Statistical Science
1998, Vol. 13, No. 2, 163–185

Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling

Andrew Gelman and Xiao-Li Meng

Abstract. Computing (ratios of) normalizing constants of probability models is a fundamental computational problem for many statistical and scientific studies. Monte Carlo simulation is an effective technique, especially with complex and high-dimensional models. This paper aims to bring to the attention of general statistical audiences of some effective methods originating from theoretical physics and at the same time to explore these methods from a more statistical perspective, through establishing theoretical connections and illustrating their uses with statistical problems. We show that the *acceptance ratio method* and *thermodynamic integration* are natural generalizations of importance sampling, which is most familiar to statistical audiences. The former generalizes importance sampling through the use of a single "bridge" density and is thus a case of *bridge sampling* in the sense of Meng and Wong. Thermodynamic integration, which is also known in the numerical analysis literature as Ogata's method for high-dimensional integration, corresponds to the use of infinitely many and continuously connected bridges (and thus a "path"). Our *path sampling* formulation offers more flexibility and thus potential efficiency to thermodynamic integration, and the search of optimal paths turns out to have close connections with the Jeffreys prior density and the Rao and Hellinger distances between two densities. We provide an informative theoretical example as well as two empirical examples (involving 17- to 70-dimensional integrations) to illustrate the potential and implementation of path sampling. We also discuss some open problems.

Key words and phrases: Acceptance ratio method, Hellinger distance, Jeffreys prior density, Markov chain Monte Carlo, numerical integration, Rao distance, thermodynamic integration.

1. THE NEED FOR COMPUTING NORMALIZING CONSTANTS

Thanks to powerful Markov chain Monte Carlo (MCMC) methods, we can now simulate from a complex probability model $p(\omega)$, where ω is in general a high-dimensional variable, without knowing its normalizing constant. That is, one can evaluate $q(\omega)$, an *unnormalized density function*, but cannot directly calculate $z = \int q(\omega)\mu(d\omega)$, the *normalizing constant*, where μ can be a counting measure, a Lebesgue measure or a mixture of them. Distributions for which $q(\omega)$ can be easily computed but z is intractable arise in many statistical models, such as spatial models, Bayesian hierarchical models and models for incomplete data. In addition, sometimes a quantity of interest is deliberately formulated as a normalizing constant of a density from which draws can be made.

For example, in likelihood analysis with missing data, it commonly occurs that if one had all the observations, denoted by y_{com}, the computation of the complete-data likelihood for parameters ψ,

Andrew Gelman is Associate Professor, Department of Statistics, Columbia University, New York, New York 10027 (e-mail: gelman@stat.columbia.edu). Xiao-Li Meng is Associate Professor, Department of Statistics, University of Chicago, Chicago, Illinois 60637 (e-mail: meng@galton.uchicago.edu).

$L(\psi|y_{\text{com}}) = p(y_{\text{com}}|\psi)$, would be straightforward. This suggests the following method for simulating the observed-data likelihood $L(\psi|y_{\text{obs}}) = p(y_{\text{obs}}|\psi)$ in the cases where it is difficult to calculate $L(\psi|y_{\text{obs}})$ directly (an example is given in Section 5.2). Because

$$(1) \qquad p(y_{\text{com}}|y_{\text{obs}}, \psi) = \frac{p(y_{\text{com}}|\psi)}{p(y_{\text{obs}}|\psi)} \equiv \frac{L(\psi|y_{\text{com}})}{L(\psi|y_{\text{obs}})},$$

we can treat the likelihood of interest $L(\psi|y_{\text{obs}})$ as the normalizing constant of $p(y_{\text{com}}|y_{\text{obs}}, \psi)$, with the complete-data likelihood $L(\psi|y_{\text{com}})$ serving as the unnormalized density. In this formulation, y_{com} plays the role of ω in our general notation.

For instance, in genetic linkage analysis a key step is the computation of the likelihood of ψ, the locations of disease genes relative to a set of markers, based on the observed data y_{obs} from a pedigree. The problem turns out to be very difficult for a large pedigree with many loci, because of the missing observations (e.g., allele types inherited from parents) from some members of the pedigree. In this example, simulating y_{com} from $p(y_{\text{com}}|y_{\text{obs}}, \psi)$ is feasible though far from trivial, for example, by using the sequential imputation method (see Irwin, Cox and Kong, 1994, and Kong, Liu and Wong, 1994). Because of (1), we can use draws from $p(y_{\text{com}}|y_{\text{obs}}, \psi)$ to estimate $L(\psi|y_{\text{obs}})$ as a normalizing constant; this is essentially the only known effective method for dealing with this problem (see, e.g., Thompson, 1996). An application of bridge sampling, which we discuss in Section 3, in linkage analysis with large pedigrees is given by Jensen and Kong (1997).

A related general problem is, given an unnormalized joint density $q(\omega, \theta)$, to evaluate the marginal density $p(\theta) = \int p(\omega, \theta)\mu(d\omega)$. Marginal densities can be of interest in physical models (e.g., evaluating the distribution of the energy in a Gibbs model at a specified temperature) or in statistics, as marginal likelihoods or marginal posterior densities (e.g., if θ is a parameter of interest and ω is a vector of nuisance parameters; see Section 5.3 for an example). The computation of a Bayes factor, which requires the calculation of two probabilities, each of which is the marginal density under an individual model, $p(y) = \int p(y|\phi)p(\phi)\,d\phi$, is another problem of this sort. This problem has received much attention in recent literature; for example, Gelfand and Dey (1994), Chib (1995), Raftery (1996), Lewis and Raftery (1997) and DiCiccio, Kass, Raftery and Wasserman (1997). In particular, DiCiccio et al. (1997) provide a comparative study on a variety of methods, from Laplace approximation to bridge sampling, for computing Bayes factors. Their main conclusion is that bridge sampling typically provides an order of magnitude of improvement. The path sampling, which was not part of their study, has potentials for even more dramatic improvement, as we demonstrate in the current paper.

In physics and chemistry, a well-studied problem of computing normalizing constants is known as free energy estimation. The problem starts with an unnormalized density, the *system density*:

$$(2) \qquad q(\omega|T, \alpha) = \exp\left(-\frac{H(\omega, \alpha)}{kT}\right),$$

where $H(\omega, \alpha)$ is the energy function of state ω, k is Boltzmann's constant, T is the temperature and α is a vector of system characteristics. The *free energy* F of the system is defined as

$$(3) \qquad F(T, \alpha) = -kT\log\big(z(T, \alpha)\big),$$

where $z(T, \alpha)$ is the normalizing constant of the system density. Simulation of ω from $p(\omega|T, \alpha) = q(\omega|T, \alpha)/z(T, \alpha)$ is typically carried out via MCMC methods. For detailed discussions of this and related topics, see, among others, Ciccotti and Hoover (1986), Ceperley (1995) and Frankel and Smit (1996). A more statistically oriented review is given in Neal (1993).

In applications in both genetics and physics, the real interest is not a single normalizing constant itself, but rather ratios, or equivalently differences of the logarithms, of them (i.e., differences of log-likelihoods; free energy differences). This is also true in many other applications, such as computing observed-data likelihood ratios for the purpose of monitoring convergence of Monte Carlo EM algorithms (Meng and Schilling, 1996). Even when it appears that we need to deal with a single normalizing constant, we can almost always bring in a convenient completely known density on the same space as a reference point, as done in DiCiccio et al. (1997). Therefore, without loss of generality, we can consider a class of densities on the same space, which we denote either by a numerical index t or by a continuous parameter θ; that is,

$$(4) \quad p_t(\omega) = \frac{1}{z_t}q_t(\omega) \quad \text{or} \quad p(\omega|\theta) = \frac{1}{z(\theta)}q(\omega|\theta).$$

We make a convention that whenever one of the triplet $\{p, q, z\}$ is defined with a proper index, so are the other two with the same index. We also use λ as a generic notation for the log ratio (e.g., $\lambda = \log(z_1/z_0)$). For some examples, we are interested in a particular log ratio λ; for others, we wish to evaluate $z(\theta)$, up to an arbitrary multiplicative constant, for a continuous range of θ.

There are three common approaches for approximating analytically intractable normalizing con-

stants: analytic approximation (e.g., DiCiccio et al., 1997), numerical integration (e.g., Evans and Swartz, 1995) and Monte Carlo simulation. Of these, Monte Carlo simulation is widely used in statistics, mainly because of its general applicability and its familiarity to statisticians. Arguably, it is also the only general method available for dealing with complex, high-dimensional problems. Current routine simulation methods in statistics rely on the scheme of importance sampling, either using draws from an approximate density or from one of $p_t(\omega)$ (or $p(\omega|\theta)$); see Section 3. However, the theoretical evidence provided in Meng and Wong (1996) and the empirical evidence provided in DiCiccio et al. (1997) and in Meng and Schilling (1996) in the context of bridge sampling, show that substantial reductions of Monte Carlo errors can be achieved with little or minor increase in computational effort, by using draws from more than one $p_t(\omega)$. The key idea here is to use "bridge" densities to effectively shorten the distances among target densities, distances that are responsible for large Monte Carlo errors with the standard importance sampling methods.

The purpose of this paper is fourfold. First, we describe the method of *path sampling* for estimating λ unbiasedly (Section 2); the method is a general formulation, with the introduction of flexible paths aiming at reduction of Monte Carlo errors, of the *thermodynamic integration* method for simulating free energy differences. Second, we show that importance sampling, bridge sampling and path sampling represent a natural methodological evolution, from using no bridge densities to using an infinite number of them (Section 3); we thus show that thermodynamic integration is a natural generalization of the *acceptance ratio method*, another well-known method for free energy estimation, since the latter corresponds to bridge sampling with a single bridge. Third, we investigate the problem of optimal paths, which turns out to be closely related to the Jeffreys prior distribution and the Rao and Hellinger distances between two distributions; we illustrate the theoretical results by a simple yet informative example (Section 4). Fourth, we provide two applications (Section 5) to illustrate the implementation and potential of path sampling for statistical problems.

2. A GENERAL FRAMEWORK FOR PATH SAMPLING

2.1 Basic Identities for Path Sampling

Unless otherwise stated, we assume that densities are indexed by a continuous (vector) parameter θ. This may come naturally from a parametric family, as in many statistical applications. In general, given two unnormalized densities with the same support (not necessarily from the same family), $q_0(\omega)$ and $q_1(\omega)$, we can always construct a *continuous path* to link them (the issue of optimizing over the choice of path is discussed later in this paper). For example, as suggested in statistical physics (e.g., Neal, 1993, page 96), we can construct a geometric path using a scalar parameter $\theta \in [0, 1]$,

$$\text{(5)} \quad \text{geometric path,} \quad q(\omega|\theta) = q_0^{1-\theta}(\omega)q_1^{\theta}(\omega),$$

or a harmonic path by analogy to the harmonic mean. (As we show in Section 4.3, the geometric path is in general suboptimal for the purpose of estimating the ratio of normalizing constants.)

To derive the basic identity for path sampling, we first assume that θ is a scalar quantity; without loss of generality, we assume that $\theta \in [0, 1]$ and that we are interested in computing the ratio $r = z(1)/z(0)$. Taking logarithms and then differentiating both sides of the second equation in (4) with respect to θ yields the standard formula (e.g., Ripley, 1988, page 64), assuming the legitimacy of interchange of integration with differentiation,

$$\begin{aligned} \frac{d}{d\theta}\log z(\theta) &= \int \frac{1}{z(\theta)}\frac{d}{d\theta}q(\omega|\theta)\mu(d\omega) \\ &= E_\theta\left[\frac{d}{d\theta}\log q(\omega|\theta)\right], \end{aligned} \tag{6}$$

where E_θ denotes the expectation with respect to the sampling distribution $p(\omega|\theta)$. Identity (6) is a consequence of the fact that the expected score function is zero for any θ. By analogy to the potential in statistical physics, we label

$$U(\omega, \theta) = \frac{d}{d\theta}\log q(\omega|\theta).$$

Integrating (6) from 0 to 1 yields

$$\lambda = \log\left[\frac{z(1)}{z(0)}\right] = \int_0^1 E_\theta[U(\omega, \theta)]\,d\theta. \tag{7}$$

Now, if we consider θ as a random variable (as in Bayesian analysis) with a uniform distribution on $[0, 1]$, we can interpret the right-hand side of (7) as the expectation of $U(\omega, \theta)$ over the joint distribution of (ω, θ). More generally, we can introduce a prior density $p(\theta)$ for $\theta \in [0, 1]$ and rewrite (7) as

$$\lambda = E\left[\frac{U(\omega, \theta)}{p(\theta)}\right], \tag{8}$$

where the expectation is with respect to the joint density $p(\omega, \theta) = p(\omega|\theta)p(\theta)$.

Identity (8) immediately suggests an unbiased estimator of λ:

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\frac{U(\omega_i, \theta_i)}{p(\theta_i)} \tag{9}$$

using n (not necessarily independent) draws (ω_i, θ_i) from $p(\omega, \theta)$. In addition, we can estimate $\log(z(b)/z(a))$ for intermediate values $a, b \in [0, 1]$ by just using the sample points i for which $\theta_i \in [a, b]$. The simulation error of $\hat{\lambda}$ depends both on the choice of $p(\theta)$ and how the samples are actually drawn. A key advantage of (8) or (9) is that the summand is on the log scale, which is generally more stable than the ratio scale. This is particularly important when computing the log-likelihood ratio as a weighted sum of log-ratios of normalizing constants, as in Meng and Schilling (1996).

Extensions of (8) to multivariate θ are straightforward and in fact suggested to us the term *path sampling*. Suppose θ is now a d-dimensional parameter vector and we are interested in the ratio $z(\theta_1)/z(\theta_0)$ for given vectors θ_0 and θ_1. We first select a continuous *path* in the d-dimensional parameter space that links θ_0 and θ_1: $\theta(t) = (\theta_1(t), \ldots, \theta_d(t))$, for $t \in [0, 1]$, with $\theta(0) = \theta_0$ and $\theta(1) = \theta_1$. Defining

$$U_k(\omega, \theta) = \frac{\partial \log q(\omega|\theta)}{\partial \theta_k},$$

$$\dot{\theta}_k(t) = \frac{d\theta_k(t)}{dt}, \quad k = 1, \ldots, d$$

and applying the same argument as with (7) for t going from 0 to 1, we obtain

$$\begin{aligned}\lambda &= \int_0^1 E_{\theta(t)}\left[\frac{d}{dt}\log q(\omega|\theta(t))\right]dt \\ &= \int_0^1 E_{\theta(t)}\left[\sum_{k=1}^{d}\dot{\theta}_k(t)U_k(\omega, \theta(t))\right]dt.\end{aligned} \tag{10}$$

From (10), we can easily construct the corresponding path sampling estimator for λ,

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{d}\dot{\theta}_k(t_i)U_k(\omega_i, \theta(t_i))\right], \tag{11}$$

where the t_i's are sampled uniformly from [0,1] and ω_i is a draw from $p(\omega|\theta(t_i))$. For any given path, (11) is a consistent (and unbiased) estimator of λ as long as the sample average converges to its population average, a requirement that is met by many MCMC methods. The choice of the path obviously affects the Monte Carlo error, as we shall illustrate later. In searching for optimal estimators, the introduction of a nonuniform density for t on $[0, 1]$ is unnecessary, as such a density can be absorbed by the path function $\theta(t)$. In fact, even in the univariate case (i.e., (8) and (9)), we can reexpress the prior density via a path function by solving $\dot{\theta}(t) = 1/p(\theta(t))$.

2.2 Thermodynamic Integration and Ogata's Method

Using identity (7) for calculating λ is not a new idea. For example, the thermodynamic integration method uses (7) for computing the free energy difference between two molecular-dynamic systems. As a simple example, using the notation in (2)–(3), we can calculate the free energy difference between two systems with the same temperature T as

$$\begin{aligned}&F(T, \alpha_1) - F(T, \alpha_0) \\ &\quad = \int_{\alpha_0}^{\alpha_1} E_{T,\alpha}\left[\frac{\partial H(\omega, \alpha)}{\partial \alpha}\right]d\alpha,\end{aligned} \tag{12}$$

where $E_{T,\alpha}$ denotes the expectation with respect to the system density $p(\omega|T, \alpha)$ (here α is a scalar quantity, such as the volume). Equation (12) is an application of (7) in conjunction with (3) using $\log q(\omega|\theta = \alpha) = -H(\omega, \alpha)/(kT)$. Similarly, we can calculate free energy difference for systems with different temperatures but the same α; identity (10) also allows for different α's and different T's simultaneously. See Frenkel (1986), Frankel and Smit (1996) and Neal (1993, Section 6.2) for more discussions of thermodynamic integration—so named because identities such as (12) were originally derived from differential equations for describing thermodynamic relationships.

Applying (7), Ogata (1989; also see Ogata and Tanemura, 1984) proposed an innovative method for high-dimensional integrations. For simplicity, suppose we are interested in integrating a positive function $q(\omega_1, \ldots, \omega_k)$ on the k-dimensional cube $[a, b]^k$ that includes the origin $(0, \ldots, 0)$, where k can be very large (e.g., $k = 1000$). To apply (7), we construct a family of densities indexed by a scale parameter σ,

$$p(\omega|\sigma) = q(\sigma\omega_1, \ldots, \sigma\omega_k)/z_k(\sigma), \tag{13}$$

where

$$\begin{aligned}&z_k(\sigma) \\ &\quad = \int_a^b \int_a^b \cdots \int_a^b q(\sigma\omega_1, \ldots, \sigma\omega_k)\, d\omega_1\, d\omega_2 \cdots d\omega_k.\end{aligned}$$

Treating $q(\sigma\omega_1, \ldots, \sigma\omega_k) \equiv q(\sigma\omega)$ as the unnormalized density, we obtain from (7) that

$$\begin{aligned}&\log z_k(1) - \log z_k(0) \\ &\quad = \int_0^1 E_\sigma\left[\frac{d}{d\sigma}\log q(\sigma\omega)\right]d\sigma,\end{aligned} \tag{14}$$

where E_σ is with respect to the density given in (13). Since $z_k(1)$ is exactly the integration we want, and $z_k(0) = (b-a)^k q(0)$, (14) allows us to estimate $z_k(1)$ by using draws $\{(\omega^{(i)}, \sigma_i), i = 1, \ldots, n\}$ from $p(\omega|\sigma)p(\sigma)$, where $p(\omega|\sigma)$ is given by (13) and $p(\sigma) = 1$ for $\sigma \in [0, 1]$. Simulations from (13) can be accomplished via the Metropolis algorithm (Metropolis et al., 1953), as illustrated in Ogata (1989). In view of (8), we do not have to simulate σ from a uniform distribution; other densities may provide better Monte Carlo errors (see Section 4). Ogata's (1989) original proposals include the use of deterministic choices of σ_i (e.g., equal-spaced) and the use of numerical integration techniques (e.g., trapezoidal rule) to carry out the one-dimensional integration in (14), in which cases one needs multiple draws of ω for any given σ_i; see Sections 2.3 and 5.1.

It appears that Ogata (1989) had independently discovered the thermodynamic integration method. In a subsequent paper, Ogata (1990, page 408) wrote: "Recently, I learned that such an estimation method of $\log Z_N(\sigma)$, which is called *free energy*, by the derivative of a suitable scalar parameter σ has been commonly used in the field of statistical physics since late the 1970's (see Binder (1986) for example)." On the other hand, although Ogata's work was motivated by high-dimensional integrations for Bayesian computations (Ogata, 1990), there is no mention of his method in Evans and Swartz's (1995) review article on methods for approximating integrals with special emphasis on Bayesian integration problems, nor is there a mention of thermodynamic integration or other popular MCMC-based methods in physics, such as the acceptance ratio method (see Section 3).

We note these lack of citations not to criticize any author, but rather to emphasize the great need of communications among researchers, especially from different fields. Indeed, when we initially worked on this problem we also started from scratch (Gelman and Meng, 1994) because we were not aware of thermodynamic integration or Ogata's method. The lack of communication is particularly unfortunate in this case, because many of us have missed perhaps some most effective methods for high-dimensional integrations, in view of their routine and successful use in physics. A main purpose of this paper is to bring to the attention of statisticians some of these powerful methods, and at the same time to explore more flexible and statistical formulations aiming at potential further improvements as well as more general applicability. In particular, the formulation given in Section 2.1 allows arbitrary construction of a path, even in distribution spaces, as we explore in Section 4.3.

2.3 Path Sampling Estimates Using Numerical Integration over θ

An alternative to using (9) for estimating λ is to numerically evaluate the integral over θ, which essentially amounts to replacing $p(\theta)$ in (9) by inverses of spacings. For example, as in Ogata (1989), one can use the trapezoidal rule when θ is univariate. Specifically, we first order the unique values of the simulation draws θ_i so that $\theta_{(1)} < \theta_{(2)} < \theta_{(3)} < \cdots$, excluding any duplicates (such as occur if θ is updated using the Metropolis algorithm). For each newly labeled $\theta_{(j)}$, we then compute $\overline{U}_{(j)}$ as the average of the values of $U(\omega_i, \theta_i)$ for all simulation draws i for which $\theta_i = \theta_{(j)}$. Suppose we want to estimate the log density ratio $\lambda(a, b) = \log[z(b)/z(a)]$ for $0 \le a < b \le 1$. Let j_a and j_b be the indexes such that $\theta_{(j_a)} \le a < \theta_{(j_a+1)} < \cdots < \theta_{(j_b-1)} < b \le \theta_{(j_b)}$. Applying the trapezoidal rule, we estimate $\lambda(a, b)$ by

$$\begin{aligned}\hat{\lambda}(a, b) &= \tfrac{1}{2}\big(\theta_{(j_a+1)} - a\big)\big(\overline{U}_{(j_a+1)} + \overline{U}_a\big) \\ &\quad + \tfrac{1}{2}\sum_{j=j_a+1}^{j_b-2}\big(\theta_{(j+1)} - \theta_{(j)}\big)\big(\overline{U}_{(j+1)} + \overline{U}_{(j)}\big) \\ &\quad + \tfrac{1}{2}\big(b - \theta_{(j_b-1)}\big)\big(\overline{U}_b + \overline{U}_{(j_b-1)}\big),\end{aligned} \tag{15}$$

where $\overline{U}_a$ and $\overline{U}_b$ are obtained via interpolation or extrapolation, wherever necessary. Similarly, one can apply Simpson's rule.

Estimating λ using (15) is particularly useful when θ is evaluated on a fixed grid or when $p(\theta)$ is not known. The latter happens, for example, when the draws of (ω, θ) are made jointly via a Metropolis–Hastings algorithm (Hastings, 1970) using our ability to evaluate $q(\omega|\theta)$, which we now view as an unnormalized density on the joint space (ω, θ). In this case, $p(\theta)$ is proportional to $\int q(\omega|\theta)\mu(d\omega)$, which is the unknown normalizing constant $z(\theta)$ that we want to estimate. In such cases, (9) is not applicable but (15) is. See Section 5.1 for more discussion of this issue.

In the case that $z(\theta)$ is interpreted as a (unnormalized) marginal density, a similar method can be applied to estimate its corresponding cumulative distribution function (cdf). We first estimate, for any $0 < a \le 1$, the *unnormalized* cdf $G(a) = \int_0^a z(\theta)\,d\theta$ by

$$\begin{aligned}\widehat{G}(a) &= \tfrac{1}{2}\sum_{j=0}^{j_a-1}\Big\{\big(\theta_{(j+1)} - \theta_{(j)}\big) \\ &\qquad \cdot \big(\exp\big[\hat{\lambda}(0, \theta_{(j+1)})\big] + \exp\big[\hat{\lambda}(0, \theta_{(j)})\big]\big)\Big\} \\ &\quad + \tfrac{1}{2}\big(a - \theta_{(j_a)}\big) \\ &\quad \cdot \big(\exp\big[\hat{\lambda}(0, a)\big] + \exp\big[\hat{\lambda}(0, \theta_{(j_a)})\big]\big),\end{aligned} \tag{16}$$

where $\theta_0 = 0$ and $\hat{\lambda}(\cdot, \cdot)$ is defined by (15). We then estimate the cdf by

$$\widehat{F}(a) = \frac{\widehat{G}(a)}{\widehat{G}(1)}. \tag{17}$$

For multivariate θ, just as in (10), there is no unique way of performing the numerical integrations; we can apply (15) with many different choices of path, and we can even consider combining (e.g., by weighted averages) estimators from different paths (see Section 4). Here, we present a simple method, based on averaging over one component of θ at a time, that turns out to be effective in our example of Section 5.2. For simplicity, we describe the method when θ is two-dimensional and evaluated on a rectangular grid of values (θ_1^i, θ_2^j), $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$. We first estimate the following functions on the grid:

$$g_1(\theta_1, \theta_2) = \log z(\theta_1, \theta_2) - \log z(\theta_1^0, \theta_2),$$
$$g_2(\theta_1, \theta_2) = \log z(\theta_1, \theta_2) - \log z(\theta_1, \theta_2^0),$$

where (θ_1^0, θ_2^0) can be any fixed point on the grid. For each θ_2^j, the function $g_1(\theta_1, \theta_2^j)$ can be estimated as a function of θ_1 using the path sampling estimate (15), averaging along θ_1. Similarly, $g_2(\theta_1^i, \theta_2)$ can be estimated by path sampling along θ_2, for each θ_1^i. These estimates can be combined using the following identity:

$$\begin{aligned} &\log z(\theta_1, \theta_2) - \log z(\theta_1^i, \theta_2^0) \\ &\quad = g_1(\theta_1, \theta_2) + g_2(\theta_1^i, \theta_2) \\ &\qquad - g_1(\theta_1^i, \theta_2) \quad \text{for any } \theta_1^i. \end{aligned} \tag{18}$$

Averaging over all values of θ_1^i yields

$$\begin{aligned} \log z(\theta_1, \theta_2) &= g_1(\theta_1, \theta_2) \\ &\quad + \frac{1}{m_1} \sum_{i=1}^{m_1} \big(g_2(\theta_1^i, \theta_2) - g_1(\theta_1^i, \theta_2)\big) \\ &\quad + \text{constant}. \end{aligned} \tag{19}$$

Of course, the order of θ_1 and θ_2 can be reversed in the above expression, giving an alternative estimate; we find in the example of Section 5.2 that the order of integration can make a practical difference. Section 4 provides a theoretical investigation of the choices of paths.

3. A METHODOLOGICAL EVOLUTION

3.1 Direct Importance Sampling Methods

Two different importance sampling schemes are commonly used for computing normalizing constants. The first approach uses draws from a trial density $\tilde{p}(\omega)$ that is *completely* known (e.g., an analytic approximation of the target density $p(\omega) = q(\omega)/z$). The importance sampling estimator of z is based on the identity

$$z = E_{\tilde{p}}\left[\frac{q(\omega)}{\tilde{p}(\omega)}\right],$$

and the corresponding Monte Carlo estimator is

$$\hat{z} = \frac{1}{n} \sum_{i=1}^{n} \frac{q(\omega_i)}{\tilde{p}(\omega_i)}, \tag{20}$$

where $\omega_1, \ldots, \omega_n$ are draws from $\tilde{p}(\omega)$. For example, Dempster, Selwyn and Weeks (1983) use this method to check an analytic approximation of z for a logistic regression likelihood. As usual with importance sampling, this method is effective only if $\tilde{p}$ is a fairly good approximation to p. For complex models, such as those encountered in free-energy estimations, finding an acceptable importance sampling density is often out of the question. In fact, even with various variance-reduction techniques (e.g., using control variates), importance sampling does not provide usable answers for these complex problems—otherwise, the more advanced methods would not be so popular.

The second kind of importance sampling method uses draws from densities that themselves are only known in unnormalized forms, and thus (20) cannot be applied directly. This is typically the case with iterative simulation (e.g., the Metropolis algorithm) where one can produce draws from $p_t(\omega)$ while only knowing $q_t(\omega) = z_t p_t(\omega)$, with z_t being the unknown quantity of interest. This is the situation we address in this paper. In such a case, various methods are based on special cases of the following identity studied in detail by Meng and Wong (1996):

$$r \equiv \frac{z_1}{z_0} = \frac{E_0[q_1(\omega)\alpha(\omega)]}{E_1[q_0(\omega)\alpha(\omega)]}, \tag{21}$$

where E_t denotes the expectation with respect to $p_t(\omega)$ $(t = 0, 1)$, $\alpha(\omega)$ is an arbitrary function satisfying

$$0 < \left| \int_{\Omega_0 \cap \Omega_1} \alpha(\omega) p_0(\omega) p_1(\omega) \mu(d\omega) \right| < \infty, \tag{22}$$

and Ω_t is the support of $p_t(\omega)$, and we assume $\mu(\Omega_0 \cap \Omega_1) > 0$. For example, taking $\alpha = q_0^{-1}$ in (21) leads to the commonly used identity (e.g., Ott, 1979; Geyer and Thompson, 1992; Green, 1992):

$$\frac{z_1}{z_0} = E_0\left[\frac{q_1(\omega)}{q_0(\omega)}\right] \quad \text{assuming } \Omega_1 \subset \Omega_0, \tag{23}$$

and taking $\alpha = [q_0 q_1]^{-1}$ leads to a generalization of the "harmonic rule" of Newton and Raftery (1994). When $z_1 = 1$, that is, when $q_1(w) = p_1(w)$, (23) leads to the so-called *reciprocal importance sam-*

pling method (see Gelfand and Dey, 1994, and DiCiccio et al., 1997).

3.2 Acceptance Ratio Method and Bridge Sampling

While (21) is trivial to verify, it was the key identity underlying the powerful *acceptance ratio* method of Bennett (1976), who motivated (21) by considering a Metropolis algorithm that allows moves between p_0 and p_1. Here we recast his derivation under the more general Metropolis–Hastings algorithm in order to reveal an explicit relationship between the α function in (21) and the corresponding *proposal*, or jumping distribution, $J(\cdot|\cdot)$, for the Metropolis–Hastings algorithm.

We start by considering a Metropolis–Hastings algorithm on the joint space of (ω, t), where $t = 0$ or 1, with target density $p(\omega, t) \propto q_t(\omega)$. Clearly, $p(\omega|t) = p_t(\omega)$, and marginally

$$\frac{p(t=1)}{p(t=0)} = \frac{z_1}{z_0} = r, \tag{24}$$

which is the ratio in which we are interested. Now consider all moves where ω stays the same but t changes (i.e., we are switching from one density to the other with the same argument ω). By the detailed balance requirement of the Metropolis–Hastings algorithm, we have

$$q_1(\omega)T\big((\omega,0)|(\omega,1)\big) = q_0(\omega)T\big((\omega,1)|(\omega,0)\big), \tag{25}$$

where the *transition kernel* $T(\cdot|\cdot)$ is given by (when $u \neq v$)

$$\begin{aligned} &T\big((\omega,u)|(\omega,v)\big) \\ &\quad = J\big((\omega,u)|(\omega,v)\big) \\ &\qquad \cdot \min\left\{1, \frac{q_u(\omega)J((\omega,v)|(\omega,u))}{q_v(\omega)J((\omega,u)|(\omega,v))}\right\} \end{aligned} \tag{26}$$

$$\begin{aligned} &= q_u(\omega)\min\left\{\frac{J((\omega,u)|(\omega,v))}{q_u(\omega)}, \right. \\ &\qquad \left. \frac{J((\omega,v)|(\omega,u))}{q_v(\omega)}\right\}. \end{aligned} \tag{27}$$

It follows then, by integrating (or summing) both sides of (25) with respect to $\mu(d\omega)$,

$$\frac{z_1}{z_0} = \frac{E_0[T((\omega,1)|(\omega,0))]}{E_1[T((\omega,0)|(\omega,1))]}, \tag{28}$$

which is the same as (21), in view of (27), when we let

$$\begin{aligned} \alpha(\omega) = \min&\left\{\frac{J\big((\omega,1)|(\omega,0)\big)}{q_1(\omega)}, \right. \\ &\left. \frac{J\big((\omega,0)|(\omega,1)\big)}{q_0(\omega)}\right\}. \end{aligned} \tag{29}$$

The derivation of Bennett (1976) corresponds to choosing $J((\omega,1)|(\omega,0)) = J((\omega,0)|(\omega,1)) = 1$, that is, always proposing to switch. The probability that a proposal is accepted is then the same as the transition kernel (see (26)), and thus the right-hand side of (28) is the ratio of the (marginal) acceptance probabilities—hence the name *acceptance ratio* given by Bennett. For general J (and thus α, which corresponds to Bennett's weight function W), Meng and Wong (1996) suggested the name *bridge sampling*, a term that we explain in Section 3.3. Note that although one could implement the above switching algorithm and then use the empirical proportions to estimate r via (24), it is not necessary and in fact typically not desirable to do so because one can generally estimate r more accurately by using (28) and averaging the acceptance *probabilities* rather than the acceptance rates—this is a case of "Rao–Blackwellization" (as in Gelfand and Smith, 1990).

Given draws $(\omega_{0i}, i = 1, \ldots, n_0)$ from $p_0(\omega)$, draws $(\omega_{1i}, i = 1, \ldots, n_1)$ from $p_1(\omega)$ and a choice of α, the sample version of (21) is

$$\hat{r}_\alpha = \frac{(1/n_0)\sum_{i=1}^{n_0} q_1(\omega_{0i})\alpha(\omega_{0i})}{(1/n_1)\sum_{i=1}^{n_1} q_0(\omega_{1i})\alpha(\omega_{1i})}. \tag{30}$$

Whereas $\hat{r}_\alpha$ is a consistent estimator of r as long as the sample averages in (30) converge to their corresponding population means, its variance obviously varies with α and how the draws are made. The question of optimal choice of α, however, is difficult to answer in general due to the correlations among the draws. A case where the answer is easily obtained is when we have independent draws from both p_0 and p_1; although this assumption is typically violated in practice, it permits useful theoretical explorations and in fact the optimal estimator obtained under this assumption performs rather well in general (see Bennett, 1976; Meng and Schilling, 1996).

Specifically, under the independence assumption, the optimal α in the sense of minimizing the asymptotic variance of $\log(\hat{r}_\alpha)$ (Bennett, 1976) or equivalently the asymptotic relative variance of $\hat{r}_\alpha$ (Meng and Wong, 1996) is given by

$$\begin{aligned} \alpha_{\text{opt}}(\omega) &\propto \frac{1}{s_0 p_0(\omega) + s_1 p_1(\omega)} \\ &\propto \frac{1}{s_0 r q_0(\omega) + s_1 q_1(\omega)}, \quad \omega \in \Omega_0 \cap \Omega_1, \end{aligned} \tag{31}$$

where $s_t = n_t/(n_0 + n_1)$, $t = 0, 1$, are assumed to be asymptotically bounded away from 0 and 1. The

corresponding asymptotic minimal error is

$$\frac{1}{ns_0s_1}\left[\left(\int_{\Omega_0\cap\Omega_1}\frac{p_0p_1}{s_0p_0+s_1p_1}\mu(d\omega)\right)^{-1}-1\right]. \tag{32}$$

Since the optimal α_{opt} is not directly usable as it depends on the unknown ratio r, Meng and Wong (1996) construct an iterative estimator,

$$\hat{r}_{\text{opt}}^{(t+1)}=\frac{(1/n_0)\sum_{i=1}^{n_0}[l_{0i}/(s_0\hat{r}_{\text{opt}}^{(t)}+s_1l_{0i})]}{(1/n_1)\sum_{i=1}^{n_1}[1/(s_0\hat{r}_{\text{opt}}^{(t)}+s_1l_{1i})]}, \quad t=0,1,\ldots, \tag{33}$$

where $l_{mi}=q_1(\omega_{mi})/q_0(\omega_{mi})$, $m=0,1$, are calculated before the iteration. They show that each iterate in (33), $\hat{r}^{(t+1)}$, $t\geq 0$, provides a consistent estimator of r, and that the unique limit $\hat{r}_{\text{opt}}$ achieves the asymptotic minimal error given in (32). They also study noniterative choices of α such as $\alpha=1$ and $\alpha=(q_0q_1)^{-1/2}$. The empirical results presented in Meng and Schilling (1996) show that these estimators can substantially (e.g., by a factor of 5 to 30) reduce the relative mean-squared errors compared to estimators based on (the same amount of) draws from only one density (e.g., p_0), such as when using (23). Note that Bennett (1976) suggested a graphical method for obtaining $\hat{r}_{\text{opt}}$, and Geyer (1994) proposed an interesting "profile-likelihood" derivation for $\hat{r}_{\text{opt}}$.

3.3 Connecting Bridge and Path Sampling to Importance Sampling

The fundamental identity (21) underlying bridge sampling can also be motivated easily from importance sampling, which is familiar to most statistical readers. To see this, define

$$\alpha(\omega)=\frac{q_{1/2}(\omega)}{q_0(\omega)q_1(\omega)}, \quad \omega\in\Omega_0\cap\Omega_1, \tag{34}$$

where $q_{1/2}(\omega)$ is an arbitrary unnormalized density having support $\Omega_0\cap\Omega_1$ (thus, condition (22) is satisfied). We use the subscript "1/2" to indicate that we intend to use a density that is "between" q_0 and q_1, in the sense of being overlapped by both of them. Substituting this α into (21) yields

$$r\equiv\frac{z_1}{z_0}=\frac{z_{1/2}/z_0}{z_{1/2}/z_1}=\frac{E_0[q_{1/2}(\omega)/q_0(\omega)]}{E_1[q_{1/2}(\omega)/q_1(\omega)]}, \tag{35}$$

with the corresponding estimator

$$\hat{r}=\frac{(1/n_0)\sum_{i=1}^{n_0}[q_{1/2}(\omega_{0i})/q_0(\omega_{0i})]}{(1/n_1)\sum_{i=1}^{n_1}[q_{1/2}(\omega_{1i})/q_1(\omega_{1i})]}, \tag{36}$$

based on n_0 draws ω_{0i} from p_0 and n_1 draws ω_{1i} from p_1. That is, instead of applying (23) to directly estimate z_1/z_0, we apply it to first estimate $z_{1/2}/z_0$ and $z_{1/2}/z_1$ and then take the ratio to cancel $z_{1/2}$. The gain of efficiency arises because with a sensible choice of the "bridge" density $p_{1/2}$, there is less nonoverlap between p_t $(t=0,1)$ and $p_{1/2}$ than that between p_0 and p_1. That is, $p_{1/2}$ serves as a bridge between p_0 and p_1, hence the name *bridge sampling*. In terms of $q_{1/2}$, we see from (31) and (34) that the best bridge density is the (weighted) harmonic mean of p_0 and p_1:

$$q_{1/2}^{\text{opt}}=(s_1p_0^{-1}+s_0p_1^{-1})^{-1}=\frac{p_0p_1}{s_0p_0+s_1p_1},$$

and, interestingly, its normalizing constant determines the (asymptotic) minimal error given in (32). See Meng and Wong (1996) for a discussion of the relationship between bridge sampling and *umbrella sampling*, another method developed in computational physics (Torrie and Valleau, 1977), which has been termed *ratio importance sampling* by Chen and Shao (1997a, b) in the statistical literature. Also see Neal (1993, page 98) for a nice graphical representation of (35).

The idea of creating a bridge can obviously be pushed further. It is possible that the two densities $q_0(\omega)$ and $q_1(\omega)$ are so far separated that, even with the optimal bridge density $q_{1/2}^{\text{opt}}$, the estimator (36) is too variable to use in practice (or even does not exist if p_1 and p_0 are completely separated). In such cases, it is useful to construct a finite series of $L-1$ intermediate densities, from which we can make draws. For simplicity of later derivations, we label the corresponding unnormalized densities as $q(\omega|\theta_l)$, $l=0,1,\ldots,L$, including the two endpoints. For each pair of consecutive functions $q(\omega|\theta_l)$ and $q(\omega|\theta_{l+1})$, $l=0,\ldots,L-1$, we label the intermediate unnormalized density by $q(\omega|\theta_{l+1/2})$, which will be computed but not sampled from. With these $2L+1$ (unnormalized) densities, we can apply identity (35) in a telescoping fashion:

$$\begin{aligned}\frac{z_1}{z_0}\equiv\frac{z(1)}{z(0)}&=\prod_{l=1}^{L}\frac{z(\theta_{l-1/2})/z(\theta_{l-1})}{z(\theta_{l-1/2})/z(\theta_l)}\\&=\prod_{l=1}^{L}\frac{E_{\theta_{l-1}}[q(\omega|\theta_{l-1/2})/q(\omega|\theta_{l-1})]}{E_{\theta_l}[q(\omega|\theta_{l-1/2})/q(\omega|\theta_l)]}.\end{aligned} \tag{37}$$

As a generalization of (35), this is bridge sampling with $2L-1$ spans. (Meng and Wong, 1996, also present multiple-bridge identities for estimating more than one ratio of normalizing constants simultaneously; also see Geyer, 1994, for a "profile-likelihood" approach for estimating several ratios simultaneously.) Using intermediate systems (distributions) to implement importance sampling is also a well-known idea in computational physics (e.g., Neal, 1993, Section 6.2; Ceperley, 1995).

Given (37), one is tempted to study the limiting case when $L \to \infty$, that is, an infinite number of bridges. This can be easily done by considering the indexes θ_l as corresponding to a parameter $\theta \in [0, 1]$, indexing a parametric family $\{q(\omega|\theta), 0 \le \theta \le 1\}$, with $\theta_a = a/L$ for any $a \in [0, L]$. With this setup, taking logarithms of both sides of (37) yields

$$\log \frac{z_1}{z_0} = \sum_{l=1}^{L} \left[G_{l-1}\left(\frac{1}{2L}\right) - G_l\left(-\frac{1}{2L}\right) \right], \tag{38}$$

where the functions G_l are defined by

$$G_l(\xi) \equiv \log \int \frac{q(\omega|\theta_l + \xi)}{q(\omega|\theta_l)} p(\omega|\theta_l)\mu(d\omega), \qquad l = 0, 1, \ldots, L.$$

It is easy to verify that, for any l, $G_l(0) = 0$ and $G_l'(0) = E_{\theta_l}[U(\omega, \theta_l)]$, using the notation of Section 2.1, under the regularity condition that the support of $p(\omega|\theta)$ does not depend on θ. Thus, when $L \to \infty$, by the Taylor expansion of the right-hand side of (38), we have

$$\begin{aligned} \log \frac{z_1}{z_0} &= \lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} \frac{E_{\theta_{l-1}}[U(\omega, \theta_{l-1})] + E_{\theta_l}[U(\omega, \theta_l)]}{2} \\ &= \int_0^1 E_\theta[U(\omega, \theta)] d\theta, \end{aligned}$$

which is exactly the basic identity (7) underlying path sampling.

The foregoing derivation may be also helpful in studying the trade-off of implementation efficiency versus Monte Carlo efficiency in adopting multibridge sampling and path sampling, an open issue of practical interest.

4. A THEORETICAL INVESTIGATION OF PATH SAMPLING

4.1 Optimal Prior Density in One Dimension

The arbitrariness of the prior density $p(\theta)$ in (9) allows us to search for optimal estimators in the sense of achieving minimal Monte Carlo variances. Due to the difficulty of establishing general results under arbitrary sampling schemes, we shall assume independent draws for theoretical explorations and guidelines (but not for real implementations, as presented in Section 5).

If (ω_i, θ_i), $i = 1, \ldots, n$, in (9) are n independent draws from the joint distribution $p(\omega, \theta) = p(\omega|\theta)p(\theta)$, then the Monte Carlo variance of $\hat{\lambda}$ is

$$\begin{aligned} \operatorname{var}(\hat{\lambda}) &= \frac{1}{n}\left[\int_0^1 \int \frac{U^2(\omega, \theta)}{p^2(\theta)} \cdot p(\omega|\theta)p(\theta)\mu(d\omega)\, d\theta - \lambda^2 \right] \\ &= \frac{1}{n}\left[\int_0^1 \frac{E_\theta[U^2(\omega, \theta)]}{p(\theta)} d\theta - \lambda^2 \right]. \end{aligned} \tag{39}$$

Assume for now that $p(\omega|\theta)$ is given. Then we seek the the marginal (or prior) density $p(\theta)$ that minimizes (39), which is equivalent to minimizing the first term in (39).

By the Cauchy–Schwarz inequality,

$$\begin{aligned} &\int_0^1 \frac{E_\theta[U^2(\omega, \theta)]}{p(\theta)}\, d\theta \\ &\quad = \int_0^1 \left(\sqrt{p(\theta)}\right)^2 d\theta \int_0^1 \left(\frac{\sqrt{E_\theta[U^2(\omega, \theta)]}}{\sqrt{p(\theta)}} \right)^2 d\theta \\ &\quad \ge \left(\int_0^1 \sqrt{E_\theta[U^2(\omega, \theta)]}\, d\theta \right)^2. \end{aligned}$$

The right-hand side above does not depend on $p(\theta)$, and the equality holds when

$$p(\theta) = \frac{\sqrt{E_\theta[U^2(\omega, \theta)]}}{\int_0^1 \sqrt{E_{\theta'}[U^2(\omega, \theta')]}\, d\theta'}. \tag{40}$$

It follows that $p(\theta)$ of (40) is the optimal prior density, and the optimal variance of $\hat{\lambda}$ is

$$\operatorname{var}_{\text{opt}} = \frac{1}{n}\left[\left(\int_0^1 \sqrt{E_\theta[U^2(\omega, \theta)]}\, d\theta \right)^2 - \lambda^2 \right]. \tag{41}$$

Interestingly, when $z(\theta)$ is independent of θ, in which case $\lambda = 0$, the optimal density given in (40) is exactly the Jeffreys prior density (based on $p(\omega|\theta)$) restricted to $\theta \in [0, 1]$. In general, (40) can be viewed as a generalized local Jeffreys prior density based on the unnormalized density $q(\omega|\theta)$ (the expectation E_θ is with respect to the normalized density $p(\omega|\theta)$), and it is proper whenever $\int_0^1 \sqrt{E_\theta[U^2(\omega, \theta)]}\, d\theta < +\infty$. One can also view (40) as a "variance-stabilizing" transformation via the equation $p(\theta(t))\dot{\theta}(t) = 1$ between the path function and the prior density, although in the current setting the term "second-moment-stabilizing" transformation would be more appropriate because $E_\theta[U(\omega, \theta)]$ is generally not zero. The second-moment-stabilizing property can be seen by noticing that with the optimal prior (40), the second moment of each term of (9) is free of θ. This makes sense: the optimal procedure should balance the second sampling moment of $U(\omega, \theta)/p(\theta)$ at differ-

ent locations of θ, since we intend to minimize the average of them.

The variance given in (39) is not appropriate for the estimator $\hat{\lambda}(0, 1)$ given by (15), because of the use of an estimated $p(\theta)$. The derivations of the asymptotic variances for (15) and (17) are quite involved even under the independence assumption due the presence of (linear and nonlinear) functions of order statistics, and thus we do not discuss them here.

4.2 Optimal Path in Many Dimensions

Generalization of the above result to multivariate θ is immediate. For any *given* path, the optimal density for θ over that path is the generalized local Jeffreys prior density on that path. This does not, however, answer the question of which path is optimal in the sense of yielding minimal Monte Carlo variance of (11) among all possible paths. The answer to this question is a problem in the calculus of variations. Specifically, the variance of (11), under independent sampling, is

$$\begin{aligned}\operatorname{var}(\hat{\lambda}) &= \frac{1}{n}\Bigg[\int_0^1 \int\Bigg(\sum_{k=1}^{d} \dot{\theta}_k(t) U_k(\omega, \theta(t))\Bigg)^2 \\ &\qquad \cdot p(\omega|\theta)\mu(d\omega)\, dt - \lambda^2\Bigg] \\ &= \frac{1}{n}\Bigg[\int_0^1 \Bigg(\sum_{i,j=1}^{d} g_{ij}(\theta(t))\dot{\theta}_i(t)\dot{\theta}_j(t)\Bigg) dt - \lambda^2\Bigg], \qquad (42)\end{aligned}$$

where $g_{ij}(\theta) = E_\theta[U_i(\omega, \theta)U_j(\omega, \theta)]$. The path function $\theta(t)$ that minimizes the first term on the right-hand side of (42) is the solution of the following Euler–Lagrange equations (e.g., Atkinson and Mitchell, 1981) with the boundary condition $\theta(t) = \theta_t$, for $t = 0, 1$:

$$\sum_{i=1}^{d} g_{ik}(\theta(t))\ddot{\theta}_i(t) + \sum_{i,j=1}^{d} [ij, k]\dot{\theta}_i(t)\dot{\theta}_j(t) = 0, \qquad (43)$$
$$k = 1, \ldots, d,$$

where $\ddot{\theta}(t)$ denotes the second derivative with respect to t, and $[ij, k]$ is the Christoffel symbol of the first kind:

$$[ij, k] = \frac{1}{2}\left[\frac{\partial g_{ik}(\theta)}{\partial \theta_j} + \frac{\partial g_{jk}(\theta)}{\partial \theta_i} - \frac{\partial g_{ij}(\theta)}{\partial \theta_k}\right], \quad i, j, k = 1, \ldots, d.$$

Similar calculus of variations problems arise in the literature on finding the Rao distance between two densities (e.g., Rao, 1945, 1949; Atkinson and Mitchell, 1981; Mitchell, 1992). The Rao distance and the minimal variance of (11) are naturally related because the accuracy of the path sampling estimator depends crucially on the distance between two unnormalized densities, $q_0(\omega)$ and $q_1(\omega)$, and the Rao distance provides the appropriate measure. The Rao distance is constructed by considering the variance of the score function projected to a particular path, and thus the only difference between the Rao distance and the current calculation is that we are dealing with unnormalized densities. For example, (43) differs from (2.7) of Atkinson and Mitchell (1981) only by using $\log q(\omega|\theta)$ instead of $\log p(\omega|\theta)$ in defining the U functions inside g_{ij}. In fact, in the next section we show that the Rao distance in distribution space is directly related to the optimal path in distribution space.

As explored in the literature on the Rao distance, solving (43) is typically difficult. Atkinson and Mitchell (1981) suggested two alternative ways of expressing solutions via Hamilton's equations and Hamilton–Jacobi equations (e.g., Courant and Hilbert, 1961) and provided a differential geometry argument for finding the Rao distance between two normal densities. Despite these efforts, the general problem remains difficult. Section 4.4 provides a theoretical example for the normal distribution, where (43) has an analytic (but nontrivial) solution.

4.3 Optimal Path in Distribution Space

In the previous section, the family of distributions $p(\omega|\theta)$ is given. A different problem is to find an optimal path in the space of integrable nonnegative functions that connects the two unnormalized density functions, $q_0(\omega)$ and $q_1(\omega)$. That is, we seek to optimize over all nonnegative functions $q(\omega|\theta)$, with θ a scalar parameter having the range $[0, 1]$, subject to the boundary conditions, $q(\omega|0) = cq_0(\omega)$ and $q(\omega|1) = cq_1(\omega)$, where c is an arbitrary positive constant. Without loss of generality we can assume θ has a uniform distribution over $[0, 1]$ because of the "absorption" transformation discussed at the end of Section 2.1. In practice, it might be necessary to define a family of functions because $q_0(\omega)$ and $q_1(\omega)$ are not part of a common parametric family. Or, the two distributions might have a common parametric form, but a more efficient path may be possible by leaving the parametric form and moving through general distribution space. Possible general constructions include the geometric path (5) suggested in physics and the *scaling path* proposed by Ogata (1990, 1994) as an example of possible paths in the general distribution space of the form $q(\omega|\theta) = q_0(\omega)h_\theta(\omega)$:

$$\text{scaling path,} \quad q(\omega|\theta) = q_0(\omega)\frac{q_1(\theta\omega)}{q_0(\theta\omega)}. \qquad (44)$$

When using this path to estimate λ, we need to adjust for a known bias, $\log(q_0(0)/q_1(0))$. In our theoretical example in Section 4.4, the geometric path and scaling path lead to identical Monte Carlo error, which compares favorably to that from optimal bridge sampling but can be improved substantially within the path sampling framework. It is of great practical interest to find general simple paths with good properties.

Finding the optimal path in the whole distribution space turns out to be an easier mathematical problem than the optimization problem described in the previous section. We start by writing the path density as $q(\omega|\theta) = p(\omega|\theta)z(\theta)$ and expressing

$$\begin{aligned}\int_0^1 E_\theta\left[\frac{d}{d\theta}\log q(\omega|\theta)\right]^2 d\theta \\ = \int_0^1\left[\frac{d}{d\theta}\log z(\theta)\right]^2 d\theta \\ + \int_0^1 E_\theta\left[\frac{d}{d\theta}\log p(\omega|\theta)\right]^2 d\theta.\end{aligned} \tag{45}$$

To minimize the left-hand side of (45) over $q(\omega|\theta)$ we can separately minimize the two terms on the right-hand side, both under the appropriate boundary conditions. For the first term on the right-hand side, the following simple result, a consequence of the Cauchy–Schwarz inequality, provides the answer.

LEMMA 1. *If $z(\theta)$ is a positive function on $\theta \in [0, 1]$ such that $\log z(1) - \log z(0) = \lambda$, then*

$$\int_0^1\left[\frac{d}{d\theta}\log z(\theta)\right]^2 d\theta \geq \lambda^2,$$

with the equality holding if and only if $z(\theta)$ equals

$$z_{\text{opt}}(\theta) = z_0^{1-\theta}z_1^\theta \propto \exp(\lambda\theta) \tag{46}$$

almost surely with respect to the Lebesgue measure on $[0, 1]$.

This result implies that any $q(\omega|\theta)$ that yields $z(\theta) = \int q(\omega|\theta)\mu(d\omega)$ different from (46) cannot be optimal in the distributional space because $\tilde{q}(\omega|\theta) = q(\omega|\theta)[z_{\text{opt}}(\theta)/z(\theta)]$ dominates $q(\omega|\theta)$ (here we are discussing theoretical optimality, not the implementation feasibility). For example, the geometric path (5) is suboptimal in general because, for that path, $z(\theta) = \int q_0^{1-\theta}(\omega)q_1^\theta(\omega)\mu(d\omega) < z_0^{1-\theta}z_1^\theta$ for $0 < \theta < 1$.

The second term on the right-hand side of (45) is simply $\int_0^1 I(\theta)d\theta$, where $I(\theta)$ is the Fisher information for $p(\omega|\theta)$. Thus minimizing $\int_0^1 I(\theta)\,d\theta$ is the same as finding the Rao geodesic distance in the distribution space, a problem that can be solved by a differential geometry approach, as reviewed in Burbea (1989); also see Kass and Vos (1997). It turns out that, somewhat unexpectedly, the problem can also be solved via the Cauchy–Schwarz inequality, as we show in the Appendix. The result is (of course) identical to the result on the Rao distance in distribution space, though our expressions are more convenient for the path sampling application.

LEMMA 2. *Let $I(\theta)$ be the Fisher information for $p(\omega|\theta)$, where $p(\omega|0) \equiv p_0(\omega)$ and $p(\omega|1) \equiv p_1(\omega)$ are given. Let*

$$\alpha_H = \arctan\left[\frac{H(p_0, p_1)}{\sqrt{4 - H^2(p_0, p_1)}}\right], \tag{47}$$

where $H(p_0, p_1) = [\int(\sqrt{p_1(\omega)} - \sqrt{p_0(\omega)})^2\mu(d\omega)]^{1/2}$ is the Hellinger distance between p_0 and p_1. Then

$$\int_0^1 I(\theta)\,d\theta \geq 16\alpha_H^2, \tag{48}$$

and the equality in (48) holds if and only if

$$\begin{aligned}p(\omega|\theta) = \Bigg[\sqrt{p_0(\omega)}\left(\frac{\cos[(2\theta-1)\alpha_H]}{2\cos(\alpha_H)} - \frac{\sin[(2\theta-1)\alpha_H]}{2\sin(\alpha_H)}\right) \\ + \sqrt{p_1(\omega)}\left(\frac{\cos[(2\theta-1)\alpha_H]}{2\cos(\alpha_H)} + \frac{\sin[(2\theta-1)\alpha_H]}{2\sin(\alpha_H)}\right)\Bigg]^2,\end{aligned} \tag{49}$$

almost surely with respect to the product measure formed by μ and the Lebesgue measure on $[0, 1]$.

Applying the lemma with (45) and (46), we see that the optimal $q(\omega|\theta) = p(\omega|\theta)z(\theta)$ in the distributional space is given by

$$\begin{aligned}q(\omega|\theta) \propto e^{\lambda\theta}\Bigg[\sqrt{p_0(\omega)}\left(\frac{\cos[(2\theta-1)\alpha_H]}{2\cos(\alpha_H)} - \frac{\sin[(2\theta-1)\alpha_H]}{2\sin(\alpha_H)}\right) \\ + \sqrt{p_1(\omega)}\left(\frac{\cos[(2\theta-1)\alpha_H]}{2\cos(\alpha_H)} + \frac{\sin[(2\theta-1)\alpha_H]}{2\sin(\alpha_H)}\right)\Bigg]^2.\end{aligned} \tag{50}$$

The corresponding minimal variance is given by

$$\text{var}(\hat{\lambda}) = 16\alpha_H^2/n, \tag{51}$$

which is a simple function of $H(p_0, p_1)$. This intrinsic connection with the Hellinger distance also ap-

pears in the bridge sampling context, as Meng and Wong (1996) show that the optimal bridge-sampling error given in (32) is bounded below and above by simple functions of $H(p_0, p_1)$ when $s_0 = s_1$. Unlike bridge sampling where the optimal error is achieved by the iterative solution found with (33), however, it is unclear whether the optimal error in (51) is achievable (asymptotically) in practice since the optimal solution given in (50) assumes the knowledge of the unknown normalizing constants (and the ability to make independent draws from (50)). Using an adaptive method (e.g., iteratively estimating λ) may lead to an increase in variance. In fact, we doubt that (51) is achievable as it is bounded above by π^2/n even if p_0 and p_1 are infinitely apart. An interesting and empirically relevant problem is to figure out the achievable minimal error and how it varies with $H(p_0, p_1)$ (or some other distance measures). This is an important issue because an unachievable theoretical optimal error could misguide the choices of methods (e.g., contrast the theoretical comparisons in Chen and Shao, 1997a, with the empirical comparisons in Chen and Shao, 1997b, when the actual computational time needed by the ratio importance sampling method is taken into account).

Our interests in exploring these theoretical results lie in finding useful insights and practical guidelines about the potential and limits of path sampling. As we shall demonstrate in Section 4.4, where we solve (43) for a family of normal distributions, an optimal path can reduce the Monte Carlo variance by orders of magnitude when compared to some "natural" nonoptimal choices, a gain that is especially important when the two densities are far apart. We emphasize, however, it is not necessary to find an optimal path in order to gain substantial reduction in variance; for example, in the normal example, very simple paths reduce the variance by orders of magnitude compared to previous methods. The empirical implementations presented in Section 5 also demonstrate the superiority of path sampling with simple choices of paths.

4.4 A Theoretical Illustration

To illustrate theoretically the potential of path sampling for reducing Monte Carlo variances, we adopt the following example, which was used by Meng and Wong (1996) for illustrating bridge sampling. The example is of "toy" nature, but the findings are not and in fact somewhat surprised us. Let $q_0(\omega) = \exp(-\omega^2/2)$ and $q_1(\omega) = \exp(-(\omega - D)^2/2)$, where $D > 0$, and thus the true λ being "estimated" is zero. For the purpose of path sampling, we consider p_0 and p_1 as two points in the family of unnormalized normal densities:

$$q(\omega|\theta) = \exp\left(-\frac{(\omega-\mu)^2}{2\sigma^2}\right), \tag{52}$$

with $\theta = (\mu, \sigma)$, $\theta_0 = (0, 1)$ and $\theta_1 = (D, 1)$.

In order to make (nearly) fair comparisons, we assume that (i) with importance sampling, we make n draws from $N(0, 1)$, (ii) with bridge sampling, we make $n/2$ (assume n is even) draws from each of $N(0, 1)$ and $N(D, 1)$ and (iii) with path sampling, we first draw t_i, $i = 1, \ldots, n$, uniformly from [0,1]; then for each t_i we make one draw ω from $N(\mu(t_i), \sigma^2(t_i))$, where $\theta(t) = (\mu(t), \sigma(t))$ is a given path. All draws are independent within each scheme. In addition, since importance sampling and bridge sampling estimate the ratio r whereas path sampling estimates the log-ratio λ, we convert the estimates of r to the scale of λ by letting $\hat{\lambda} = \log \hat{r}$. Under this conversion, the variance of $\hat{\lambda}$ is asymptotically the same as the squared relative error of $\hat{r}$ (i.e., $E(\hat{r} - r)^2/r^2$). Under such a setting, Table 1 compares six estimators of λ, where the computations of $\sqrt{nE(\hat{\lambda} - \lambda)^2}$ are exact for the path sampling estimators and correct to terms of $O(n^{-1})$ for the others.

In Table 1, estimator (I) is the importance sampling estimator using (23), estimators (II) and (III) are bridge sampling using (35) with $q_{1/2} = \sqrt{q_0 q_1}$ and $q_{1/2} = (p_0^{-1} + p_1^{-1})^{-1}$, respectively, and the corresponding variance computations are from Meng

TABLE 1
Comparison of theoretical Monte Carlo errors of importance, bridge and path sampling estimators for two normal densities spaced D standard deviations apart

	Method	$\sqrt{nE(\hat{\lambda} - \lambda)^2}$
(I)	Importance sampling	$[\exp(D^2) - 1]^{1/2}$
(II)	Bridge sampling with geometric bridge	$2\left[\exp\left(\frac{D^2}{4}\right) - 1\right]^{1/2}$
(III)	Bridge sampling with (optimal) harmonic bridge	$2\left[\frac{1}{\sqrt{2\pi}\beta(D)} D \exp\left(\frac{D^2}{8}\right) - 1\right]^{1/2}$
(IV)	Path sampling with geometric (and scaling) path	$D^2\sqrt{\frac{1}{12} + \frac{1}{D^2}}$
(V)	Path sampling with optimal path in μ-space	D
(VI)	Path sampling with optimal path in (μ, σ)-space	$\sqrt{12}\left[\log\left(\frac{D}{\sqrt{12}} + \sqrt{1 + \frac{D^2}{12}}\right)\right]$

Note: In (III), $\beta(D) = (1/\pi)\int_0^\infty (\exp[-x^2/(2D^2)]/\cosh(x/2))\,dx$, with the property $\beta(D) \le 1$ and $\lim_{D\to\infty} \beta(D) = 1$.

and Wong (1996). Estimator (IV) is the path sampling estimator using the geometric path (5), which in this case leads to the identical estimator as that from the scaling path (44). Estimator (V) is the optimal univariate path sampling estimator based on (7) by considering σ to be fixed at 1 and letting μ vary in $[0, D]$. It is easy to verify that, in the current example, the (generalized) local Jeffreys prior density defined by (40) is uniform, so the optimal path function is given by $\mu(t) = Dt$. Estimator (VI) is the optimal multivariate path sampling estimator based on (10) when both μ and σ are allowed to very freely in their two-dimensional space; the form of the optimal path will be discussed shortly.

Figure 1 plots the six expressions in Table 1 as functions of $D \in [0, 10]$, where the dotted line plots $4\alpha_H$ (see (47)) with $H^2(p_0, p_1) = 2(1-\exp(-D^2/8))$, which is the lower bound from (51). As we discussed in Section 4.3, we doubt this bound can be achieved in reality, though we can easily improve estimator (VI) by using the optimal normalizing constants, given by (46), for the densities in the path. This entails using $\sigma^{-1}\exp\{-(\omega-\mu)^2/(2\sigma^2)\}$ in place of (52), and the resulting Monte Carlo error is obtained by replacing the values "12" with "8" in row (VI) of Table 1; the result is lower for all values of D but still is not optimal in distribution space.

The optimal path in (μ, σ)-space, denoted by $\theta(t) = (\mu(t), \sigma(t))$ (with $t \in [0, 1]$), turns out to be quite interesting and informative. Figure 2a, b plots $\mu(t)$ and $\sigma(t)$ with $D = 5$ as the boldfaced segments; the general expressions for $\mu(t)$ and $\sigma(t)$ and their derivations are given in the Appendix. This amounts to a half-ellipsoid curve in (μ, σ)-space:

$$(53) \qquad \left(\mu - \frac{D}{2}\right)^2 + 3\sigma^2 = 3 + \frac{D^2}{4}, \quad \sigma > 0,$$

as displayed, in boldface, in Figure 2c with $D = 5$. The optimal path thus increases the variances of the normal densities in the middle, which makes sense as we want the middle densities to have large overlaps with the two endpoint densities. However, the variance of the intermediate densities are not allowed to be arbitrarily large because that would introduce too much sampling variability at each given value of σ. The optimal path is a result of such a trade-off. This can be seen more clearly from Figure 2d, which displays the normalized normal densities corresponding to $q(\omega|\mu(t), \sigma(t))$ for $t = 0, 0.1, 0.2, \ldots, 1$, with the two end densities, $N(0, 1)$ and $N(5, 1)$, shown as boldfaced lines.

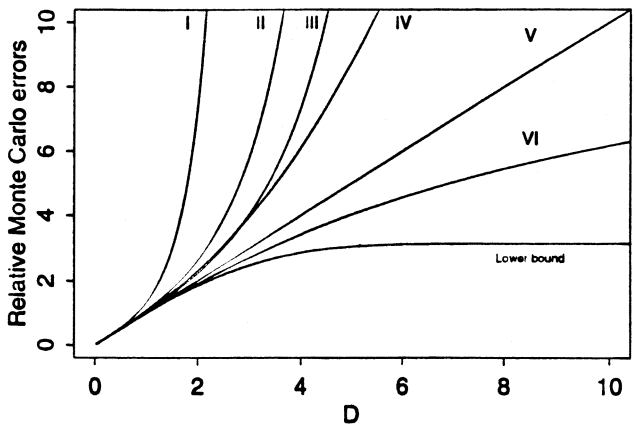

FIG. 1. *Relative Monte Carlo errors for various simulation-based estimates of* $\log(z_1/z_0)$, *comparing* $N(0, 1)$ *to* $N(D, 1)$ *densities, using* (I) *importance sampling*, (II) *bridge sampling with geometric bridge*, (III) *bridge sampling with optimal bridge*, (IV) *path sampling with geometric* (*or scaling*) *path*, (V) *optimal path sampling in* μ*-space and* (VI) *optimal path sampling in* (μ, σ)*-space. The dotted line is the lower bound given by* (51).

5. PRACTICAL IMPLEMENTATION AND EXAMPLES

5.1 Issues in Implementing Path Sampling

To implement path sampling, we need draws (ω, θ) from a joint distribution that can be written as $p(\omega, \theta) = q(\omega|\theta)/c(\theta)$. The marginal distribution of θ in these draws is

$$p(\theta) = \int p(\omega, \theta)\mu(d\omega) = \frac{z(\theta)}{c(\theta)}.$$

We have the freedom to specify $p(\theta)$ or $c(\theta)$, but not both (or else $z(\theta)$ would already be known). We emphasize that, depending on the nature of $c(\theta)$, $p(\theta)$ can be completely unrelated to $z(\theta)$, which we want to compute, or can be proportional or even identical to $z(\theta)$.

We distinguish between two kinds of implementations of path sampling. In the first kind, which is the subject of most of the preceding discussion, we specify $p(\theta)$, with the particular form chosen typically for reasons of convenience and perceived optimality. The simplest method of obtaining simulation draws is *direct sampling*, in which we draw (ω, θ) by first drawing θ from the known $p(\theta)$ (or choosing θ systematically over a grid, as in Ogata, 1989, and in our example of Section 5.2), then drawing ω from $q(\omega|\theta)$. The step of sampling θ is easy when we have the freedom to specify $p(\theta)$. Drawing ω given θ is usually more difficult, however: in many problems for which we would like to apply path sampling, there is no easy way to directly sample ω. Instead, the preferred method is some form of iterative simulation, such as the Metropolis algorithm. To implement path sampling using iterative

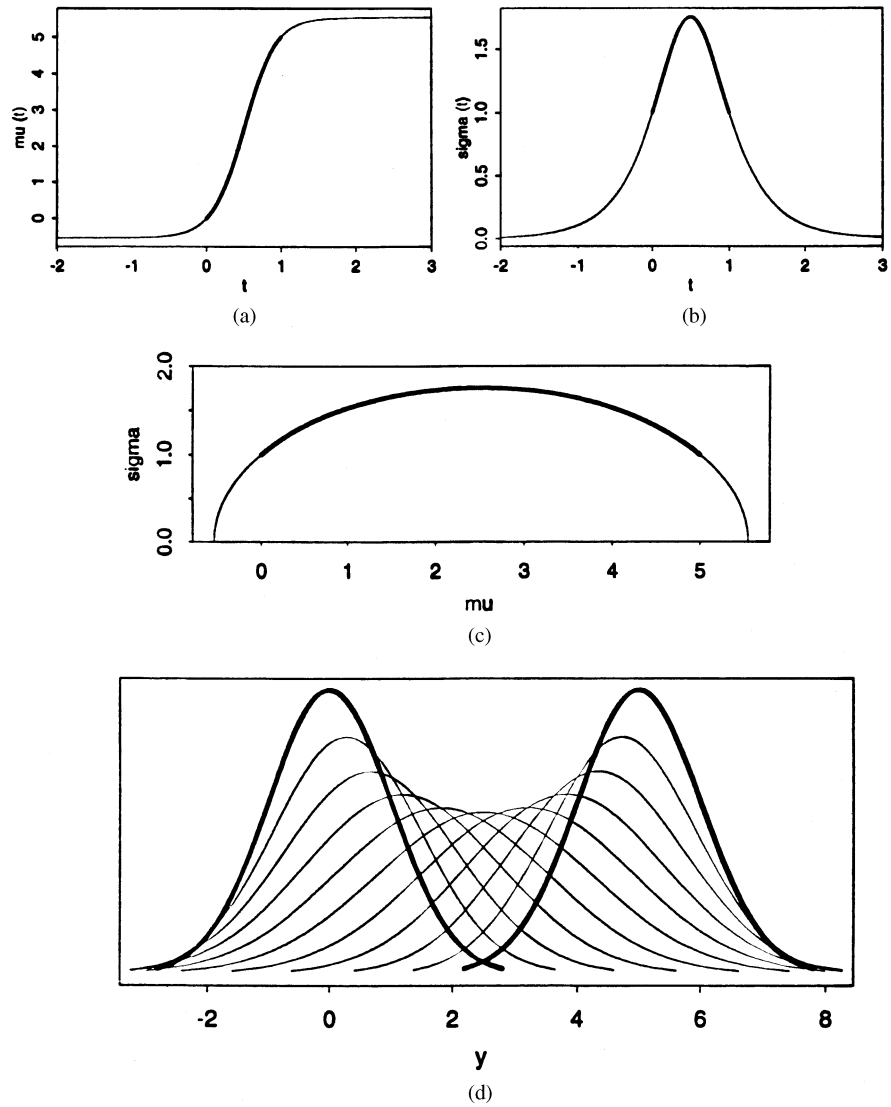

FIG. 2. *Optimal path from* $N(0, 1)$ *to* $N(5, 1)$ *in* (μ, σ)*-space*: (a) *parameterization of* $\mu(t)$; (b) *parameterization of* $\sigma(t)$; (c) *optimal path*; (d) *normalized densities along the optimal path.*

simulation, we can proceed directly by using *nested loops*: for each simulated (or chosen) θ, run an iterative simulation algorithm (until approximate convergence). The result is a large number of draws ω for each θ, but the estimator (9) is still applicable. The nested simulation approach may be attractive in a parallel computing environment. The nested-loop method has a flavor of the more elaborate *Metropolis-coupled Markov chain* method of Geyer (1991).

In the second kind of implementation, we specify $c(\theta)$ (up to a multiplicative constant). This means we can write the joint density of (ω, θ), up to a constant, and so can draw from this density, combining the simulations of ω and θ in a *single loop* of iterative simulation. The most natural approach here is to alternately update ω and θ in a Gibbs or Metropolis-type algorithm. This is essentially a special case of *simulated tempering* (see Marinari and Parisi, 1992, and Geyer and Thompson, 1995) with θ being viewed as the temperature variable, which is also similar to the *multicanonical* algorithms proposed in the statistical physics literature (e.g., Berg and Neuhaus, 1991; Berg and Celik, 1992); see Neal (1993, page 94) and Geyer and Thompson (1995) for more discussion.

Once the draws are made, we can apply (15) (in conjunction with (19) when θ is multivariate) to estimate $z(\theta)$ as a function of θ, as discussed in Section 2.3. (Interestingly, we can still directly estimate relative values of $z(\theta)$ using (15), without having to make any adjustment for $c(\theta)$.) A special case of

this kind of implementation is when drawing from a joint density given an unnormalized density $q(\omega, \theta)$, which means we have set $c(\theta) \equiv$ constant (and thus $p(\theta) \propto z(\theta)$); we illustrate this implementation with an example in Section 5.3.

When using the single-loop method, a new problem can arise when $z(\theta)$ is not to be used as a marginal density. The difficulty is that $z(\theta)$ can vary over several orders of magnitude in the region of θ of interest—this is not much a problem when $z(\theta)$ is used as a (unnormalized) marginal density if only regions of relatively high marginal mass are of interest. Simply sampling (ω, θ) from the joint distribution proportional to $q(\omega|\theta)$ (i.e., setting $c(\theta) \equiv$ constant) would leave very few draws of θ in regions of low marginal density and thus very little ability to compute $z(\theta)$ in those regions using path sampling or any other method. (For example, to estimate $z(b)/z(a)$, both (9) and (15) require draws of θ in the interval $[a, b]$.) Fortunately, in single-loop sampling we have the ability to choose $c(\theta)$ to reduce the variance of our estimators. Since we cannot, in general, easily compute the optimal $p(\theta)$—the generalized Jeffreys prior density—we aim for the simpler goal of a uniform $p(\theta)$ [i.e., the goal is $c(\theta)$ proportional to $z(\theta)$], which at least avoids the problem that some regions have far fewer draws than others. Several noniterative approaches are available for creating an approximation to $z(\theta)$, including Laplace's method (e.g., DiCiccio et al., 1997), the method of coding for conditional distributions (Besag, 1974) and various numerical methods (e.g., Evans and Swartz, 1995). The approximation here is used as $c(\theta)$ for making draws, not as our final estimate of $z(\theta)$, so the inaccuracy in the approximation does not bias our estimates from path sampling.

In cases where a reasonable approximation to $z(\theta)$ is not immediately available, we can update the function $c(\theta)$ iteratively. We start with some initial guess, say, $c(\theta) \equiv 1$. We then run the simulation of (ω, θ) using the Metropolis–Hastings algorithm. Occasionally, say, every few hundred iterations, we stop and estimate the function $z(\theta)$, either using path sampling (15) or some other density estimate of $p(\theta) = z(\theta)/c(\theta)$ based directly on the simulated values of θ. In either case, we update $c(\theta)$ to equal the current estimate of $z(\theta)$ and then continue the iteration. In the limit, under suitable mixing conditions, $c \to z$, and so $p(\theta)$ converges to uniformity. This iterative scheme is similar to the iterative method proposed in Geyer and Thompson (1995) for adjusting a pseudoprior needed for implementing simulated tempering. We emphasize that because we are using the estimator (15), which does not require $p(\theta)$ to be computed, the convergence of $c(\theta)$ is not actually required for the path-sampling estimators to be valid. This is reminiscent of the iterative sequence (33) for the optimal single-bridge estimator, where each iterate provides a valid estimate of r, and the iteration is needed only for the purpose of optimality.

One nice feature of the above iterative procedure is that the empirical distribution of the simulated θ values converges to a known distribution—uniform on $[0, 1]$. We can thus monitor the convergence of the simulations by comparing to this known distribution, which is far easier than the usual task of monitoring convergence to an unknown target distribution. In general, one can construct checks on the convergence of an iterative simulation as a by-product of path sampling. For example, when $c(\theta) = 1$, we can compare the estimate of $F(a)$ in (17) to the empirical distribution of the simulated values of θ (see Section 5.3); a discordance between these two distributions (as measured by some criterion of practical concern, such as a comparison of the 95% central posterior intervals) indicates a lack of convergence in the simulation (or an error in the implementation of the sampler or the path sampling estimate). Similar procedures are available with arbitrary chosen (known) $c(\theta)$. We can use such procedures to check the convergence of any parameter in the model (i.e., any parameter can take the role of θ with the others taking the role of ω in the analysis) by merely changing the derivative in U and recomputing the path sampling estimate.

5.2 Example 1: Censored Data in Spatial Statistics

The problem of high-dimensional integration commonly arises with missing or censored data. It is often the case that, given the uncensored data ω, the likelihood function $L(\theta|\omega) = p(\omega|\theta)$, is easy to compute. On the other hand, the likelihood based on the censored data y cannot be calculated directly. To fix ideas, we assume that $\omega = (\omega_1, \ldots, \omega_d)$ is a vector of real numbers, and the censored data $y = (y_1, \ldots, y_d)$ are given by $y_j = \max(\omega_j, 0)$ for $j = 1, \ldots, d$. The likelihood based on the censored data is then

$$p(y|\theta) = \int_{-\infty}^{0} \cdots \int_{-\infty}^{0} p(\omega|\theta) \prod_{j:\, y_j = 0} d\omega_j, \tag{54}$$

integrating over all the censored components. Treating $p(y|\theta)$ as the normalizing constant of $p(\omega|y, \theta)$ with the complete-data likelihood $p(\omega|\theta)$ as the unnormalized density, we are in the setting of (1).

For a particular example, we consider a stationary model in spatial statistics described by Stein (1992). In this example, each y_j is observed at a

location x_j in two-dimensional space. The vector of uncensored data ω is modeled by a joint normal distribution, in which each component ω_j has mean m and variance c, and the correlation between any two components, ω_i and ω_j, is $\exp(-|x_i - x_j|)$. Figure 1 of Stein (1992) presents a set of simulated data on a 6×6 grid evenly spread over the square $[0, 1]^2$, in which 17 of the 36 components y_j equal 0. The goal is to compute the likelihood of the parameter vector $\theta = (m, \log c)$. Were it not for the censored data, the likelihood would be trivial to compute from the joint normal density; however, because of the spatial dependence among the 36 observations, the 17-dimensional integral (54) cannot be calculated analytically.

Stein (1992) used importance sampling to compute the relative values of the marginal likelihood $p(y|\theta)$ on a 21×21 grid in the space of θ. At each point θ on the grid, Stein used a decomposition of a truncated multivariate normal distribution to construct an approximation $h_\theta(\omega)$ to $p(\omega|y, \theta)$, with known normalizing constant. He sampled 1000 draws of ω at each point of θ and estimated $p(y|\theta)$ by importance sampling. A crucial step that makes Stein's method work is that he used the *same* 1000 pseudorandom numbers for all draws at each θ, which was feasible because Stein sampled from the approximate densities h using an inverse-cdf approach. This introduces desirable positive dependence between the importance sampling estimates at the different points on the grid of θ, which, as Stein noted, greatly reduces the Monte Carlo error of the resulting ratio estimator.

Here we replicate Stein's results using path sampling with nested loops, which is computationally straightforward thanks to the simplicity of Gibbs sampler in this case. For each value of $\theta = (m, \log c)$ in the 21×21 grid, we use the Gibbs sampler to simulate from the conditional distribution of the uncensored values, $p(\omega|y, \theta)$. We monitored the convergence of parallel runs of the Gibbs sampler using the method of Gelman and Rubin (1992) and found that the simulations had reached approximate convergence after 100 iterations. We discard the first half of each simulation (i.e., we use 50 draws at each value of θ). We then use (15) in conjunction with (19) to estimate the function $\log[p(y|\theta)/p(y|\theta_0)]$ on the 21×21 grid of $(m, \log c)$, where θ_0 is the maximum likelihood estimate. Figure 3a gives the contour plot of the estimated negative log-likelihood ratio when we integrate $\log c$ first in applying (19). Figure 3b shows the corresponding plot when we integrate m first. The effects of the two different paths are quite visible in this case, as Figure 3b gives a much smoother answer and is almost identical to the plot given in Figure 3 of Stein (1992). The path sampling method used here does not involve constructing approximate densities and does not need to use the same pseudorandom numbers at different points of θ.

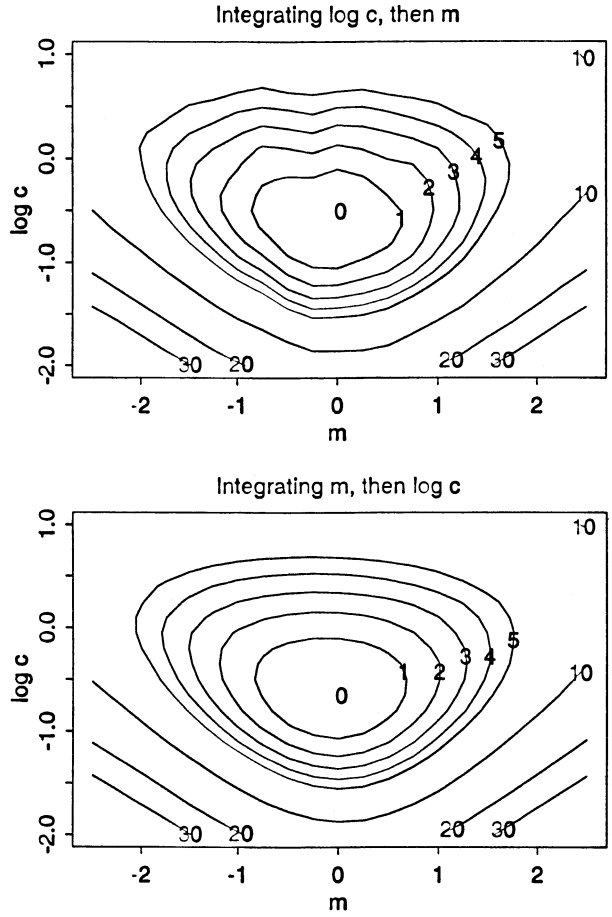

FIG. 3. *Estimated negative loglikelihood for spatial statistics example (replication of Figure 3 of Stein, 1992) using path sampling, with a 21×21 grid and only 100 draws of ω at each point. The two plots show the estimates based on two different paths.*

5.3 Example 2: Heteroscedastic Regression Models for Election Forecasting

5.3.1 *Statistical model and substantive background.* Consider the heteroscedastic regression model,

$$y \mid \beta, \sigma, \theta \sim N(X\beta, \sigma^2 r^{-\theta}), \tag{55}$$

where r is a vector of weights, and $\theta \in [0, 1]$ is a model parameter. Boscardin and Gelman (1996) use this model for forecasting U.S. Presidential elections, with units i representing states and election years, y_i the Democratic Party's share of the vote in the state in that year, X a matrix of predictors and r_i proportional to the number of voters in the state in that year. The predictors X used in the regres-

sion are chosen based on existing regression models used in political science.

The values $\theta = 0$ and 1 represent two extreme models that have been considered, implicitly or explicitly, in political science. Setting $\theta = 0$ corresponds to equal residual variances for the 50 states, which is generally assumed in forecasting and regression models of elections in political science research, perhaps for convenience as much as any other reason. Setting $\theta = 1$, so that variance is inversely proportional to the number of voters, has theoretical appeal as a generalization of the binomial model, which is implicit in many game-theoretic models of voting. One of the major trends in recent research in political science is to unify empirical and theoretical analyses; here, the value of θ is an issue that needs to be resolved, for reasons both applied (obtaining efficient forecasts and regression estimates) and theoretical (understanding the variability of voters in the aggregate). See Gelman, King and Boscardin (1998) for a discussion of these issues, along with many references from political science and economics on these models.

At this point, statistical practice suggests several different ways of using the data to assess the information of the data about the parameter θ. Classical approaches include (a) using significance tests to accept or reject the null hypothesis $\theta = 0$ against the alternative, $\theta = 1$ (or vice-versa); and (b) obtaining an approximately unbiased point estimate of θ, considering it as a nonlinear estimation problem with nuisance parameters. Bayesian approaches include (c) choosing between $\theta = 0$ and $\theta = 1$, using the Bayes factor to assess the relative evidence in favor of the two possibilities; and (d) including θ as a continuous parameter in the model (taking the range [0,1]) and computing its posterior distribution.

Of these four approaches, (a) and (c) involve nearly identical computations, since both are based on the distribution of the likelihood ratio under the two candidate models. In the context of simulation-based inference, approach (c) would be more natural, and it would involve the computation of a ratio of marginal densities, which, as discussed in Section 1, is equivalent to a ratio of normalizing constants. This computation could be done using any of the methods described in this paper; to the extent that the likelihoods under the two models ($\theta = 0$ and $\theta = 1$) are far apart (which will generally be the case as the number of data points increases), it would be advisable to consider path sampling. A natural choice of path is (55) with θ varying from 0 to 1.

However, in this application, we prefer to consider θ to be a continuous parameter from the start, because we wish to consider the possibilities of models that fall between the two extremes $\theta = 0$ and $\theta = 1$: that is, perhaps there is some truth in both of the existing approaches. (A theoretical argument for allowing θ to vary is that its appropriate value might very well depend on the set of explanatory variables X used in the model, so that, e.g., the game-theoretic descriptions might be more or less accurate depending on what information is assumed to be known.) Now that we are allowing θ to be uncertain on a continuous range, the classical estimation approach has some serious problems, most notably that the likelihood can be extremely flat (parameters of the variance model, such as θ in this example, can often be poorly identified in data sets of moderate size) so that no point estimate is an accurate summary. Along with this is the possibility that the point estimate could be outside [0,1] just due to high variability or, if the estimate is constrained, that it could be on the boundary. For example, it is possible to have a point estimate at $\theta = 0$ even though $\theta = 1$ is also well-supported by the data. These problems get more serious in the presence of nuisance parameters (e.g., when β contains random effects components). For all these reasons, we prefer a Bayesian approach of summarizing the information about θ by a posterior distribution (or, to use non-Bayesian terminology, a marginal likelihood, since we shall use the uniform prior distribution, $p(\theta) = 1$). As discussed in Section 1, determining the marginal posterior density of θ is mathematically equivalent to computing a normalizing constant parameterized by $\theta \in [0, 1]$. Because we are constraining $\theta \in [0, 1]$, it is also important to examine the behavior of the likelihood near the boundary to see if there is evidence that $\theta < 0$ or $\theta > 1$.

We consider two methods of computing the posterior density, or marginal likelihood, or normalizing constant, as a function of θ: (i) the usual approach of Bayesian simulation, which is to consider θ as a parameter in the model and then summarize its posterior distribution by the empirical distribution of its simulation draws, as was done by Boscardin and Gelman (1996); and (ii) path sampling. In fact, we shall use the simulation draws from (i) to implement path sampling and then compare the estimated marginal posterior distribution for θ under path sampling to the direct estimates from the simulation draws.

5.3.2 *Path sampling for the nonhierarchical model.* To check the performance of path sampling, we first consider the simple nonhierarchical model, which assigns a uniform prior distribution to $(\beta, \log \sigma, \theta)$. As discussed in Boscardin and Gelman

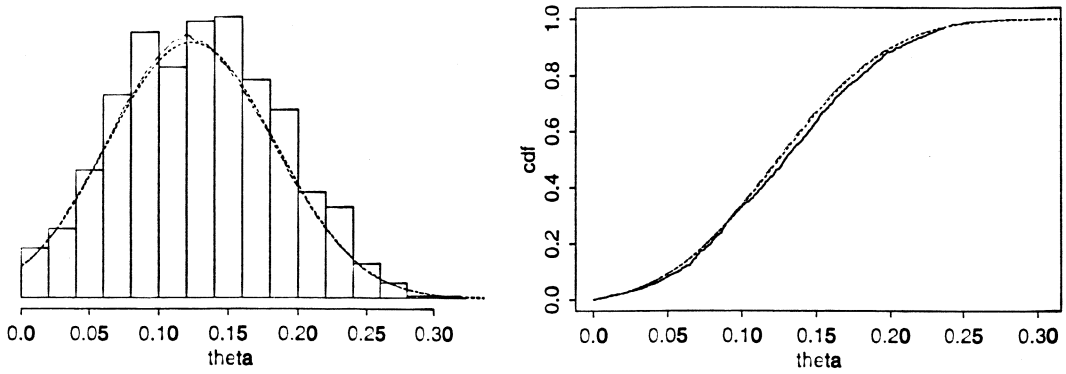

FIG. 4. *Estimates of the density function and cdf of the heteroscedasticity parameter θ for the election forecasting example with nonhierarchical model*: (a) *dashed line is exact density, dotted line is estimated density from path sampling and histogram is from* 1000 *iid simulation draws*; (b) *dashed line is exact cdf, dotted line* (*almost exactly on top of dashed line*) *is estimated cdf from path sampling and solid line is empirical cdf from simulation draws.*

(1996), the (unnormalized) marginal posterior density for θ in this model can be written analytically, and posterior draws for the vector of parameters can be obtained directly by first drawing θ from a discrete approximation to its numerically calculated marginal posterior distribution, then drawing (β, σ) from their normal-inverse-χ^2 posterior distribution conditional on the drawn θ. For the election example, this was done using 1000 independent draws of θ and 2 draws of (β, σ) for each draw of θ. In our general notation, $\omega = (\beta, \sigma)$, which is 20-dimensional because β has 19 components. To compute the path sampling estimate of $p(\theta|y)$, we must first determine the function $U(\omega, \theta)$; the differentiation is easy and yields

$$U(\omega, \theta) = -\frac{1}{2\sigma^2} \sum_i (\log r_i) r_i^{\theta} \big(y_i - (X\beta)_i\big)^2. \tag{56}$$

We use the simulation draws, which have the nested-loop form, to compute the path sampling estimate of $p(\theta|y)$. Figure 4a shows the results, comparing the exact density (smooth line), path sampling estimate (slightly jagged line) and the histogram from the 1000 simulation draws. The path sampling estimate is (of course) worse than the exact density but compares much more favorably to the histogram estimate. Another comparison is afforded by Figure 4b, which shows the corresponding cdf's. Here, the jagged line is the empirical cdf of the 1000 draws, and the smooth line represents both the path sampling estimate using (17) and the exact cdf—the differences are barely visible! Obviously, in this case, one can simply use the exact formula, but it is informative and encouraging to be able to confirm that path sampling is capable of producing such an accurate approximation to a 20-dimensional integration indexed by an entire curve with only 2000 draws in total.

5.3.3 *Path sampling for the hierarchical model.* We now move to a more realistic, and thus more complicated, model fitted by Boscardin and Gelman (1996) in which the marginal density for θ cannot be computed analytically. In this model, 50 additional components of β are added (and thus we are now dealing with a 70-dimensional integration), along with a hierarchical regression model and additional variance components. For this expanded model, the marginal posterior density of θ cannot be computed exactly, and posterior simulations are obtained using the Gibbs sampler and the Metropolis algorithm, alternating between Metropolis jumps for θ and Gibbs draws for the remaining parameters. Approximately overdispersed starting points for the algorithm are obtained by a t_4 approximation to the posterior distribution. The Gibbs sampler draws are performed using linear regression operations and simulations of normal and χ^2 random variables, and the Metropolis steps use a univariate normal jumping kernel with a scale set to 2.38 times the estimated standard deviation of θ from the initial approximation (motivated by Gelman, Roberts and Gilks, 1996).

For the election example, 10 sequences, each of length 500, were sufficient for approximate convergence of the simulations as monitored using the methods of Gelman and Rubin (1992). To compute the path sampling estimate of $p(\theta|y)$, we again need the function U, which actually has the same form (56) as before, since the added hierarchical part of the model does not involve the parameter θ. In this case, the simulation draws have the single-loop

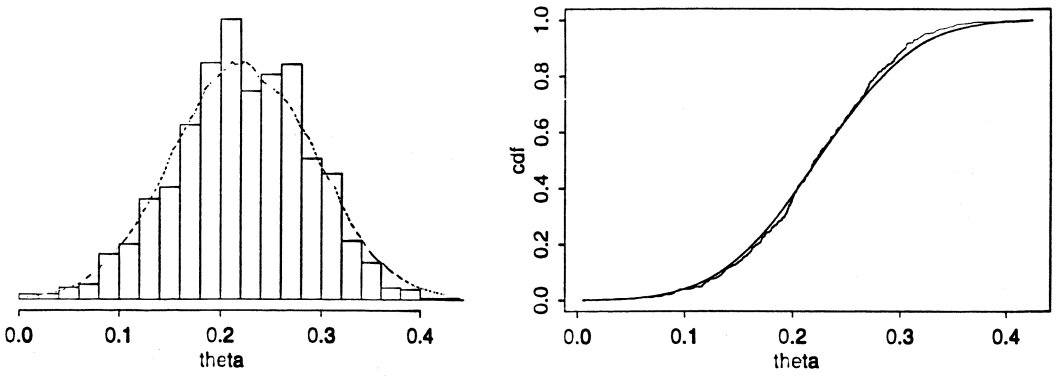

FIG. 5. *Estimates of the density function and cdf of the heteroscedasticity parameter θ for the election forecasting example with hierarchical model*: (a) *dotted line is estimated density from path sampling and histogram is from* 1250 *simulation draws from Metropolis algorithm*; (b) *smooth line is estimated cdf from path sampling, and jagged line is empirical cdf from simulation draws.*

form, meaning that we can only learn about $p(\theta|y)$ for the range of θ's that were obtained in the simulation. Figure 5a shows the estimated marginal posterior density from path sampling and the histogram of simulation draws, and Figure 5b shows the corresponding estimated cdf's. The path sampling estimates are far smoother. Based on the evidence given in Figure 4a, b we can be quite confident that the path sampling estimates are very close to the truth, especially for the cdf. For both these models, the smoothness of the path sampling estimates is an intrinsic property of the estimation procedure—despite their appearance, no smoothing was used in creating the estimates.

6. SUMMARY AND FURTHER RESEARCH

This paper attempts to bring to the attention of statistical researchers some useful methods for computing normalizing constants for complex, high-dimensional probability models, or more generally computing high-dimensional integrations with complicated integrands. Both bridge and path sampling are rooted in popular methods in theoretical physics, namely, the acceptance ratio method and thermodynamic integration. Due to extremely challenging and important computational problems in physics and chemistry, some of which are far from being resolved (e.g., minimum energy configurations of protein molecules), there is a huge literature in theoretical and computational physics and chemistry on creative methods for high-dimensional integration and optimization. For an excellent, though not necessarily "statistician friendly," recent review of a good number of these powerful methods, see Ceperley's (1995) long review article on path integrals in the theory of condensed helium. In particular, the methods discussed in Section V of Ceperley (1995) are potentially very useful for implementing path sampling in general—indeed, the phrase "path sampling" is used there to describe sampling methods for simulating path integrals for the so-called thermal density matrix.

Given its great success in theoretical physics for dealing with complex integrations, as well as Ogata's (1990, 1994) successful applications to Bayesian computations, we believe that path sampling can be generally useful in statistical computations for dealing with complex integrations. The method is not only capable of producing remarkably accurate results but is also quite straightforward, in the class of methods that are useful for high-dimensional complex integrations. As Frenkel (1986, page 169) states in his review chapter on free-energy estimation: "Thermodynamic integration (TI) is undoubtedly the method most widely used to compute absolute free energies and free-energy differences. The reason is that, although it may be more time consuming than some of the sophisticated methods described above, it is straightforward, accurate and does not run into special problems at high densities or for large system sizes." We hope our simple (though not trivial) empirical illustrations, as well as the theoretical example, have helped to convey these messages to statistical researchers; the quote also makes it clear that thermodynamic integration (and hence path sampling) does not dominate other methods. Furthermore, we hope our derivation of how path sampling relates to importance sampling via bridge sampling will help general statistical readers to understand the method intuitively and thus be able to apply it with more confidence. In a statistical context, path sampling also gives an alternative

method of estimating marginal distributions and offers an effective check on the convergence of Monte Carlo simulations.

Our general formulation and investigation also reveals that further research is needed in order to explore fully the potential of path sampling. For example, the construction of efficient yet simple general paths is of great importance for routine application of path sampling. Our theoretical results (e.g., the general suboptimality of the geometric path and, for the normal example, the optimal path that curves through the space of (μ, σ)) show that the best paths are not always obvious. The question of achievable optimal error is not only of theoretical interest but also of practical relevance if we can construct an easily implementable iterative procedure, just as with bridge sampling, to compute the optimal estimate. Such questions are inherently statistical, and thus we statisticians should be able to contribute substantially to the study of efficient implementation of path sampling, especially in view of the theoretical relations between optimal paths and the Jeffreys prior and the Rao and Hellinger distances. With path sampling, as with Markov chain Monte Carlo methods—another statistical tool that originated in computational physics—there is the potential not only to benefit from a powerful method but also to make it more efficient and applicable, thus broadening the range of statistical models that we can use routinely.

APPENDIX

A.1 An Elementary Proof of Lemma 2

PROOF. Let $g(\theta)$ be a differentiable positive function on $[0, 1]$ such that $g(0) = g(1) = 1$, and let $h(\omega|\theta) = p(\omega|\theta)g(\theta)$. Then, by Fubini's theorem, we can verify that

$$\begin{aligned}\int_0^1 I(\theta)\,d\theta &= 4\int\left[\int_0^1\left(\frac{\partial\sqrt{h(\omega|\theta)}}{\partial\theta}\frac{1}{\sqrt{g(\theta)}}\right)^2 d\theta\right]\mu(d\omega) \\ &\quad - \int_0^1\left(\frac{d\log g(\theta)}{d\theta}\right)^2 d\theta. \end{aligned} \tag{57}$$

By the Cauchy–Schwarz inequality, the first term on the right-hand side of (57) is bounded below by $4H^2(p_0, p_1)/\int_0^1 g(\theta)\,d\theta$ with the bound achieved if and only if

$$\frac{\partial\sqrt{h(\omega|\theta)}}{\partial\theta}\frac{1}{\sqrt{g(\theta)}} = b(\omega)\sqrt{g(\theta)} \quad \text{a.s. } (\mu \times \text{Lebesgue on } [0,1]), \tag{58}$$

where $b(\omega)$ is a positive function to be determined. Solving (58) for $p(\omega|\theta) = h(\omega|\theta)/g(\theta)$ with the given boundary condition yields

$$\begin{aligned}p(\omega|\theta) = \Big\{&\big[\sqrt{p_0(\omega)}(1 - G(\theta)/G(1)) \\ &+ \sqrt{p_1(\omega)}(G(\theta)/G(1))\big]^2/g(\theta)\Big\},\end{aligned} \tag{59}$$

where $G(\theta) = \int_0^\theta g(\xi)\,d\xi$. The freedom in choosing $g(\theta)$ allows us to ensure that $p(\omega|\theta)$ of (59) is a proper density for any $\theta \in [0, 1]$, a requirement that leads to a differential equation for $G(\theta)$:

$$\begin{aligned}G'(\theta) = 1 &- \frac{H^2(p_0, p_1)}{4} \\ &+ H^2(p_0, p_1)\left(\frac{G(\theta)}{G(1)} - \frac{1}{2}\right)^2.\end{aligned} \tag{60}$$

Solving (60) for $g(\theta) = G'(\theta)$ with the boundary condition $g(0) = g(1) = 1$ yields

$$g(\theta) = \frac{\cos^2(\alpha_H)}{\cos^2[(2\theta - 1)\alpha_H]}.$$

The rest of the proof follows by simple algebraic manipulation. □

A.2 Derivation of the Optimal Path in (μ, σ)-Space for the Normal Example

Here we derive the optimal path in the normal family example of Section 4.4 in the general case of any endpoints $\theta_0 = (\mu_0, \sigma_0)$ and $\theta_1 = (\mu_1, \sigma_1)$; without loss of any generality, we assume $\mu_1 \geq \mu_0$. We start by noting that, with the normal family (52), the variance formula (42) becomes

$$\operatorname{var}(\hat{\lambda}) = \frac{1}{n}\left[\int_0^1 \frac{\dot{\mu}^2(t) + 3\dot{\sigma}^2(t)}{\sigma^2(t)}\,dt - \lambda^2\right]. \tag{61}$$

The corresponding Euler–Lagrange equations (43) for the optimal path can be simplified into

$$\dot{\mu}(t) - c_0\sigma^2(t) = 0, \tag{62}$$

$$3\ddot{\sigma}(t)\sigma(t) - 3\dot{\sigma}^2(t) + \dot{\mu}^2(t) = 0, \tag{63}$$

where c_0 is a constant to be determined by the boundary conditions: $\mu(t) = \mu_t$, $\sigma(t) = \sigma_t$, $t = 0, 1$. This differential equation can be solved using a differential geometric argument developed in Atkinson and Mitchell (1981) or directly as follows.

We first substitute (62) into (63) and obtain

$$3\ddot{\sigma}(t)\sigma(t) - 3\dot{\sigma}^2(t) + c_0^2\sigma^4(t) = 0. \tag{64}$$

We then let $v = \dot{\sigma}(t)$ and express $v = v(\sigma)$ via $t = t^{-1}(\sigma)$, which yields

$$\ddot{\sigma}(t) = \frac{dv}{dt} = \frac{dv}{d\sigma}\frac{d\sigma}{dt} = \dot{v}(\sigma)v(\sigma). \tag{65}$$

Combining (65) with (64) gives

$$\frac{d}{d\sigma}\left[\frac{v^2(\sigma)}{\sigma^2}\right] + \frac{2}{3}c_0^2\sigma = 0, \tag{66}$$

which implies

$$\dot{\sigma}(t) \equiv v(\sigma) = \sqrt{c_1\sigma^2 - \frac{c_0^2}{3}\sigma^4}, \tag{67}$$

where c_1 is a constant to be determined.

When $c_0 \neq 0$, (67) leads to

$$\begin{aligned} t &= \int \frac{d\sigma}{\sqrt{c_1\sigma^2 - (c_0^2/3)\sigma^4}} \\ &= \frac{1}{\sqrt{c_1}}\cosh^{-1}\left[\frac{\sqrt{3c_1}}{c_0\sigma(t)}\right] + c_2, \end{aligned}$$

where c_2 is another constant to be determined. It follows that

$$\sigma(t) = \frac{\sqrt{3c_1}}{c_0}\operatorname{sech}[\sqrt{c_1}(t - c_2)]. \tag{68}$$

Finally, combining (68) with (62) yields

$$\begin{aligned} \mu(t) &= c_0\int \sigma^2(t)\,dt \\ &= \frac{3\sqrt{c_1}}{c_0}\tanh[\sqrt{c_1}(t - c_2)] + c_3, \end{aligned} \tag{69}$$

where the constant c_3 is determined, along with c_0, c_1 and c_2, by the boundary conditions. The solution is then given by

$$\begin{aligned} \mu(t) &= R\tanh[\phi_0(1-t) + \phi_1 t] + C, \\ \sigma(t) &= \frac{R}{\sqrt{3}}\operatorname{sech}[\phi_0(1-t) + \phi_1 t], \end{aligned} \tag{70}$$

where

$$\begin{aligned} R^2 &= \left(\frac{\mu_0 - \mu_1}{2}\right)^2 + \frac{3}{2}(\sigma_1^2 + \sigma_0^2) + \frac{9}{4}\left(\frac{\sigma_1^2 - \sigma_0^2}{\mu_1 - \mu_0}\right)^2, \\ C &= \frac{\mu_0 + \mu_1}{2} + \frac{3}{2}\frac{\sigma_1^2 - \sigma_0^2}{\mu_1 - \mu_0}, \\ \phi_t &= \tanh^{-1}\left(\frac{\mu_t - C}{R}\right) \\ &= \frac{1}{2}\log\frac{R + \mu_t - C}{R - \mu_t + C} \quad \text{for } t = 0, 1. \end{aligned} \tag{71}$$

This implies a path in (μ, σ)-space of the form

$$(\mu - C)^2 + 3\sigma^2 = R^2, \tag{72}$$

which reduces to (53) when $\sigma_0 = \sigma_1 = 1$, $\mu_0 = 0$ and $\mu_1 = D$. The case of $c_0 = 0$ corresponds to the special case $\mu_0 = \mu_1$, in which case the solution is

$$\mu(t) \equiv \mu_0, \quad \sigma(t) = \sigma_1^t\sigma_0^{1-t} \quad \text{for } 0 \le t \le 1.$$

The solution (70) also induces the optimal prior densities on the optimal path; for μ, it is

$$p(\mu) = \frac{1}{R(\phi_1 - \phi_0)}\,\frac{1}{1 - [(\mu - C)/R]^2}, \qquad \mu_0 \le \mu \le \mu_1, \tag{73}$$

and for σ,

$$p(\sigma) = \frac{2}{\phi_1 - \phi_0}\,\frac{1}{\sigma\sqrt{1 - (3\sigma^2/R^2)}}, \qquad \min(\sigma_0, \sigma_1) \le \sigma \le \sigma_{\max}, \tag{74}$$

where

$$\sigma_{\max} = \begin{cases} \dfrac{R}{\sqrt{3}}, & \text{if } \mu_0 \le C \le \mu_1, \\ \max(\sigma_0, \sigma_1), & \text{otherwise.} \end{cases}$$

The density of σ has an asymptote at $\sigma_{\max}$ because $\dot{\sigma}(t) = 0$ at $t = \sigma_{\max}$. (In our example, $\sigma_{\max}^2 = 1 + D^2/12$.) The optimal error associated with the optimal path can be easily obtained, using the fact that, on the path, the integrand inside (61) is free of t due to the "second-moment-stabilizing" transformation (see Section 4.1). The optimal error is given, in general, by

$$\text{var}_{\text{opt}} = \frac{1}{n}\left[3(\phi_1 - \phi_0)^2 - \left(\log\frac{\sigma_1}{\sigma_0}\right)^2\right],$$

where ϕ_t, $t = 0, 1$, are given by (71). In our example, it is simplified to

$$\text{var}_{\text{opt}} = \frac{1}{n}\left[\sqrt{12}\log\left(\frac{D}{\sqrt{12}} + \sqrt{1 + \frac{D^2}{12}}\right)\right]^2,$$

because $\mu_0 = 0$, $\mu_1 = D$ and $\sigma_0 = \sigma_1 = 1$.

ACKNOWLEDGMENTS

We thank D. Ceperley, P. McCullagh, R. Neal, M. Stein and W. Wong for helpful conversations, and Y. Ogata for sending (pre)reprints. Gelman's research was supported in part by NSF Grants DMS-94-04305, SBR-97-08424 and Young Investigator Award DMS-94-57824, and by fellowship F/96/9 of Katholike Universiteit Leuven. Meng's research was supported in part by NSF Grants DMS-95-05043 and DMS-96-26691, and NSA Grant MDA 904-96-1-0007.

REFERENCES

ATKINSON, C. and MITCHELL, A. F. S. (1981). Rao's distance measure. *Sankhyā Ser. A* **43** 345–365.

BENNETT, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **22** 245–268.

BERG, B. and CELIK, T. (1992). New approach to spin-glass simulations. *Phys. Rev. Lett.* **69** 2292–2295.

BERG, B. and NEUHAUS, T. (1991). Multicanonical algorithms for the first order phase transitions. *Phys. Lett. B* **267** 249–253.

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

BINDER, K. (1986). Introduction: theory and technical aspects of Monte Carlo simulations. In *Monte Carlo Methods in Statistical Physics* (K. Binder, ed.). *Topics in Current Physics* **7**. Springer, Berlin.

BOSCARDIN, W. J. and GELMAN, A. (1996). Bayesian regression with parametric models for heteroscedasticity. *Advances in Econometrics* **11A** 87–109.

BURBEA, J. (1989). Rao distance. In *Encyclopedia of Statistical Science*, supplement volume, 128–130. Wiley, New York.

CEPERLEY, D. M. (1995). Path integrals in the theory of condensed helium. *Rev. Modern Phys.* **67** 279–355.

CHEN, M.-H. and SHAO, Q. M. (1997a). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.* **25** 1563–1594.

CHEN, M.-H. and SHAO, Q. M. (1997b). Estimating ratios of normalizing constants for densities with different dimensions. *Statist. Sinica* **7** 607–630.

CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90** 1313–1321.

CICCOTTI, G. and HOOVER, W. G., eds. (1986). *Molecular-Dynamics Simulation of Statistical-Mechanical Systems.* North-Holland, Amsterdam.

COURANT, R. and HILBERT, D. (1961). *Methods of Mathematical Physics* **2**. Wiley, New York.

DEMPSTER, A. P., SELWYN, M. R. and WEEKS, B. J. (1983). Combining historical and randomized controls for assessing trends in proportions. *J. Amer. Statist. Assoc.* **78** 221–227.

DICICCIO, T. J., KASS, R. E., RAFTERY, A. and WASSERMAN, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.*. **92** 903–915.

EVANS, M. and SWARTZ, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statist. Sci.* **10** 254–272.

FRANKEL, D. and SMIT, B. (1996). *Understanding Molecular Simulation*. Academic Press, New York.

FRENKEL, D. (1986). Free-energy computation and first-order phase transition. In *Molecular-Dynamics Simulation of Statistical-Mechanical Systems* (G. Ciccotti and W. G. Hoover, eds.) 151–188. North-Holland, Amsterdam.

GELFAND, A. E. and DEY, D. K. (1994). Bayesian model choice: asymptotic and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56** 501–514.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

GELMAN, A., KING, G. and BOSCARDIN, J. (1998). Estimating the probability of events that have never occurred: when is your vote decisive? *J. Amer. Statist. Assoc.* **93** 1–9.

GELMAN, A. and MENG, X. L. (1994). Path sampling for computing normalizing constants: identities and theory. Technical Report 376, Dept. Statistics, Univ. Chicago.

GELMAN, A., ROBERTS, G. O. and GILKS, W. R. (1996). Efficient Metropolis jumping rules. In *Bayesian Statistics 5* (J. O. Berger, J. M. Bernardo, D. V. Lindley and A. F. M. Smith, eds.) 599–607. Oxford Univ. Press.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7** 457–511.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

GEYER, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (E. M. Keramidas, ed.) 156–163. Interface Foundation, Fairfax Station, VA.

GEYER, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical Report 568, School of Statistics, Univ. Minnesota.

GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 657–699.

GEYER, C. J. and THOMPSON, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90** 909–920.

GREEN, P. J. (1992). Discussion of "Constrained Monte Carlo maximum likelihood for dependent data" by C. J. Geyer and E. A. Thompson. *J. Roy. Statist. Soc. Ser. B* **54** 683–684.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

IRWIN, M., COX, N. and KONG, A. (1994). Sequential imputation for multilocus linkage analysis. *Proc. Nat. Acad. Sci. U.S.A.* **91** 11684–11688.

JENSEN, C. S. and KONG, A. (1998). Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops. *American Journal of Human Genetics*. To appear.

KASS, R. E. and VOS, P. W. (1997). *Geometry Foundations of Asymptotic Inference*. Wiley, New York.

KONG, A., LIU, J. and WONG, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **89** 278–288.

LEWIS, S. M. and RAFTERY, A. E. (1997). Estimating Bayes factors via posterior simulation with Laplace–Metropolis estimator. *J. Amer. Statist. Assoc.* **92** 648–663.

MARINARI, E. and PARISI, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* **19** 451–458.

MENG, X. L. and SCHILLING, S. (1996). Fitting full-information factor models and an empirical investigation of bridge sampling. *J. Amer. Statist. Assoc.* **91** 1254–1267.

MENG, X. L. and WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica* **6** 831–860.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21** 1087–1092.

MITCHELL, A. F. S. (1992). Estimative and predictive distances. *Test* **1** 105–121.

NEAL, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Dept. Computer Science, Univ. Toronto.

NEWTON, M. A. and RAFTERY, A. E. (1994). Approximate Bayesian inference and the weighted likelihood bootstrap (with discussion). *J. Roy. Statist. Soc. Ser. B* **56** 3–48.

OGATA, Y. (1989). A Monte Carlo method for high dimensional integration. *Numer. Math.* **55** 137–157.

OGATA, Y. (1990). A Monte Carlo method for an objective Bayesian procedure. *Ann. Inst. Statist. Math.* **42** 403–433.

OGATA, Y. (1994). Evaluation of Bayesian visualization models—Two computational methods. Research Memorandum 503, Institute of Statistical Mathematics, Tokyo.

OGATA, Y. and TANEMURA, M. (1984). Likelihood analysis of spatial point patterns. *J. Roy. Statist. Soc. Ser. B* **46** 496–518.

OTT, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *American Journal of Human Genetics* **31** 161–175.

RAFTERY, A. E. (1996). Hypothesis testing and model selection via posterior simulation. In *Practical Markov Chain Monte Carlo* (W. Gilks, S. Richardson and D. J. Spiegelhalter, eds) 163–187. Chapman and Hall, London.

RAO, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37** 81–91.

RAO, C. R. (1949). On the distance between two populations. *Sankhyā* **9** 246–248.

RIPLEY, B. D. (1988). *Statistical Inference for Spatial Processes*. Cambridge Univ. Press.

STEIN, M. (1992). Prediction and inference for truncated spatial data. *J. Comput. Graph. Statist.* **1** 91–110.

THOMPSON, E. A. (1996). Likelihood and linkage: from Fisher to the future. *Ann. Statist.* **24** 449–465.

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.