\rho(X) \geq t\bigr)\,dt \leq \frac{1}{l+1}.$$

Fixing l, take any such k. The above inequality implies that there exists $t \in [2^k, 2^{k+1}]$ such that

$$\mathbb{P}\bigl(\rho(X) \geq t\bigr) \leq \frac{1}{(l+1)2^k}.$$

Now take any $\varepsilon > 0$. Let $l = [1/\varepsilon^3]$, where $[1/\varepsilon^3]$ is the integer part of $1/\varepsilon^3$. Then there exists $k \leq l$ and $t \in [2^k, 2^{k+1}]$ such that the above inequality is satisfied. Let

$n = [t/\varepsilon^2] + 1$. Then

$$\begin{aligned}\mathbb{P}\Big(\max_{1\le i\le n}\rho(X_i)\ge\varepsilon^2 n\Big) &\le n\mathbb{P}\big(\rho(X)\ge\varepsilon^2 n\big)\\ &\le n\mathbb{P}\big(\rho(X)\ge t\big)\\ &\le \frac{n}{(l+1)2^k}\le\frac{\varepsilon^3 n}{2^k}\le\frac{\varepsilon(t+1)}{2^k}\le\frac{\varepsilon(2^{k+1}+1)}{2^k}\le 3\varepsilon.\end{aligned}$$

Consequently, for this n,

$$\begin{aligned}\mathbb{P}\big(v_n(f)\ge\varepsilon\big) &\le \mathbb{P}\left(\frac{1}{n^2}\sum_{i=1}^n f(X_i)^2\rho(X_i)^2\ge\varepsilon\right)\\ &\le \mathbb{P}\Big(\max_{1\le i\le n}\rho(X_i)\ge\varepsilon^2 n\Big)+\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n f(X_i)^2\rho(X_i)\ge\frac{1}{\varepsilon}\right)\\ &\le 3\varepsilon+\varepsilon\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n f(X_i)^2\rho(X_i)\right)\\ &= 3\varepsilon+\varepsilon\mathbb{E}\big(f(X)^2\rho(X)\big)=3\varepsilon+\varepsilon\|f\|^2_{L^2(\nu)}.\end{aligned}$$

To complete the proof, note that $n\le\varepsilon^{-2}t\le\varepsilon^{-2}2^{k+1}\le\varepsilon^{-2}2^{l+1}\le\varepsilon^{-2}2^{1+\varepsilon^{-3}}$. □

PROOF OF THEOREM 2.3. Since $0\le Q_n\le 1$, therefore,

$$\begin{aligned}q_n=\mathbb{E}(Q_n) &\le \frac{q_n}{2}+\mathbb{P}\left(Q_n\ge\frac{q_n}{2}\right)\\ &= \frac{q_n}{2}+\mathbb{P}\left(M_n\ge\frac{q_n S_n}{2}\right)\\ &\le \frac{q_n}{2}+\mathbb{P}\left(S_n<\frac{n}{2}\right)+\mathbb{P}\left(M_n\ge\frac{q_n n}{4}\right)\\ &\le \frac{q_n}{2}+2\varepsilon_n+\mathbb{P}\left(M_n\ge\frac{q_n n}{4}\right).\end{aligned}$$

Suppose that

$$\varepsilon_n\le\frac{q_n}{8}. \tag{4.2}$$

Then by the previous display,

$$\mathbb{P}\left(M_n\ge\frac{q_n n}{4}\right)\ge\frac{q_n}{4}. \tag{4.3}$$

Let $k := [8/q_n]$ and $l := [n/k]$. For $1\le j\le k$, define

$$M_{n,j} := \max_{(j-1)l+1\le i\le jl}\rho(X_i).$$

Then for any $x \geq 0$,

$$\begin{aligned}\mathbb{P}(S_n \geq kx) &\geq \mathbb{P}(M_{n,j} \geq x \text{ for all } 1 \leq j \leq k)\\ &= \bigl(\mathbb{P}(M_{n,1} \geq x)\bigr)^k\\ &= \bigl(1 - \bigl(\mathbb{P}\bigl(\rho(X_1) < x\bigr)\bigr)^l\bigr)^k\\ &= \bigl(1 - \bigl(\mathbb{P}(M_n < x)\bigr)^{l/n}\bigr)^k\\ &= \bigl(1 - \bigl(1 - \mathbb{P}(M_n \geq x)\bigr)^{l/n}\bigr)^k.\end{aligned}$$

Since $k \leq 8/q_n$ and $l \geq n/k - 1 \geq q_n n/8 - 1$, this gives

$$\mathbb{P}(S_n \geq kx) \geq \bigl(1 - \bigl(1 - \mathbb{P}(M_n \geq x)\bigr)^{q_n/8 - 1/n}\bigr)^{8/q_n}.$$

Suppose that

$$\frac{1}{n} \leq \frac{q_n}{16}. \tag{4.4}$$

Then the previous equation gives

$$\mathbb{P}(S_n \geq kx) \geq \bigl(1 - \bigl(1 - \mathbb{P}(M_n \geq x)\bigr)^{q_n/16}\bigr)^{8/q_n}.$$

Taking $x = q_n n/4$, assuming (4.2) and (4.4), and using (4.3), gives

$$\mathbb{P}(S_n \geq 2n) \geq \bigl(1 - (1 - q_n/4)^{q_n/16}\bigr)^{8/q_n}.$$

Now note that $1 - (1 - y)^{y/4}$ is asymptotic to $y^2/4$ as $y \to 0$, and is positive everywhere in the interval $(0, 1)$. Therefore, there is a positive constant C_1 such that $1 - (1 - y)^{y/4} \geq C_1 y^2$ for all $y \in [0, 1]$. Using this in the above inequality gives

$$\mathbb{P}(S_n \geq 2n) \geq e^{-8q_n^{-1}\log(C_2/q_n)},$$

where C_2 is a universal constant. By Markov's inequality, $\mathbb{P}(S_n \geq 2n) \leq \varepsilon_n$. Therefore,

$$e^{-8q_n^{-1}\log(C_2/q_n)} \leq \varepsilon_n.$$

This shows that as $\varepsilon_n \to 0$, q_n must also tend to zero. Using this and the monotonicity of the map $x \mapsto (\log x)/x$ for $x \geq e$, it is easy to show that

$$q_n \leq \frac{C_3 \log\log(1/\varepsilon_n)}{\log(1/\varepsilon_n)},$$

where C_3 is a universal constant. Note that this holds under (4.2) and (4.4). The maximum in the statement of the theorem accounts for these constraints. $\square$

PROOF OF THEOREM 3.1. By the integrability condition on H and the dominated convergence theorem, it is easy to see that F is infinitely differentiable. Moreover, if Y is a random variable with law G_β, then

$$F'(\beta) = -\mathbb{E}\big(H(Y)\big) \tag{4.5}$$

and

$$F''(\beta) = \operatorname{Var}\big(H(Y)\big). \tag{4.6}$$

The probability density of G_β with respect to G_{β_0} is

$$\rho(x) = \frac{Z(\beta_0)}{Z(\beta)} \exp\big(-(\beta - \beta_0)H(x)\big).$$

Therefore,

$$\frac{\hat{Z}_n(\beta)}{Z(\beta)} = \frac{1}{n}\sum_{i=1}^{n} \rho(X_i).$$

In the notation of Theorem 1.1, this is nothing but $I_n(1)$. Now note that if $Y \sim G_\beta$, then by (4.5) and (4.6),

$$\begin{aligned}\mathbb{E}\big(\log \rho(Y)\big) &= F(\beta_0) - F(\beta) - (\beta - \beta_0)\mathbb{E}\big(H(Y)\big) \\ &= F(\beta_0) - F(\beta) - (\beta_0 - \beta)F'(\beta) = L\end{aligned}$$

and

$$\operatorname{Var}\big(\log \rho(Y)\big) = (\beta_0 - \beta)^2 \operatorname{Var}\big(H(Y)\big) = (\beta_0 - \beta)^2 F''(\beta) = \sigma^2.$$

The proof is now easily completed by an application of Theorem 1.1, together with Chebychev's inequality for bounding the tail probabilities occurring in the statement of Theorem 1.1. □

PROOF OF THEOREM 3.4. Let ρ_N be the probability density of $G_{N,\beta}$ with respect to G_{N,β_0}. As in the proof of Theorem 3.1, we have

$$\log \rho_N(x) = F_N(\beta_0) - F_N(\beta) - (\beta - \beta_0)H_N(x). \tag{4.7}$$

For each γ, let $Y_{N,\gamma}$ be a random variable with law $G_{N,\gamma}$. A simple computation shows that for any bounded measurable function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\frac{d}{d\gamma}\mathbb{E}\big(\phi\big(H_N(Y_{N,\gamma})\big)\big) = \operatorname{Cov}\big(\phi\big(H_N(Y_{N,\gamma})\big), H_N(Y_{N,\gamma})\big).$$

It is an easy fact that if X is a real-valued random variable and f and g are two increasing functions, then $\operatorname{Cov}(f(X), g(X)) \geq 0$. From this and the above identity, it follows that for any bounded increasing function ϕ,

$$\frac{d}{d\gamma}\mathbb{E}\big(\phi\big(H_N(Y_{N,\gamma})\big)\big) \geq 0.$$

In particular, for any $t \in \mathbb{R}$, $\mathbb{P}(H_N(Y_{N,\gamma}) \geq t)$ is an increasing function of γ. This is an important observation that will be used below.

Take any γ such that p is well defined and differentiable in an open neighborhood of γ. Note that F_N is a convex function, since F_N'' is nonnegative by (4.6). Therefore, for any $h > 0$,

$$F_N'(\gamma) \leq \frac{F_N(\gamma + h) - F_N(\gamma)}{h}.$$

Consequently, if h is small enough, then

$$\limsup_{N\to\infty} \frac{F_N'(\gamma)}{L_N} \leq \frac{p(\gamma + h) - p(\gamma)}{h}.$$

Taking $h \to 0$, we get

$$\limsup_{N\to\infty} \frac{F_N'(\gamma)}{L_N} \leq p'(\gamma).$$

Similarly,

$$\liminf_{N\to\infty} \frac{F_N'(\gamma)}{L_N} \geq p'(\gamma).$$

This proves that for all γ in an open neighborhood of β,

$$\lim_{N\to\infty} \frac{F_N'(\gamma)}{L_N} = p'(\gamma).$$

Using the monotonicity of F_N' and p' and the continuity of p' at β, it is easy to conclude from the above identity that for any sequence $\gamma_N \to \beta$,

$$\lim_{N\to\infty} \frac{F_N'(\gamma_N)}{L_N} = p'(\beta). \tag{4.8}$$

By (4.5), note that for any γ

$$\bigl|F_N'(\gamma)\bigr| = \bigl|\mathbb{E}\bigl(H_N(Y_{N,\gamma})\bigr)\bigr| \leq CL_N.$$

Therefore,

$$\int_\beta^{\beta + L_N^{-1/2}} F_N''(\gamma)\,d\gamma = F_N'\bigl(\beta + L_N^{-1/2}\bigr) - F_N'(\beta) \leq 2CL_N.$$

Thus, there exists $\gamma_N \in [\beta, \beta + L_N^{-1/2}]$ such that

$$F_N''(\gamma_N) \leq 2CL_N^{3/2}. \tag{4.9}$$

Since $L_N \to \infty$, therefore, $\gamma_N \to \beta$. Hence by (4.5), (4.6), (4.8) and (4.9),

$$\lim_{N\to\infty} \mathbb{E}\biggl(\frac{-H_N(Y_{N,\gamma_N})}{L_N}\biggr) = \lim_{N\to\infty} \frac{F_N'(\gamma_N)}{L_N} = p'(\beta) \tag{4.10}$$

and

$$\lim_{N\to\infty} \operatorname{Var}\left(\frac{-H_N(Y_{N,\gamma_N})}{L_N}\right) = \lim_{N\to\infty} \frac{F_N''(\gamma_N)}{L_N^2} = 0.$$

This implies that $-H_N(Y_{N,\gamma_N})/L_N \to p'(\beta)$ in probability. Therefore, by our previous observation about the monotonicity of tail probabilities,

$$\lim_{N\to\infty} \mathbb{P}\left(\frac{-H_N(Y_{N,\beta})}{L_N} \geq p'(\beta) + \delta\right) = 0$$

for any $\delta > 0$. In a similar manner, one can show that

$$\lim_{N\to\infty} \mathbb{P}\left(\frac{-H_N(Y_{N,\beta})}{L_N} \leq p'(\beta) - \delta\right) = 0.$$

Thus, $-H_N(Y_{N,\beta})/L_N \to p'(\beta)$ in probability. Consequently,

$$\frac{\log \rho_N(Y_{N,\beta})}{L_N} \to p(\beta_0) - p(\beta) - (\beta_0 - \beta)p'(\beta) = q(\beta)$$

in probability. The proofs of parts (i) and (ii) are now easily completed by applying Theorem 1.1. To prove part (iii), take any $\gamma \in (\beta_0, \beta)$. Since p' is nonconstant in any neighborhood of β and p' is an increasing function due to the convexity of p, therefore $p'(\gamma) < p'(\beta)$. Thus, by the convexity of p,

$$\begin{aligned} q(\beta) - q(\gamma) &= (\beta - \beta_0)\big(p'(\beta) - p'(\gamma)\big) + (\beta - \gamma)p'(\gamma) + p(\gamma) - p(\beta) \\ &\geq (\beta - \beta_0)\big(p'(\beta) - p'(\gamma)\big) + (\beta - \gamma)\big(p'(\gamma) - p'(\beta)\big) \\ &= (\gamma - \beta_0)\big(p'(\beta) - p'(\gamma)\big) > 0. \end{aligned} \tag{4.11}$$

By part (i) of the theorem, this implies that if $b = q(\beta)$, then

$$\frac{\hat{Z}_{n,N}(\gamma)}{Z_N(\gamma)} \to 1$$

in probability and, therefore,

$$\frac{\log \hat{Z}_{n,N}(\gamma)}{L_N} \to p(\gamma) \tag{4.12}$$

in probability. Now note that for any β',

$$\left|\frac{d}{d\beta'} \log \hat{Z}_{n,N}(\beta')\right| = \left|\frac{\sum_{i=1}^n H_N(X_i)\exp(-(\beta' - \beta_0)H_N(X_i))}{\sum_{i=1}^n \exp(-(\beta' - \beta_0)H_N(X_i))}\right| \leq CL_N.$$

Therefore

$$\left|\log \hat{Z}_{n,N}(\beta) - \log \hat{Z}_{n,N}(\gamma)\right| \leq CL_N(\beta - \gamma). \tag{4.13}$$

Since γ is an arbitrary point in (β_0, β), it is now easy to complete the proof of part (iii) using (4.12), (4.13) and the continuity of p. $\square$

PROOF OF THEOREM 3.5. Suppose that $\{W_N\}_{N\ge 1}$ is a sequence of real-valued random variables and c is a real number. In this proof, we will use the notation

$$\text{P-}\liminf_{N\to\infty} W_N \ge c$$

to mean that for any $\varepsilon > 0$, $\lim_{N\to\infty} \mathbb{P}(W_N \ge c - \varepsilon) = 1$. Similarly, $\text{P-}\limsup_{N\to\infty} W_N \le c$ means that for any $\varepsilon > 0$, $\lim_{N\to\infty} \mathbb{P}(W_N \le c + \varepsilon) = 1$, and $\text{P-}\lim_{N\to\infty} W_N = c$ means that both of these hold, that is, $W_N \to c$ in probability.

First, suppose that $b \le q(\beta)$. Since p has no interval of linear behavior in the interval I, therefore, the convexity of p implies that p' is strictly increasing in I. From this and a variant of (4.11), it is easy to see that in the interval $I \cap [\beta_0, \infty)$, q is continuous and strictly increasing. Moreover, $q(\beta_0) = 0$. It follows that for any $a \in [0, q(\beta)]$, there exists $\gamma \in [\beta_0, \beta]$ such that $q(\gamma) = a$. Therefore, since $b \le q(\beta)$, therefore $b = q(\gamma)$ for some $\gamma \in [\beta_0, \beta]$. Suppose that $\gamma > \beta_0$. Then by part (i) and part (iii) of Theorem 3.4,

$$\begin{aligned}\text{P-}\lim_{N\to\infty} \frac{1}{L_N} \log\biggl(\sum_{i=1}^n \exp\bigl(-(\gamma-\beta_0)H_N(X_i)\bigr)\biggr) \\ = q(\gamma) + p(\gamma) - p(\beta_0) = (\gamma-\beta_0)p'(\gamma).\end{aligned} \tag{4.14}$$

Let $U_N(\gamma)$ denote the left-hand side of (4.14), without the limit. Using the positivity of the second derivative, it is easy to see that U_N is a convex function of γ. Take any $\gamma' \in (\beta_0, \gamma)$. Then by the convexity of U_N, we have

$$\begin{aligned}&\frac{\max_{1\le i\le n}(-H_N(X_i))}{L_N} \\ &\quad\ge \frac{1}{L_N}\frac{\sum_{i=1}^n (-H_N(X_i))\exp(-(\gamma-\beta_0)H_N(X_i))}{\sum_{i=1}^n \exp(-(\gamma-\beta_0)H_N(X_i))} \\ &\quad= U_N'(\gamma) \\ &\quad\ge \frac{U_N(\gamma) - U_N(\gamma')}{\gamma-\gamma'}.\end{aligned}$$

Now let $N \to \infty$ on both sides and apply (4.14), and then let $\gamma' \to \gamma$ on the right. This gives

$$\text{P-}\liminf_{N\to\infty} \frac{\max_{1\le i\le n}(-H_N(X_i))}{L_N} \ge p'(\gamma). \tag{4.15}$$

Next, note that

$$\log\biggl(\sum_{i=1}^n \exp\bigl(-(\beta-\beta_0)H_N(X_i)\bigr)\biggr)$$

$$\leq (\beta-\gamma)\max_{1\leq i\leq n}\bigl(-H_N(X_i)\bigr)+\log\biggl(\sum_{i=1}^{n}\exp\bigl(-(\gamma-\beta_0)H(X_i)\bigr)\biggr).$$

By (4.14) and (4.15), this implies that

$$\text{P-}\limsup_{N\to\infty}\frac{-\log Q_{n,N}}{L_N}\leq 0. \tag{4.16}$$

Note that this inequality was proved under the assumption that $\gamma>\beta_0$. Next, suppose that $\gamma=\beta_0$. Observe the easy inequality

$$\log\biggl(\sum_{i=1}^{n}\exp\bigl(-(\beta-\beta_0)H_N(X_i)\bigr)\biggr)\leq(\beta-\beta_0)\max_{1\leq i\leq n}\bigl(-H_N(X_i)\bigr)+\log n.$$

From this and the fact that $L_N^{-1}\log n\to q(\beta_0)=0$, it follows that (4.16) holds even if $\gamma=\beta_0$. Next, note that we trivially have

$$\log\biggl(\sum_{i=1}^{n}\exp\bigl(-(\beta-\beta_0)H_N(X_i)\bigr)\biggr)\geq\log\Bigl(\max_{1\leq i\leq n}\exp\bigl(-(\beta-\beta_0)H_N(X_i)\bigr)\Bigr)$$

which is same as

$$\text{P-}\liminf_{N\to\infty}\frac{-\log Q_{n,N}(\beta)}{L_N}\geq 0. \tag{4.17}$$

Equations (4.16) and (4.17) prove that if $b\leq q(\beta)$, then $L_N^{-1}\log Q_{n,N}\to 0$ in probability. Next, note that $Q_{n,N}\in[0,1]$, which implies that $\mathbb{E}(Q_{n,N})\in[0,1]$, and hence

$$\frac{\log\mathbb{E}(Q_{n,N})}{L_N}\leq 0. \tag{4.18}$$

On the other hand, Jensen's inequality gives

$$\frac{\log\mathbb{E}(Q_{n,N})}{L_N}\geq\frac{\mathbb{E}(\log Q_{n,N})}{L_N}. \tag{4.19}$$

It is not difficult to see that since $|H_N|\leq CL_N$ and $L_N^{-1}\log n\to b<\infty$, therefore, the random variable $|L_N^{-1}\log Q_{n,N}|$ is bounded by a nonrandom constant that does not vary with N. Since we already know that

$$L_N^{-1}\log Q_{n,N}\to 0$$

in probability, this shows that

$$\lim_{N\to\infty}\frac{\mathbb{E}(\log Q_{n,N})}{L_N}=0.$$

Combining this with (4.18) and (4.19), we get

$$\lim_{N\to\infty}\frac{\log\mathbb{E}(Q_{n,N})}{L_N}=0.$$

This completes the proof of part (i) of the theorem. Next, suppose that $b > q(\beta)$. Then note that by Theorem 3.4,

$$\text{P-}\lim_{N\to\infty} \frac{Z_N(\beta_0)}{nZ_N(\beta)} \sum_{i=1}^{n} \exp\bigl(-(\beta-\beta_0)H_N(X_i)\bigr) = 1. \tag{4.20}$$

Next, let

$$M_N := \max_{1\le i\le n} \exp\bigl(-(\beta-\beta_0)H_N(X_i)\bigr).$$

Since p is continuously differentiable in the interval I and p' is strictly increasing, therefore, there exists $\gamma \in I \cap (\beta,\infty)$ such that $b > q(\gamma)$. If $M_N > \exp(L_N(\beta-\beta_0)p'(\gamma))$, then

$$M_N \le \sum_{i=1}^{n} \exp\bigl(-(\beta-\beta_0)H_N(X_i)\bigr)1_{\{-H_N(X_i)>L_N p'(\gamma)\}} =: M'_N.$$

Therefore,

$$M_N \le \max\bigl\{\exp\bigl(L_N(\beta-\beta_0)p'(\gamma)\bigr), M'_N\bigr\}. \tag{4.21}$$

Define

$$R_N := \frac{Z_N(\beta_0)M'_N}{nZ_N(\beta)}. \tag{4.22}$$

Note that if Y is a random variable with law $G_{N,\beta}$, then for any $\theta > 0$,

$$\begin{aligned}
\mathbb{E}(R_N) &= \mathbb{E}\biggl(\frac{Z_N(\beta_0)}{Z_N(\beta)} \exp\bigl(-(\beta-\beta_0)H_N(X_1)\bigr)1_{\{-H_N(X_1)>L_N p'(\gamma)\}}\biggr)\\
&= \mathbb{P}\bigl(-H_N(Y) > L_N p'(\gamma)\bigr)\\
&\le e^{-\theta L_N p'(\gamma)}\mathbb{E}\bigl(e^{-\theta H_N(Y)}\bigr)\\
&= \exp\bigl(-\theta L_N p'(\gamma) + F_N(\beta+\theta) - F_N(\beta)\bigr).
\end{aligned} \tag{4.23}$$

Let

$$c(\theta) := p(\beta+\theta) - p(\beta) - \theta p'(\gamma),$$

and choose $\theta = (\gamma-\beta)/2$. Then by the strict convexity of p in I,

$$c(\theta) \le \theta\bigl(p'(\beta+\theta) - p'(\gamma)\bigr) < 0. \tag{4.24}$$

By (4.23) and Markov's inequality,

$$\begin{aligned}
\mathbb{P}\biggl(\frac{\log R_N}{L_N} \ge \frac{c(\theta)}{2}\biggr) &\le e^{-L_N c(\theta)/2}\mathbb{E}(R_N)\\
&\le \exp\biggl(-\frac{L_N c(\theta)}{2} - \theta L_N p'(\gamma) + F_N(\beta+\theta) - F_N(\beta)\biggr).
\end{aligned}$$

Taking logarithm on both sides, dividing by L_N and sending $N \to \infty$, we get

$$\limsup_{N\to\infty} \frac{1}{L_N} \log \mathbb{P}\Big(\frac{\log R_N}{L_N} \geq \frac{c(\theta)}{2}\Big) < 0. \tag{4.25}$$

In particular,

$$\text{P-}\limsup_{N\to\infty} \frac{\log R_N}{L_N} \leq \frac{c(\theta)}{2}. \tag{4.26}$$

Next, note that

$$\begin{aligned} &\lim_{N\to\infty} \frac{1}{L_N} \log\Big(\frac{Z_N(\beta_0)}{nZ_N(\beta)} \exp\big(L_N(\beta-\beta_0)p'(\gamma)\big)\Big) \\ &\quad = p(\beta_0) - p(\beta) - b + (\beta-\beta_0)p'(\gamma) \\ &\quad \leq q(\gamma) - b. \end{aligned} \tag{4.27}$$

From (4.21), (4.26) and (4.27), we get

$$\text{P-}\limsup_{N\to\infty} \frac{1}{L_N} \log\Big(\frac{Z_N(\beta_0)M_N}{nZ_N(\beta)}\Big) \leq \max\Big\{q(\gamma) - b, \frac{c(\theta)}{2}\Big\}. \tag{4.28}$$

By combining (4.20), (4.28), (4.24) and the fact that $q(\gamma) < b$, this shows that there exists $c < 0$ such that $\mathbb{P}(L_N^{-1} \log Q_{n,N} \leq c) \to 1$ as $N \to \infty$.

Next, let

$$V_N := \frac{Z_N(\beta_0)}{nZ_N(\beta)} \sum_{i=1}^{n} \exp\big(-(\beta-\beta_0)H_N(X_i)\big).$$

Then V_N is nothing but the importance sampling estimate $I_n(1)$ when the sampling measure is G_{N,β_0} and the target measure is $G_{N,\beta}$. In this setting, we have already seen in the proof of Theorem 3.4 that the quantity L of Theorem 1.1 is asymptotic to $L_N q(\beta)$ [to see this, simply combine equations (4.7) and (4.10)]. Combined with the fact that $L_N^{-1} \log n \to b$, this implies that the quantity t of Theorem 1.1 is asymptotic to $L_N(b - q(\beta))$ in the present setting.

Next, let $Y \sim G_{N,\beta}$ and ρ_N be the probability density of $G_{N,\beta}$ with respect to G_{N,β_0}. The formula (4.7) implies that $\log \rho_N(Y)$ is asymptotic to $L_N(p(\beta_0) - p(\beta)) - (\beta-\beta_0)H_N(Y)$. Combining all of these observations and applying Theorem 1.1, it follows that there is a positive constant c (which may depend on β, β_0 and b) such that for all large enough N,

$$\mathbb{E}|V_N - 1| \leq e^{-cL_N} + \sqrt{\mathbb{P}\big(-H_N(Y) \geq L_N\big(p'(\beta) + c\big)\big)}.$$

Take any $\theta > 0$. Then

$$\mathbb{P}\big(-H_N(Y) \geq L_N\big(p'(\beta) + c\big)\big) \leq e^{-\theta L_N(p'(\beta)+c)} \mathbb{E}\big(e^{-\theta H_N(Y)}\big).$$

It is easy to see that $\log \mathbb{E}(e^{-\theta H_N(Y)})$ is asymptotic to $L_N(p(\beta+\theta) - p(\beta))$. Thus, the logarithm of the right-hand side in the above display is asymptotic to $-\theta c L_N +$

$L_N(p(\beta+\theta)-p(\beta)-\theta p'(\beta))$. Since p' is continuous in a neighborhood of β, we can choose a θ small enough so that $\theta c > p(\beta+\theta)-p(\beta)-\theta p'(\beta)$. Therefore, there exists $C_1 > 0$ such that $\mathbb{P}(-H_N(Y) \geq L_N(p'(\beta)+c)) \leq e^{-C_1 L_N}$ for all large enough N. Combining these steps, we see that there is a positive constant C_2 such that $\mathbb{E}|V_N - 1| \leq e^{-C_2 L_N}$ for all large N, and hence

$$\mathbb{P}(V_N < 1/2) \leq 2e^{-C_2 L_N}. \tag{4.29}$$

Now note that by (4.21),

$$Q_{n,N} \leq \frac{\max\{S_N, R_N\}}{V_N}, \tag{4.30}$$

where R_N is defined in (4.22) and

$$S_N := \frac{Z_N(\beta_0)}{nZ_N(\beta)} \exp\bigl(L_N(\beta-\beta_0)p'(\gamma)\bigr).$$

Recall that by (4.25), there are positive constants C_3 and C_4 such that for all large enough N,

$$\mathbb{P}\bigl(R_N \geq e^{-C_3 L_N}\bigr) \leq e^{-C_4 L_N}. \tag{4.31}$$

Since $Q_{n,N} \in [0,1]$, (4.29), (4.30) and (4.31) imply that

$$\begin{aligned}\mathbb{E}(Q_{n,N}) &\leq \mathbb{P}(V_N < 1/2) + \mathbb{P}\bigl(R_N \geq e^{-C_3 L_N}\bigr) + 2\max\bigl\{S_N, e^{-C_3 L_N}\bigr\} \\ &\leq 2e^{-C_2 L_N} + e^{-C_4 L_N} + 2\max\bigl\{S_N, e^{-C_3 L_N}\bigr\}.\end{aligned}$$

However, we have already seen in (4.27) that there is a constant $C_5 > 0$ such that $S_N \leq e^{-C_5 L_N}$ for all large enough N. Thus,

$$\limsup_{N\to\infty} L_N^{-1} \log \mathbb{E}(Q_{n,N}) < 0.$$

This completes the proof of the theorem. □

Acknowledgments. We thank Ben Bond, Brad Efron, Jonathan Huggins, Don Knuth, Shuangning Li, Art Owen, Daniel Roy and David Siegmund for helpful comments. We also thank the referees and the Associate Editor for many helpful suggestions, and the editorial board of AAP for their patience with our long delay in submitting the revision.

REFERENCES

[1] AGAPIOU, S., PAPASPILIOPOULOS, O., SANZ-ALONSO, D. and STUART, A. M. (2017). Importance sampling: Computational complexity and intrinsic dimension. Preprint. Available at arXiv:1511.06196.

[2] ASMUSSEN, S. and GLYNN, P. W. (2007). *Stochastic Simulation*: *Algorithms and Analysis*. *Stochastic Modelling and Applied Probability* **57**. Springer, New York. MR2331321

[3] BAHADUR, R. R. (1960). Some approximations to the binomial distribution function. *Ann. Math. Stat.* **31** 43–54.
[4] BASSETTI, F. and DIACONIS, P. (2006). Examples comparing importance sampling and the Metropolis algorithm. *Illinois J. Math.* **50** 67–91. MR2247824
[5] BAXTER, R. J. (1982). *Exactly Solved Models in Statistical Mechanics*. Academic Press, London. MR0690578
[6] BHATTACHARYA, B. B., GANGULY, S., LUBETZKY, E. and ZHAO, Y. (2015). Upper tails and independence polynomials in random graphs. Preprint. Available at arXiv:1507.04074.
[7] BLANCHET, J. and GLYNN, P. (2008). Efficient rare-event simulation for the maximum of heavy-tailed random walks. *Ann. Appl. Probab.* **18** 1351–1378. MR2434174
[8] BLANCHET, J., GLYNN, P. and LEDER, K. (2012). On Lyapunov inequalities and subsolutions for efficient importance sampling. *ACM Trans. Model. Comput. Simul.* **22** 1104–1128.
[9] BLANCHET, J. and LIU, J. (2008). State-dependent importance sampling for regularly varying random walks. *Adv. Appl. Probab.* **40** 1104–1128. MR488534
[10] BLANCHET, J. and LIU, J. (2010). Efficient importance sampling in ruin problems for multidimensional regularly varying random walks. *J. Appl. Probab.* **47** 301–322. MR2668490
[11] BLITZSTEIN, J. and DIACONIS, P. (2010). A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Math.* **6** 489–522.
[12] BOUSQUET-MÉLOU, M. (2014). On the importance sampling of self-avoiding walks. *Combin. Probab. Comput.* **23** 725–748.
[13] CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in Hidden Markov Models*. Springer, New York. MR2159833
[14] CHAN, H. P. and LAI, T. L. (2007). Efficient importance sampling for Monte Carlo evaluation of exceedance probabilities. *Ann. Appl. Probab.* **17** 440–473. MR2308332
[15] CHAN, H. P. and LAI, T. L. (2011). A sequential Monte Carlo approach to computing tail probabilities in stochastic models. *Ann. Appl. Probab.* **21** 2315–2342. MR2895417
[16] CHATTERJEE, S. and DIACONIS, P. (2013). Estimating and understanding exponential random graph models. *Ann. Statist.* **41** 2428–2461. MR3127871
[17] CHEN, Y., DIACONIS, P., HOLMES, S. P. and LIU, J. S. (2005). Sequential Monte Carlo methods for statistical analysis of tables. *J. Amer. Statist. Assoc.* **100** 109–120.
[18] CHEN, Y. and LIU, J. S. (2007). Sequential Monte Carlo methods for permutation tests on truncated data. *Statist. Sinica* **17** 857–872. MR2397385
[19] DEL MORAL, P. (2004). *Feynman–Kac Formulae*: *Genealogical and Interacting Particle Systems with Applications*. Springer, New York.
[20] DEL MORAL, P. (2013). *Mean Field Simulation for Monte Carlo Integration*. CRC Press, Boca Raton, FL.
[21] DEL MORAL, P., KOHN, R. and PATRAS, F. (2015). A duality formula for Feynman–Kac path particle models. *C. R. Math. Acad. Sci. Paris* **353** 465–469.
[22] DIACONIS, P. and ZABELL, S. (1991). Closed form summation for classical distributions: Variations on a theme of de Moivre. *Statist. Sci.* **6** 284–302. MR1144242
[23] DOUCET, A., DE FREITAS, N. and GORDON, N., eds. (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York.
[24] DUPUIS, P., SPILIOPOULOS, K. and WANG, H. (2012). Importance sampling for multiscale diffusions. *Multiscale Model. Simul.* **10** 1–27.
[25] DUPUIS, P. and WANG, H. (2004). Importance sampling, large deviations, and differential games. *Stoch. Stoch. Rep.* **76** 481–508.
[26] EFRON, B. (2012). Bayesian inference and the parametric bootstrap. *Ann. Appl. Stat.* **6** 1971–1997. MR3058690
[27] FREER, C. E., MANSINGHKA, V. K. and ROY, D. M. (2010). When are probabilistic programs probably computationally tractable? Presented at the *NIPS Workshop on Monte*

Carlo Methods for Modern Applications, 2010. Available at http://danroy.org/papers/FreerManRoy-NIPSMC-2010.pdf.

[28] GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185. MR1647507

[29] HAMMERSLEY, J. M. and HANDSCOMB, D. C. (1965). *Monte Carlo Methods*. Methuen & Co., Ltd., London.

[30] HESTERBERG, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37** 185–194.

[31] HUGGINS, J. H. and ROY, D. M. (2015). Convergence of sequential Monte Carlo-based sampling methods. Preprint. Available at arXiv:1503.00966.

[32] HULT, H. and NYQUIST, P. (2016). Large deviations for weighted empirical measures arising in importance sampling. *Stochastic Process. Appl.* **126** 138–170.

[33] KAHN, H. and MARSHALL, A. W. (1953). Methods of reducing sample size in Monte Carlo computations. *J. Oper. Res. Soc. Am.* **1** 263–278.

[34] KENYON, R., KRAL, D., RADIN, C. and WINKLER, P. (2015). A variational principle for permutations. Preprint. Available at arXiv:1506.02340.

[35] KENYON, R., RADIN, C., REN, K. and SADUN, L. (2014). Multipodal structure and phase transitions in large constrained graphs. Preprint. Available at arXiv:1405.0599.

[36] KENYON, R. and YIN, M. (2014). On the asymptotics of constrained exponential random graphs. Preprint. Available at arXiv:1406.3662.

[37] KNUTH, D. E. (1976). Mathematics and computer science: Coping with finiteness. *Science* **194** 1235–1242. MR0534161

[38] KNUTH, D. E. (1996). *Selected Papers on Computer Science*. *CSLI Lecture Notes* **59**. CSLI Publications, Stanford, CA; Cambridge University Press, Cambridge.

[39] LELIÈVRE, T., ROUSSET, M. and STOLTZ, G. (2010). *Free Energy Computations*: *A Mathematical Perspective*. World Scientific, Singapore.

[40] LIU, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

[41] LIU, J. S. and CHEN, R. (1995). Blind deconvolution via sequential imputations. *J. Amer. Statist. Assoc.* **90** 567–576. MR3363399

[42] MADRAS, N. (1998). Umbrella sampling and simulated tempering. In *Numerical Methods for Polymeric Systems* (*Minneapolis*, *MN*, 1996). *IMA Vol. Math. Appl.* **102** 19–32. Springer, New York. MR1655577

[43] MADRAS, N. and PICCIONI, M. (1999). Importance sampling for families of distributions. *Ann. Appl. Probab.* **9** 1202–1225. MR1728560

[44] MCCOY, B. M. (2010). *Advanced Statistical Mechanics*. *International Series of Monographs on Physics* **146**. Oxford Univ. Press, Oxford. MR2583103

[45] MUKHERJEE, S. (2013). Estimation in exponential families on permutations. Preprint. Available at arXiv:1307.0978.

[46] NAIMAN, D. Q. and WYNN, H. P. (1997). Abstract tubes, improved inclusion-exclusion identities and inequalities and importance sampling. *Ann. Statist.* **25** 1954–1983. MR1474076

[47] OWEN, A. and ZHOU, Y. (1999). Adaptive importance sampling by mixtures of products of beta distributions. Technical report No. 1999–25, Dept. Statistics, Stanford Univ., Stanford, CA.

[48] OWEN, A. and ZHOU, Y. (2000). Safe and effective importance sampling. *J. Amer. Statist. Assoc.* **95** 135–143.

[49] OWEN, A. B. (2005). Multidimensional variation for quasi-Monte Carlo. In *Contemporary Multivariate Analysis and Design of Experiments*. *Ser. Biostat.* **2** 49–74. World Sci. Publ., Hackensack, NJ.

[50] OWEN, A. B. (2006). Quasi-Monte Carlo for integrands with point singularities at unknown locations. In *Monte Carlo and Quasi-Monte Carlo Methods* 2004 403–417. Springer, Berlin. MR2208721

[51] ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York.

[52] ROSENBLUTH, M. N. and ROSENBLUTH, A. W. (1955). Monte Carlo calculation of the average extension of molecular chains. *J. Chem. Phys.* **23** 356–359.

[53] SHI, J., SIEGMUND, D. and YAKIR, B. (2007). Importance sampling for estimating p values in linkage analysis. *J. Amer. Statist. Assoc.* **102** 929–937.

[54] SIEGMUND, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.* **4** 673–684. MR0418369

[55] SRINIVASAN, R. (2002). *Importance Sampling*: *Applications in Communications and Detection*. Springer, Berlin. MR1949250

[56] STARR, S. (2009). Thermodynamic limit for the Mallows model on S_n. *J. Math. Phys.* **50** 095208.

[57] TORRIE, G. M. and VALLEAU, J. P. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23** 187–199.

[58] WHITELEY, N., LEE, A. and HEINE, K. (2016). On the role of interaction in sequential Monte Carlo algorithms. *Bernoulli* **22** 494–529.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL, 390 SERRA MALL
STANFORD, CALIFORNIA 94305
USA
E-MAIL: souravc@stanford.edu
diaconis@math.stanford.edu