A computational problem in many fields is to evaluate multiple integrals and expectations simultaneously. Consider probability distributions with unnormalized density functions indexed by parameters on a 2-dimensional grid, and assume that samples are simulated from distributions on a subgrid. Examples of such unnormalized density functions include the observed-data likelihoods in the presence of missing data and the prior times the likelihood in Bayesian inference. There are various methods using a single sample only or multiple samples jointly to compute each integral. Path sampling seems a compromise, using samples along a 1-dimensional path to compute each integral. However, different choices of the path lead to different estimators, which should ideally be identical. We propose calibrated estimators by the method of control variates to exploit such constraints for variance reduction. We also propose biquadratic interpolation to approximate integrals with parameters outside the subgrid, consistently with the calibrated estimators on the subgrid. These methods can be extended to compute differences of expectations through an auxiliary identity for path sampling. Furthermore, we develop stepwise bridge-sampling methods in parallel but complementary to path sampling. In three simulation studies, the proposed methods lead to substantially reduced mean squared errors compared with existing methods.

1. Introduction

Consider the following problem of computing multiple integrals and expectations simultaneously. Let \( \mu \) be a baseline measure on a sample space \( \mathcal{X} \) and \( q(\mathbf{x}; \theta) \) be a nonnegative function indexed by a parameter \( \theta \) in a set \( \Theta \), such that the integral
\[
Z_\theta = \int q(\mathbf{x}; \theta) \, d\mu
\]
is finite and positive. For concreteness, assume that the set \( \Theta \) is a regular grid of discrete points. Consider the probability distribution
\[
dP_\theta = \frac{q(\mathbf{x}; \theta)}{Z_\theta} \, d\mu.
\]
Then \( q(\mathbf{x}; \theta) \) is called an unnormalized density function and \( Z_\theta \) is called the normalizing constant. For each \( \theta \in \Theta \), let \( \phi(\mathbf{x}; \theta) \) be a real-valued function on \( \mathcal{X} \) such that the following expectation is finite:
\[
E_\theta(\phi_\theta) = \int \phi(\mathbf{x}; \theta) \, dP_\theta.
\]
The objective is to compute $Z_\theta$ and $E_\theta(\phi|\theta)$ for all $\theta \in \Theta$. As shall be seen from examples below, it is typically sufficient to compute $(Z_\theta : \theta \in \Theta)$ up to a positive multiple, or equivalently to compute the ratios $(Z_\theta/Z_{\theta_0} : \theta \in \Theta)$ for a fixed point $\theta_0$.

This type of problems are common to many fields including physics, chemistry, and statistics, as discussed by Gelman and Meng (1998) and Geyer (1996). We highlight two such problems in statistics. For missing data problems, write the observed data as $y_{obs}$ and the missing data as $y_{mis}$, and let $p(y_{obs}, y_{mis} | \theta)$ be their joint probability density depending on a parameter $\theta$. Usually, this joint density $p(y_{obs}, y_{mis} | \theta)$ is of analytically closed form. Then the marginal density of $y_{obs}$ (i.e., the observed-data likelihood) is of form $Z_\theta$ and the conditional density of $y_{mis}$ given $y_{obs}$ is of form $P_{\theta_{mis}}$, with $x = y_{mis}$ and $q(x, \theta) = p(y_{obs}, y_{mis} | \theta)$. Although likelihood inference is often focused on the maximum likelihood estimator and the observed Fisher information (i.e., the Hessian of the log-likelihood at the maximum), computing the observed-data likelihood function is still essential for using the likelihood ratio statistic and conducting higher-order likelihood inference in small to medium samples (e.g., Brazzale et al., 2007).

In the Bayesian paradigm, let $p(y | \beta, \theta)$ be the probability density of the data $y$ given parameters $\beta$ and hyper-parameters $\theta$ and $p(\beta | \theta)$ be the prior density of $\beta$. Then the marginal likelihood of $y$ is of form $Z_\theta$ and the posterior density of $\beta$ given $y$ is of form $P_\beta$, with $x = \beta$ and $q(\beta, \theta) = p(y | \beta, \theta)p(\beta | \theta)$. For two distinct values $\theta_0$ and $\theta_1$, the ratio $Z_\theta / Z_{\theta_0}$ is called the Bayes factor and often used to choose between models $p(y | \beta, \theta_0)$ and $p(y | \beta, \theta_1)$ in Bayesian model selection.

We investigate Monte Carlo methods for solving the aforementioned problems. A simple approach is importance sampling, i.e., simulate draws $x_1, \ldots, x_n$ from $P_{\theta_0}$, and then estimate $Z_{\theta_0}/Z_{\theta_0}$ by $n^{-1} \sum_{i=1}^n (q(x_i; \theta_1)/q(x_i; \theta_0))$ and $E_\theta(\phi|\theta)$ by $\sum_{i=1}^n (\phi(x_i; \theta)/q(x_i; \theta_0))/\sum_{i=1}^n (q(x_i; \theta)/q(x_i; \theta_0))$. These estimators may perform poorly unless $P_\theta$ is close to $P_{\theta_0}$. Let $\Theta_j$ be a subgrid of $\Theta$ and assume that $\theta_0 \in \Theta_j$. To handle multiple integrals, a reasonable approach consists of first simulating samples from multiple distributions $P_{\theta_0} (\theta \in \Theta_j)$ and then constructing estimators of $Z_\theta$ and $E_\theta(\phi|\theta)$ ($\theta \in \Theta$) using the multiple samples. For the simulation stage, various sampling algorithms can be used such as Markov chain Monte Carlo or sequential Monte Carlo. See Liu (2001) and Robert and Casella (2005) for textbook accounts. For the estimation stage, there are also various estimators available, roughly summarized as follows.

(i) For $\theta \in \Theta_s$, apply an estimator of $Z_\theta$ based on solely the sample from $P_{\theta_0}$. Examples include the estimator of Chib (1995) if the sample is obtained by Gibbs sampling or an estimator derived from importance weights (Liu, 2001, Section 3.4) if the sample is obtained by particle filtering or sequential Monte Carlo. For $\theta \in \Theta_s$, approximate $Z_\theta$ by interpolation and extrapolation.

(ii) For $\theta \in \Theta_s$, apply a path-sampling estimator of Gelman and Meng (1998) for $Z_\theta/Z_{\theta_0}$. This estimator is related to thermodynamic integration in statistical physics. For $\theta \in \Theta_s$, approximate $Z_\theta$ by interpolation and extrapolation.

(iii) For $\theta \in \Theta$, apply likelihood estimators of Kong et al. (2003) for $Z_\theta/Z_{\theta_0}$ and $E_\theta(\phi|\theta)$ based on jointly the multiple samples. These estimators have been previously derived by Geyer (1994) and Meng and Wong (1996) and called reverse logistic regression and extended bridge sampling, respectively.

(iv) For $\theta \in \Theta_s$, estimate $E_\theta(\phi|\theta)$ by the sample average of $\phi(x; \theta)$ in the sample from $P_{\theta_0}$. For $\theta \in \Theta_s$, approximate $E_\theta(\phi|\theta)$ by interpolation and extrapolation.

We emphasize that for each estimator, the quality of results depends on that of the sampling algorithm used. Considerable efforts are often required to design good sampling algorithms in specific problems. For example, Bornn et al. (2010) developed interesting sequential Monte Carlo methods for sampling from a sequence of posterior distributions for sensitivity analysis and model selection. They did not explicitly discuss estimation of marginal likelihoods $Z_\theta$, but it seems straightforward to use method (i) with the estimator of $Z_\theta$ based on importance weights. We also realize that the foregoing discussion treats sampling and estimation separately. There are other methods in which sampling and estimation are combined in a single process (e.g., Everitt, in press). For space limitation, we do not pursue a systematic comparison with those methods.

A basic difference between methods (i), (iii), and (iv) is that the multiple samples are used separately in methods (i) and (iv) but jointly in method (iii). By Tan (2004), if the draws from $P_{\theta_0}$ are independent for each $\theta \in \Theta_s$, then the likelihood estimators have no greater asymptotic variances than a large class of estimators, which includes many commonly used estimators. In general, if the draws are not independent, the likelihood estimators may still gain efficiency over those in methods (i) and (iv). On the other hand, method (iii) is computationally more costly than methods (i) and (iv). To compute the likelihood estimators for $Z_{\theta_0}/Z_{\theta_0}$ ($\theta \in \Theta_s$), method (iii) requires a total of $m^2n$ evaluations of $q(x; \theta)$ instead of typically $mn$ evaluations, where $m$ is the size of $\Theta_s$ and $n$ is the common sample size for individual samples. There is a tradeoff between statistical efficiency and computational cost. The likelihood method may not be computationally as appealing as methods (i) and (iv) when $m$ is large. To address this issue, methods have been proposed by reducing computational cost while maintaining statistical efficiency for the likelihood method (Doss, 2010; Tan, in press).

Path sampling (ii) seems a compromise between methods (i) and (iii). The path-sampling method requires a total of $mn$ evaluations of $(\partial \log q(x; \theta)/\partial \theta_0) \log q(x; \theta)$ to compute all $Z_{\theta_0}/Z_{\theta_0}$ ($\theta \in \Theta_s$), but involves multiple samples including those from $P_{\theta_0}$ and $P_{\theta_1}$ to estimate each $Z_{\theta_0}/Z_{\theta_0}$. In fact, let $(\theta_0, \theta_1, \ldots, \theta_{m-1}, \theta_0)$ be a path in $\Theta_s$ connecting $\theta_0$ and $\theta_1 \equiv \theta$. Then a single-path estimator of $Z_{\theta_0}/Z_{\theta_0}$ can be constructed on the samples from $(P_{\theta_0}, P_{\theta_1}, \ldots, P_{\theta_{m-1}})$. Different choices of the path lead to different estimators. For a 2-dimensional, rectangular grid, Gelman and Meng (1998, Section 2.3) proposed a simple method of averaging single-path estimators. In a numerical example, they found that two different ways of applying the method yield...
results of substantially different accuracy. It remains an interesting open problem whether, in general, the better way of applying the method can be automatically identified.

We develop new methods for path sampling on a 2-dimensional, rectangular grid. In Section 2.1, we derive a calibrated estimator of $\log(Z_\theta/Z_{\theta_0})$ for $\theta \in \Theta$, by the method of control variates (e.g., Hammersley and Handscomb, 1964) to exploit the constraints that different single-path estimators of $\log(Z_\theta/Z_{\theta_0})$ should ideally be identical. If the draws from $P_\theta$ are independent, then the calibrated estimator achieves the lowest asymptotic variance among linear combinations of single-path estimators with the coefficients summing to 1, and hence is asymptotically guaranteed to gain efficiency over either estimator obtained by Gelman and Meng’s averaging method. In general, if the draws are not independent, the calibrated estimator may still yield variance reduction over existing estimators, as shown by examples in Section 4.

In Section 2.2, we propose biquadratic interpolation to approximate $\log(Z_\theta/Z_{\theta_0})$ for $\theta \in \Theta \setminus \Theta_0$, where $\Theta_0$ is coarser than $\Theta$. For each rectangular cell in $\Theta_0$, the interpolant is biquadratic in $\theta$ without the cross-quadratic term, interpolating the function values and the first-order derivatives at the four corners. The function value at $\theta \in \Theta_0$ is set to the calibrated estimator of $\log(Z_\theta/Z_{\theta_0})$, and the first-order derivatives are set to a calibration of the sample-average estimators of those derivatives.

In Section 2.3, we provide a path-sampling identity for the difference of expectations $E_\theta(\phi_\theta) - E_{\theta_0}(\phi_{\theta_0})$, which is similar to that for the log difference of normalizing constants $\log(Z_\theta/Z_{\theta_0})$. With this result, path-sampling methods from Sections 2.1–2.2 can be directly extended to compute $E_\theta(\phi_\theta) - E_{\theta_0}(\phi_{\theta_0})$.

There are two subtle limitations for path-sampling methods considered here. Each single-path estimator on a fixed grid is in general inconsistent and so is the calibrated path sampling estimator. The biases can be potentially large unless the grid is sufficiently dense or $\log Z_\theta$ is approximately quadratic in $\theta$ on each edge. Moreover, path sampling does not seem to suggest a better estimator for $E_\theta(\phi_\theta)$ itself than the sample-average estimator. To address these issues, we develop stepwise bridge-sampling methods for computing $\log(Z_\theta/Z_{\theta_0})$ and $E_\theta(\phi_\theta)$ in Section 3.

For a path $(\theta_0, \theta_1, \ldots, \theta_{t-1}, \theta_t)$ in $\Theta$, with $\theta_1 \equiv \theta$, stepwise bridge-sampling is to employ a bridge-sampling estimator of $Z_{\theta_0}/Z_{\theta_{t-1}}$ based on only the samples from $P_{\theta_0}$ and $P_{\theta_t}$ and then substitute those estimators into $\log(Z_\theta/Z_{\theta_0})$. The resulting estimator is related to path sampling (Gelman and Meng, 1998, Section 3.3), and consistent as long as $P_{\theta_0}$ and $P_{\theta_t}$ have non-negligible overlap. Furthermore, for each edge $(\theta, \theta')$ in $\Theta$, a likelihood estimator of $E_\theta(\phi_\theta)$ can be obtained while applying bridge sampling on the samples from $P_{\theta}$ and $P_{\theta'}$. To combine estimators of $\log(Z_\theta/Z_{\theta_0})$ from different paths and those of $E_\theta(\phi_\theta)$ from different edges, we derive calibrated estimators by the method of control variates similarly as in Section 2.

We present simulation studies in Section 4, and provide concluding remarks in Section 5. All proofs are collected in the Appendix.

2. Path sampling

Gelman and Meng (1998) provided a formulation of path sampling, generalizing thermodynamic integration in statistical physics and Ogata’s (1989) method. As introduced in Section 1, $q(x; \theta)$ is an unnormalized density function and $Z_\theta$ is the normalizing constant. The fundamental identity for path sampling is

$$\frac{\partial}{\partial \theta} \log Z_\theta = E_\theta(U(x; \theta)),$$  \hspace{1cm} (1)

where $U(x; \theta) = (\partial/\partial \theta) \log q(x; \theta)$, and the expectation $E_\theta$ is taken with respect to $P_\theta$. An immediate estimator of $(\partial/\partial \theta) \log Z_\theta$ is the sample average of $U(x; \theta)$ in the sample from $P_\theta$, thereafter denoted by $E_\theta(U(x; \theta))$ or $E_\theta(U_{\theta})$.

To explain main ideas, suppose that $\theta$ is univariate and let $\theta_0 < \theta_1 < \cdots < \theta_t$ be a sequence of grid points in $\Theta$, with $\theta_1 \equiv \theta$. Then Eq. (1) gives

$$\log \frac{Z_{\theta_t}}{Z_{\theta_0}} = \int_{\theta_{t-1}}^{\theta_t} E_\theta(U(x; t)) \, dt.$$  \hspace{1cm} (2)

The path-sampling method is to estimate $E_{\theta_0}(U(x; \theta_0))$ by $\tilde{E}_{\theta_0}(U(x; \theta_0))$ and approximate $\log(Z_{\theta_0}/Z_{\theta_{t-1}})$ by, for example, the trapezoidal rule:

$$\frac{1}{2}(\theta_t-\theta_{t-1})[\tilde{E}_{\theta_{t-1}}(U(x; \theta_{t-1})) + \tilde{E}_{\theta_0}(U(x; \theta_0))].$$  \hspace{1cm} (3)

Consequently, $\log(Z_{\theta_0}/Z_{\theta_{t-1}})$ is approximated by the composite trapezoidal rule.

For simplicity, assume that the bias of estimator (3) for $\log(Z_{\theta_0}/Z_{\theta_{t-1}})$ is negligible when compared with the standard deviation. By the error analysis for the trapezoidal rule, the approximation bias of (3) is

$$-\left(\frac{d^3}{d\theta^3} \log Z_\theta\right)_{\theta = \theta'} (\theta_t-\theta_{t-1})^3/12,$$

where $\theta'$ is a fixed point in $[\theta_{t-1}, \theta_t]$. Therefore, this assumption requires that $(\theta_t-\theta_{t-1})$ is sufficiently small or $\log Z_\theta$ is approximately quadratic in $\theta$. Interestingly, the latter condition is likely satisfied if $\log Z_\theta$ represents a log-likelihood near its maximum as in the two types of problems outlined in Section 1.
Now assume that \( \theta \) is bivariate and \( \Theta_s \) consists of \( m_1 \times m_2 \) grid points \( \theta^i = (\theta_{1i}, \theta_{2j}) \), \( i = 1, \ldots, m_1 \), \( j = 1, \ldots, m_2 \). Path sampling can be extended by performing integration along a 1-dimensional path (Gelman and Meng, 1998, Section 2.3). Figure 1 illustrates the constraints that different single-path estimators along the four edges of a rectangular cell as \( 0 \times 0 \) should ideally be 0. Arrange all \( \xi_{ij}^{\theta} \) into a vector \( \xi \). Moreover, arrange all \( \xi_{ij}^{\theta} \) into a vector \( \xi \) and let \( C \) be a matrix such that \( \xi = C \xi \). Each row of \( C \) consists of 0s except for two pairs of \((1, -1)\).

The method of control variates is a useful technique for variance reduction in Monte Carlo computation. See Hammersley and Handscomb (1964, Section 5.7) for a textbook discussion. By construction, each element of \( \xi \) is approximately 0 and hence a control variate. Consider the regression estimator

\[
\xi_{\text{reg}} = \xi - (VC^T)(VC^T)^{-1}C\xi.
\]

where \( V \) is a “working” estimator of the asymptotic variance matrix \( \Psi \) of \( \xi \), and hence \( CVC^T \) is an estimator of that of \( \xi = C\xi \). It is not required that \( V \) be a consistent estimator of \( \Psi \). By definition, \( \xi_{\text{reg}} \) automatically satisfies \( C\xi_{\text{reg}} = 0 \).
The vector $\xi_{\text{reg}}$ can be interpreted as a calibration of $\xi$ such that every single-path estimator based on $\xi_{\text{reg}}$ along a closed path in $\Theta_\delta$ is exactly 0.

For our implementation, $V$ is specified using the sample variances and covariances of $U(x; \theta)$. The element of $V$ for the asymptotic variance of $\xi^V$ is

$$
\frac{1}{4} \left( \text{var}_\theta \left[ U_1(x; \theta^i) \right] + \text{var}_\theta \left[ U_1(x; \theta^{i+1}) \right] \right),
$$

and that of $V$ for the asymptotic covariance of $\xi_i^V$ and $\xi_j^V$ is

$$
\frac{1}{4} \left( \text{var}_\theta \left[ U_1(x; \theta^i) \right] \text{cov}_\theta \left[ U_1(x; \theta^i), U_2(x; \theta^j) \right] \right),
$$

where $\text{var}_\theta$ and $\text{cov}_\theta$ denote the sample variance and covariance in the sample from $P_\theta$, and $n^V$ is the number of draws from $P_\theta$. Other elements of $V$ are specified similarly. Both $V$ and $V^C$ are sparse, because samples from different distributions are independent. In fact, $\xi^V_i$ and $\xi^V_j$ are independent if $|i-j| \geq 2$. By the sparsity of $V^C$, the estimator $\xi_{\text{reg}}$ appears numerically stable in our examples, even though $V^C$ is a relatively large matrix and need to be inverted.

If the draws from each $P_\theta$ are independent, then $V$ is a consistent estimator of $V^*$ such that $V^* = \text{var}_\theta$. In this case, the asymptotic variance matrix of $\xi_{\text{reg}}$ is $V_{\text{reg}}^* = V^* - (V^C)^{-1}(V^C)^{-1}V^*$, smaller than $V^*$ in the order of positive definite matrices. If the draws are not independent, for example, generated by Markov chain or sequential Monte Carlo, then $V$ is not a consistent estimator of $V^*$. A direction for future work is to incorporate specialized methods to estimate $V^*$ properly (e.g., Flegal and Jones, 2010). Nevertheless, the simple choice of $V$ may be useful in many situations for the following reason. The asymptotic variance matrix of $\xi_{\text{reg}}$ is approximately $V^* = \text{var}_\theta$ as long as $(V^C)^{-1}$ is close to $(V^C)^{-1}(V^C)^{-1}V^*$, but not necessarily $V$ is close to $V^*$. This condition is satisfied if the effect of dependency within the samples from different distributions is to inflate the asymptotic variance matrix of $\xi$ by approximately a scalar factor $\alpha > 0$, i.e., $V$ is close to $\alpha V^*$.

For each $\theta \in \Theta$, a single-path estimator of $\log(Z_\theta/Z_{\theta_0})$ is a linear combination with coefficients 1 or $-1$ of $\xi^V_i$ and $\xi^V_j$ along the path. Let $D$ be a matrix with coefficients 1, -1, or 0 such that $D\xi_{\text{reg}}$ gives a vector of single-path estimators of $\log(Z_\theta/Z_{\theta_0})$ for all $\theta \in \Theta$. Different choices of $D$ lead to different $D\xi$. However, if $\xi$ is replaced by $\xi_{\text{reg}}$, the resulting vector of estimators

$$
D\xi_{\text{reg}} = D\xi - (DVC)^{-1}C\xi
$$

is invariant to different choices of $D$ because $C\xi_{\text{reg}} = 0$. We refer to $D\xi_{\text{reg}}$ as a vector of calibrated estimators of $\log(Z_\theta/Z_{\theta_0})$. By standard theory for the method of control variates (e.g., Hammersley and Handscomb, 1964), if $V$ is a consistent estimator of $V^*$, then $D\xi_{\text{reg}}$ achieves the smallest asymptotic variance matrix in the class of estimators $D_1\xi - B_1C\xi$, where $D_1$ is an arbitrary choice of $D$ and $B_1$ is a matrix of arbitrary constants with suitable dimensions. This class of estimators includes linear combinations $\sum_{j=1}^k \gamma_j D_j \xi$, where $D_1, \ldots, D_k$ are different choices of $D$ and $\gamma_1, \ldots, \gamma_k$ are constant coefficients with $\sum_{j=1}^k \gamma_j = 1$. If $V$ is not a consistent estimator of $V^*$, $D\xi_{\text{reg}}$ may still provide considerable variance reduction as shown in Section 4.

It is important to point out that even further simplified choices of $V$ may be used. For example, the asymptotic covariance of $\xi^V_i$ and $\xi^V_j$ may be set to 0 for all $i$ and $j$. In our simulation studies, this specification leads to similar results to those with the full specification of $V$. Furthermore, the averaging method of Gelman and Meng (1998) seems to coincide with a deterministic choice of $V$. For averaging over $\theta_1$-first paths, $\text{var}_\theta \left[ U_1(x; \theta^i) \right]$ is replaced by 0 and $\text{var}_\theta \left[ U_2(x; \theta^i) \right]$ by 0 for all $i$ and $j$. For averaging over $\theta_2$-first paths, $\text{var}_\theta \left[ U_1(x; \theta^i) \right]$ is replaced by 0 and $\text{var}_\theta \left[ U_2(x; \theta^i) \right]$ by 1 for all $i$ and $j$. In both cases, $\text{cov}_\theta \left[ U_1(x; \theta^i), U_2(x; \theta^j) \right]$ is replaced by 0 for all $i$ and $j$. The first (or second) specification may yield similar results to those with the full specification of $V$ when $E_{\theta_1} \left[ U_1(x; \theta^i) \right]$ has substantially greater (or smaller) asymptotic variance than $E_{\theta_2} \left[ U_2(x; \theta^i) \right]$, as observed in Section 4.1 for the example of Gelman and Meng (1998, Section 5.2). Of course, these deterministic specifications in general lead to less satisfactory results than those with the full specification of $V$, as demonstrated in the example of Section 4.2.

### 2.2. Biquadratic interpolation

So far we study estimating $\log(Z_\theta/Z_{\theta_0})$ for $\theta \in \Theta$, based on samples from $P_\theta$ ($\theta \in \Theta$). However, if $\Theta_\delta$ is coarser than $\Theta$, it remains necessary to approximate $\log(Z_\theta/Z_{\theta_0})$ for $\theta \in \Theta \setminus \Theta_\delta$. We first draw a connection between the trapezoidal rule and quadratic interpolation for univariate functions, two seemingly unrelated topics in numerical analysis. Then we propose biquadratic interpolation to approximate $\log(Z_\theta/Z_{\theta_0})$ consistently with the calibrated estimators of $\log(Z_\theta/Z_{\theta_0})$ ($\theta \in \Theta$).
We recast the trapezoidal rule, typically derived by linear interpolation of the integrand, in the form of quadratic interpolation of the integral itself. In the context of path sampling with univariate \( \theta \), write \( g(\theta) = (d/d\theta) \log Z_\theta = E_{\theta}(U(x; \theta)) \). Set \( g_{j-1} = E_{\theta_{j-1}}(U(x; \theta_{j-1})) \) and \( g_j = E_{\theta_j}(U(x; \theta_j)) \) and then approximate \( g(\theta) \) by the linear interpolant \( \bar{g}(\theta) = (\theta_j - \theta)g_{j-1} + (\theta - \theta_{j-1})g_j/(\theta_j - \theta_{j-1}) \) over the interval \([\theta_{j-1}, \theta_j]\). The resulting approximation of \( \log(Z_\theta/Z_{\theta_{j-1}}) = \int_{\theta_{j-1}}^{\theta_j} g(t) \, dt \) is
\[
\int_{\theta_{j-1}}^{\theta_j} \bar{g}(t) \, dt = \frac{(\theta_j - \theta_{j-1})}{2} \left( \frac{\theta_j - \theta+ \theta - \theta_{j-1}}{2} \right) g_{j-1} + \frac{1}{2} \frac{\theta_j - \theta_{j-1}}{\theta_j - \theta_{j-1}} g_j.
\]
which yields the trapezoidal rule (3) for \( \theta = \theta_j \). However, it is useful to derive formula (6) by directly approximating \( \int_{\theta_{j-1}}^{\theta_j} g(t) \, dt \) instead of approximating \( g(\theta) \) and then integrating the approximation to \( g(\theta) \). By simple algebra, (6) can be obtained as the unique solution to the following problem: find a quadratic function \( G(\theta) \) on \([\theta_{j-1}, \theta_j]\) such that \( G(\theta_{j-1}) = 0 \) and \( dG/d\theta \) equals \( g_{j-1} \) at \( \theta_{j-1} \) and \( g_j \) at \( \theta_j \).

Now consider the case where \( \theta \) is bivariate and the task of approximating \( \log(Z_\theta/Z_{\theta'}) \) for \( \theta \) over a rectangular cell \([\theta_{i1}, \theta_{i1+j+1}] \times [\theta_{j1}, \theta_{j2+j+1}]\). Formula (6) can be applied along the four edges, for example, along \( \theta^i \to \theta^i+j \to \theta^i+1+j \) and along \( \theta^j \to \theta^j+i \to \theta^j+1+i \) separately. In general, the two paths lead to different values at \( \theta^i+j+1 \) as discussed in Section 2.1. Moreover, no approximation is provided for \( \theta \) in the interior of the rectangle. Alternatively, bilinear interpolation can be adopted to approximate \( \log(Z_\theta/Z_{\theta'}) \) over the rectangle, using single-path or calibrated estimators at the four corners. This approximation may be too simple to be accurate. However, motivated by the foregoing discussion about quadratic interpolation, we attempt to approximate \( \log(Z_\theta/Z_{\theta'}) \) by solving the following problem:

Find a biquadratic function \( G(\theta) \) such that \( G(\theta^i) = 0 \) and \( \partial G/\partial \theta_1 \) equals \( g_1^{ij} \) and \( \partial G/\partial \theta_2 \) equals \( g_2^{ij} \) at \( \theta^i+j \)
for \((i, j) = (i, j), (i+1, j), (i, j+1), \) and \((i+1, j+1)\).

(7)

Here \( g_1^{ij} \) and \( g_2^{ij} \) can be flexibly specified. For example, \((g_1^{ij}, g_2^{ij})\) can be set to \( E_{\theta^i\theta^j}/(\mathbf{Ux}; \theta^i\theta^j)\). A biquadratic function \( G(\theta) \) over \([\theta_{i1}, \theta_{i1+j+1}] \times [\theta_{j1}, \theta_{j2+j+1}]\) is of form \( \sum_{i,j} c_{ij} \theta_i \theta_j \), where \( c_{ij} \) are real coefficients. There are nine coefficients to be determined and nine constraints imposed in problem (7). It may be easily expected that a unique solution exists for problem (7). However, a careful analysis gives the following answer.

**Proposition 1.** Let
\[
\begin{align*}
\gamma_{i1} &= \frac{1}{2} (g_1^{ij} + g_1^{i+1,j+1})(\theta_{i+1,j+1} - \theta_{i1}), \\
\gamma_{i2} &= \frac{1}{2} (g_2^{ij} + g_2^{i+1,j+1})(\theta_{i+1,j+1} - \theta_{i2}), \\
\gamma_{j1} &= \frac{1}{2} (g_1^{i+1,j} + g_1^{i+1,j+1})(\theta_{i+1,j+1} - \theta_{i1}), \\
\gamma_{j2} &= \frac{1}{2} (g_2^{i+1,j} + g_2^{i+1,j+1})(\theta_{i+1,j+1} - \theta_{i2}).
\end{align*}
\]

(i) There exists no solution to problem (7) unless
\[
\gamma_{i1} + \gamma_{i2} + \gamma_{j1} + \gamma_{j2} = 0.
\]

(ii) There exists a unique solution to problem (7) if \( c_{i2} = 0 \) and condition (8) holds. See the Appendix for an expression of the solution.

To show Proposition 1(i), we notice that if \( G(\theta) \) is a solution to problem (7), then \( G(\theta) \) is quadratic along the four edges and hence the trapezoidal rule is exact. Therefore, \( \gamma_{i1} = G(\theta^i+1) \), \( \gamma_{i2} = G(\theta^i+1)-G(\theta^i+j) \), \( \gamma_{j1} = G(\theta^i+j+1)-G(\theta^i+1) \), \( \gamma_{j2} = G(\theta^i+j+1) \), \( \gamma_{i1} + \gamma_{i2} = G(\theta^i+i+1) \), \( \gamma_{i1} + \gamma_{i2} = G(\theta^i+1) \), \( \gamma_{j1} + \gamma_{j2} = G(\theta^i+i+1) \), \( \gamma_{j1} + \gamma_{j2} = G(\theta^i+1) \). On the other hand, if condition (8) indeed holds, then there remain eight free constraints and hence one constraint must be introduced on the nine coefficients \( c_{i2} \) to obtain a unique solution to (7). It seems reasonable to set \( c_{i2} \) to 0 as in Proposition 2(ii). Problem (7) with \( c_{i2} = 0 \) and condition (8) satisfied is equivalent to the following problem:

Find a biquadratic function \( G(\theta) \) with \( c_{i2} = 0 \) such that \( G(\theta^{i+j}) = G^{ij} \)
for \((i', j') = (i, j), (i+1, j), (i, j+1), \) and \((i+1, j+1)\)
and \( \partial G/\partial \theta_1 \) equals \( g_1^{ij} \) at \( \theta^i \) and \( g_1^{ij+1} \) at \( \theta^{i+j} \)
and \( \partial G/\partial \theta_2 \) equals \( g_2^{ij} \) at \( \theta^j \) and \( g_2^{ij+1} \) at \( \theta^{i+j} \).

(9)

Here \( G^{ij}, g_1^{ij}, g_1^{ij+1}, g_2^{ij}, g_2^{ij+1} \) can be flexibly specified. The equivalence holds because if four derivatives are fixed for a biquadratic function as in problem (9), then the remaining four derivatives, by the trapezoidal rule, have a one-to-one correspondence with the four function values at the corners minus \( G(\theta^i) \). The form of problem (9) is convenient for interpolation because the function values are fixed at the corners.
We propose a two-step method to approximate \( \log(Z_{\theta}/Z_{\theta_0}) \) for \( \theta \in \Theta \), combining calibration and interpolation. First, for \( i = 1, \ldots, m_1 \) and \( j = 1, \ldots, m_2 \), set \( (g_{ij}^0, g_{ij}^1) = E_{\theta_0}^{(i)}(\mathbf{U}(\mathbf{x}; \theta_0)) \). Then \( (g_{ij}^0, g_{ij}^1) \) agree with those in Section 2.1. In general, condition \( (8) \) or equivalently \( \xi^0 = 0 \) does not hold. Arrange all \( g_{ij}^0 \) and \( g_{ij}^1 \) into a vector \( \mathbf{g} \). Then \( \xi \) is linearly related to \( \mathbf{g} \): \( \xi = \mathbf{A}\mathbf{g} \) for a constant matrix \( \mathbf{A} \). Consider the regression estimator similar to \( \xi_{\text{reg}} \).

\[
\mathbf{g}_{\text{reg}} = \mathbf{g} - (\mathbf{U}^\top \mathbf{C}^\top) (\mathbf{V} \mathbf{C}^\top)^{-1} \mathbf{C} \mathbf{g},
\]

where \( \mathbf{U} \) is a working estimator of the asymptotic variance matrix of \( \mathbf{g} \) and hence \( \mathbf{U} \mathbf{C}^\top \) is that of the asymptotic covariance matrix of \( \mathbf{g} \) and \( \mathbf{C} \). For our implementation, \( \mathbf{U} \) is specified using sample variances and covariances, similarly as \( \mathbf{V} \) is specified by \( (4) \) and \( (5) \), etc. Then \( \mathbf{V} = \mathbf{A}^\top \mathbf{A} \), and \( \xi_{\text{reg}} \) is linearly related to \( \xi_{\text{reg}} \) in the same way as \( \xi \) to \( \mathbf{g} \), i.e., \( \xi_{\text{reg}} = \mathbf{A}\xi \). The vector \( \mathbf{g}_{\text{reg}} \) can be regarded as a calibration of \( \mathbf{g} \), similarly as \( \xi_{\text{reg}} \) is that of \( \xi \), such that condition \( (8) \) holds.

Second, approximate \( \log(Z_{\theta}/Z_{\theta_0}) \) by the solution to problem \( (9) \) over each rectangular cell in \( \Theta_s \), where \( g_{ij}^2 \) is set to the calibrated estimator of \( \log(Z_{\theta}/Z_{\theta_0}) \) and the \( g_{ij}^0, g_{ij}^1 \) are set to the corresponding elements in \( \mathbf{g}_{\text{reg}} \). For every two cells sharing an edge, the corresponding solutions to problem \( (9) \) are identical on the common edge by the trapezoidal rule. Therefore, the resulting approximation of \( \log(Z_{\theta}/Z_{\theta_0}) \) is continuous over the entire \( \Theta \). The first-order derivatives are continuous inside each rectangular cell and at the corners, but may not be so elsewhere on the edges.

2.3. Path sampling for expectations

The framework of path sampling as formulated by Gelman and Meng (1998) is devoted to estimating log normalizing constants \( \log Z_{\theta} \). It is interesting to extend the framework for estimating expectations \( E_{\theta}(\phi_{\theta}) \). While little work seems to be done in this direction in statistics, there are techniques of this nature, with \( \phi_{\theta} \) free of \( \theta \), developed to compute entropy differences in physics and chemistry (e.g., Peter et al., 2004). We establish a general identity for \( \partial(\partial/\partial\theta)E_{\theta}(\phi_{\theta}) \), similar to \( (1) \), and show that methods from Sections 2.1–2.2 can be directly extended to estimate \( E_{\theta}(\phi_{\theta}) - E_{\theta_0}(\phi_{\theta_0}) \).

**Proposition 2.** Under standard regularity conditions,

\[
\frac{\partial}{\partial\theta} E_{\theta}(\phi_{\theta}) = E_{\theta} \left\{ \frac{\partial}{\partial\theta} \phi_{\theta}(\mathbf{x}; \theta) \right\} + \text{cov}_{\theta}(\phi_{\theta}(\mathbf{x}; \theta), \mathbf{U}(\mathbf{x}; \theta)),
\]

(10)

where \( \text{cov}_{\theta} \) denotes the covariance with respect to \( \mathbf{P}_{\theta} \).

We refer to \( (10) \) as an auxiliary identity for path sampling. In the special case where \( \phi_{\theta}(\mathbf{x}; \theta) = \phi_{\theta}(\mathbf{x}) \), free of \( \theta \), then \( (10) \) gives

\[
\frac{\partial}{\partial\theta} E_{\theta}(\phi_{\theta}) = \text{cov}_{\theta}(\phi_{\theta}(\mathbf{x}), \mathbf{U}(\mathbf{x}; \theta)),
\]

which agrees with Eq. (17) in Peter et al. (2004). For another interesting application, taking \( \phi_{\theta}(\mathbf{x}; \theta) = \mathbf{U}(\mathbf{x}; \theta) \) in \( (10) \) gives

\[
\frac{\partial^2}{\partial\theta^2} \log Z_{\theta} = E_{\theta} \left\{ \frac{\partial^2}{\partial\theta^2} \log \mathbf{y}_{\text{com}}(\mathbf{\theta}) \right\} + \text{var}_{\theta}(\mathbf{U}(\mathbf{x}; \theta)),
\]

(11)

where \( \text{var}_{\theta} \) denotes the variance with respect to \( \mathbf{P}_{\theta} \). Eq. (11) is equivalent to the Bartlett identity:

\[
-E_{\theta}[(\partial^2/\partial\theta^2)\log \mathbf{y}_{\text{com}}(\mathbf{\theta})/Z_{\theta}] = \text{var}_{\theta}(\mathbf{U}(\mathbf{x}; \theta))
\]

in likelihood theory. In the context of missing-data problems as described in Section 1, Eq. (11) yields the missing information principle (e.g., Louis, 1982)

\[
\frac{\partial^2}{\partial\theta^2} \log p_{\mathbf{y}_{\text{obs}} | \theta} = E_{\theta} \left\{ \frac{\partial^2}{\partial\theta^2} \log p_{\mathbf{y}_{\text{com}} | \theta} \right\} + \text{var}_{\theta} \left( \frac{\partial}{\partial\theta} \log p_{\mathbf{y}_{\text{com}} | \theta} \right),
\]

where \( \mathbf{y}_{\text{com}} = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mix}}) \) and \( \mathbf{E}_{\theta} \) and \( \text{cov}_{\theta} \) are taken with respect to \( p_{\mathbf{y}_{\text{mix}} | \mathbf{y}_{\text{obs}}, \theta} \). In the context of spatial models, Eq. (11) together with \( (1) \) is used by Gu and Zhu (2001, Equation 5) to compute maximum likelihood estimators.

The derivatives of \( E_{\theta}(\phi_{\theta}) \) are expressed by \( (10) \) through expectations with respect to \( \mathbf{P}_{\theta} \), similarly as those of \( \log Z_{\theta} \) by \( (1) \). For \( \theta \in \Theta_s \), let

\[
\mathbf{U}^{(\theta)}(\mathbf{x}; \theta) = \frac{\partial}{\partial\theta} \phi_{\theta}(\mathbf{x}; \theta) + (\phi_{\theta}(\mathbf{x}; \theta) - E_{\theta}(\phi_{\theta})) | (\mathbf{U}(\mathbf{x}; \theta) - E_{\theta}(\mathbf{U}))
\]

If \( \mathbf{U}(\mathbf{x}; \theta) \) is replaced by \( \mathbf{U}^{(\theta)}(\mathbf{x}; \theta) \) throughout in \( \xi, \xi_{\text{reg}}, \mathbf{g}, \mathbf{g}_{\text{reg}} \) and related formulas, then methods in Sections 2.1–2.2 can be adopted to estimate \( E_{\theta}(\phi_{\theta}) - E_{\theta_0}(\phi_{\theta_0}) \) (\( \theta \in \Theta_s \)) and approximate \( E_{\theta}(\phi_{\theta}) - E_{\theta_0}(\phi_{\theta_0}) \) (\( \theta \in \Theta \)). Even though \( \mathbf{U}^{(\theta)}(\mathbf{x}; \theta) \) involves \( E_{\theta}(\phi_{\theta}) \) and \( E_{\theta}(\mathbf{U}) \) instead of the true values \( E_{\theta_0}(\phi_{\theta}) \) and \( E_{\theta_0}(\mathbf{U}) \), this difference does not introduce additional variation in \( E_{\theta}(\mathbf{U}) \) up to the first order, because

\[
\text{cov}_{\theta}(\phi_{\theta}(\mathbf{x}; \theta), \mathbf{U}(\mathbf{x}; \theta)) = E_{\theta}[(\phi_{\theta}(\mathbf{x}; \theta) - E_{\theta}(\phi_{\theta})) | (\mathbf{U}(\mathbf{x}; \theta) - E_{\theta}(\mathbf{U}))]
\]

and the second term on the right side is negligible up to the first order. Therefore, \( \mathbf{V} \) based on \( (4)–(5) \) remains a consistent estimator of the asymptotic variance matrix \( \mathbf{V}^* \) if the draws from each \( \mathbf{P}_{\theta} \) are independent.
Stepwise bridge sampling

Throughout Section 2, the development assumes that the bias of estimator (3) is negligible relatively to the standard deviation. Moreover, the extension of path sampling in Section 2.3 only deals with the difference \( E_{\theta}(\phi_0) - E_{\theta}(\phi_0) \), not \( E_0(\phi_0) \) itself. To address these issues, we propose alternative methods by a stepwise application of two-sample bridge sampling (Kong et al., 2003; Meng and Wong, 1996).

3.1. Two-sample bridge sampling

The method of bridge sampling by Meng and Wong (1996) corresponds to a special case of the likelihood method of Kong et al. (2003) for two samples. We present relevant formulas in the context of estimator (3), where \( \theta \) is univariate and \( \theta_{j-1} \) and \( \theta_j \) are grid points in \( \Theta_0 \). However, the formulas are applicable in general.

Suppose that a sample of size \( n_{j-1} \) is simulated from \( P_{\theta_{j-1}} \), and a sample of size \( n_j \) from \( P_{\theta_j} \). Denote by \( \{x_1, \ldots, x_N\} \) the pooled sample of size \( N = n_{j-1} + n_j \). The likelihood estimator of \( \log(Z_0/Z_{0,1}) \) is a minimizer, \( \eta \), to the function

\[
n_j \eta + \sum_{i=1}^{N} \log(n_{j-1} q(x_i; \theta_{j-1}) + n_j e^{-\eta} q(x_i; \theta_j)),
\]

or equivalently a solution to the equation

\[
e^\eta = \sum_{i=1}^{N} \frac{q(x_i; \theta_{j-1})}{n_{j-1} q(x_i; \theta_{j-1}) + n_j e^{-\eta} q(x_i; \theta_j)},
\]

or equivalently to the equation

\[
1 = \sum_{i=1}^{N} \frac{n_{j-1} q(x_i; \theta_{j-1}) + n_j e^{-\eta} q(x_i; \theta_j)}{q(x_i; \theta_{j-1})}.
\]

Function (12) is concave and can be minimized by a fast, globally convergent algorithm such as R package trust. The estimator \( \hat{\eta} \) is consistent and asymptotically normal under a connectivity condition: \( \mu((x: q(x; \theta_{j-1}) > 0) \cap (x: q(x; \theta_j) > 0)) > 0 \), i.e., there is non-negligible overlap between \( P_{\theta_{j-1}} \) and \( P_{\theta_j} \). The asymptotic variance of \( \hat{\eta} \) increases as the degree of overlap between \( P_{\theta_{j-1}} \) and \( P_{\theta_j} \) decreases.

The baseline measure \( \mu \) is estimated as a discrete measure:

\[
\hat{\mu}(\{x_i\}) \propto (n_{j-1} q(x_i; \theta_{j-1}) + n_j e^{-\hat{\eta}} q(x_i; \theta_j))^{-1}, \quad i = 1, \ldots, N.
\]

For a general \( \theta \), the resulting estimator of \( Z_0/Z_{0,1} \) is

\[
\int \frac{q(x; \theta) d\hat{\mu}}{\int q(x; \theta_{j-1}) d\hat{\mu}} = \sum_{i=1}^{N} \frac{w(x_i; \theta)}{w_0(x_i; \theta)},
\]

where \( w(x_i; \theta) = q(x_i; \theta)/(n_{j-1} q(x_i; \theta_{j-1}) + n_j e^{-\hat{\eta}} q(x_i; \theta_j)) \). Taking \( \theta = \theta_j \) or \( \theta_{j-1} \) leads back to (13)–(14). The resulting estimator of \( E_0(\phi_0) \) is

\[
\int \frac{\phi(x; \theta) q(x; \theta) d\hat{\mu}}{\int q(x; \theta) d\hat{\mu}} = \sum_{i=1}^{N} \frac{\phi(x_i; \theta) w(x_i; \theta)}{w_0(x_i; \theta)}.
\]

The two estimators are essentially importance-sampling estimators, with the draws treated as from the mixture \((n_{j-1}/N)P_{\theta_{j-1}} + (n_j/N)P_{\theta_j} \). They are consistent and asymptotically normal under standard conditions for importance sampling, in addition to the connectivity condition. In the Appendix, for the case where the draws are independent from \( P_{\theta_{j-1}} \) and \( P_{\theta_j} \), we derive influence functions \( \nu(x; \theta) \) and \( \nu^h(x; \theta) \) such that the asymptotic variance of the log of estimator (15) can be consistently estimated by \( N^{-2} \sum_{i=1}^{N} (\nu^h(x_i; \theta))^2 \) and that of estimator (16) by \( N^{-2} \sum_{i=1}^{N} (\nu^h(x_i; \theta))^2 \).

As mentioned in Section 1, the likelihood estimators are statistically efficient if the draws are independent. By Meng and Wong (1996), \( \hat{\eta} \) achieves the minimum asymptotic variance for a class of estimators, \( \tilde{\eta} \), defined by

\[
\tilde{\eta} = \frac{E_{\theta_{j-1}} [\lambda(x) q(x; \theta_{j-1})]}{E_{\theta_j} [\lambda(x) q(x; \theta_j)]},
\]

where \( \lambda(x) \) is an arbitrary function such that \( \int \lambda(x) q(x; \theta_{j-1}) q(x; \theta_j) d\mu \) is nonzero. Many estimators are special cases of \( \tilde{\eta} \) with, for example, \( \lambda(x) = q^{-1}(x; \theta_{j-1}) \) or \( q^{-1}(x; \theta_j) \) or \( q(x; \theta_{j-1}) q(x; \theta_j) \). By an extension of Tan (2004), the likelihood estimator of the vector \((Z_0/Z_{0,1}) \), \( \lambda(x) q(x; \theta) d\mu/Z_{0,1} \) has no greater asymptotic variance matrix than any vector-valued estimator with elements

\[
E_{\theta_{j-1}} \left[ \frac{q(x; \theta_{j-1})}{q(x; \theta_{j-1})} \right] + \tilde{\eta} E_{\theta_j} \left[ \frac{q(x; \theta_j)}{q(x; \theta_j)} \right],
\]

(17)
\[
\hat{E}_{\theta_j} - \frac{\hat{\lambda}(x)q(x;\theta)}{q(x;\theta_j)} + \alpha^2 \hat{E}_{\theta_j} \left[ \frac{\phi(x;\theta)q(x;\theta)}{q(x;\theta_j)} \right],
\]

where \(\hat{\lambda}(x)\) and \(\lambda(x)\) are arbitrary functions. Therefore, estimator (16) has no greater asymptotic variance than any ratio of (18) over (17). For \(\hat{\theta} = \theta_{j-1}\) or \(\theta_j\), the sample average \(\hat{E}_{\theta_j}(\phi_{\theta_{j-1}})\) or \(\hat{E}_{\theta_j}(\phi_{\theta_j})\) corresponds to \(\hat{\lambda}(x) = \lambda(x) = 1\) or 0.

It is interesting to compare likelihood (or bridge-sampling) estimator \(\hat{\eta}\) and path-sampling estimator \((3)\). Recall that estimator (3) is based on Eq. (2), but \(E_t(U(x;\theta))\) is estimated only in \(t = \theta_{j-1}\) and \(\theta_j\) by the sample averages and then approximated by linear interpolation for \(\theta_{j-1} < t < \theta_j\). On one hand, estimator (3) is in general inconsistent unless \(\log Z_{\theta_j}\) is quadratic on \([\theta_{j-1}, \theta_j]\), whereas \(\hat{\eta}\) is consistent as long as \(P_{\theta_{j-1}}\) and \(P_{\theta_j}\) have non-negligible overlap. Moreover, if \(E_t(U(x;\theta))\) is estimated for \(\theta_{j-1} < t \leq \theta_j\) according to (16) by

\[
\int \frac{U(x;\theta)q(x;\theta)}{q(x;\theta_j)} \frac{d\hat{\mu}}{d\mu} = \frac{\int \frac{\hat{\partial}}{\partial t} q(x;\theta) \frac{d\hat{\mu}}{d\mu} \frac{d\mu}{dt} - \frac{d\mu}{dt} \log \left( \frac{\int q(x;\theta) \frac{d\hat{\mu}}{d\mu}}{\int q(x;\theta_{j-1}) \frac{d\hat{\mu}}{d\mu}} \right)},
\]

then Eq. (2) yields exactly \(\hat{\eta} = \log \left( \frac{\int q(x;\theta_j) \frac{d\hat{\mu}}{d\mu}}{\int q(x;\theta_{j-1}) \frac{d\hat{\mu}}{d\mu}} \right)\).

3.2. Computing normalizing constants

We now return to the setting where \(\theta\) is bivariate and \(\Theta\) consists of \(m_1 \times m_2\) grid points \(\theta^i = (\theta_{ij}, \theta_{ij})\), \(i = 1, \ldots, m_1, j = 1, \ldots, m_2\). To compute \(\log(Z_{\theta}/Z_{\theta_{j-1}})\), a similar method to path sampling is to pick a path \((\theta_{i_1}, \theta_{i_2}, \ldots, \theta_{i_{m_1}})\) with \(\theta_1 = \theta_{j-1}\) along the edges in \(\Theta\) and apply the two-sample likelihood (or bridge-sampling) method stepwise on the path. For every two consecutive points \(\theta_{j-1}\) and \(\theta_j\), denote

\[
\log \left( \frac{Z_{\theta_j}}{Z_{\theta_{j-1}}} \right) = \hat{\eta} \text{ based on the samples from } P_{\theta_{j-1}} \text{ and } P_{\theta_j}
\]

in parallel to \(\log(Z_{\theta}/Z_{\theta_{j-1}})\) in Section 2. The resulting single-path, bridge-sampling estimator of \(\log(Z_{\theta}/Z_{\theta_{j-1}})\) is

\[
\sum_{j=1}^{m_2} \log(Z_{\theta_j}/Z_{\theta_{j-1}}).
\]

Similarly as for path sampling, different choices of the path lead to different single-path estimators of \(\log(Z_{\theta}/Z_{\theta_{j-1}})\), which should ideally be identical. To exploit this information, we adopt the method of control variates by the same steps as in Section 2.1. For \(i = 1, \ldots, m_1\) and \(j = 1, \ldots, m_2\), redefine

\[
\xi_{i,j}^1 = \log \left( \frac{Z_{\theta_{i,j-1}}}{Z_{\theta_{i,j}}} \right) \quad (i \neq m_1),
\]

\[
\xi_{i,j}^2 = \log \left( \frac{Z_{\theta_{i,j+1}}}{Z_{\theta_{i,j}}} \right) \quad (j \neq m_2).
\]

Otherwise, define \(C, C_{\text{reg}}, \text{ and } D\) in the same way as in Section 2.1. Then \(C_{\text{reg}}\) serves as a vector of control variates, and \(C_{\text{reg}} = 0\). The vector of calibrated estimators \(D_{\text{reg}}\) for \(\log(Z_{\theta}/Z_{\theta_{j-1}})\) is invariant to different choices of \(D\), and has advantages over \(D_{\text{reg}}\) similarly as in Section 2.1.

A working estimator \(V\) of the asymptotic variance matrix of \(\xi\) is required for computing \(\xi_{\text{reg}}\). We apply the estimated influence function of \(\hat{\eta}\) from two-sample bridge sampling, derived in the Appendix under the assumption that the draws are independent. Denote by \(\nu_{ij}(x)\) the estimated influence function of \(\xi_{ij}^1\) and by \(\nu_{ij}^2(x)\) of \(\xi_{ij}^2\). The element of \(V\) for the asymptotic variance of \(\xi_{ij}^1\) is

\[
\frac{n^i\hat{E}_{\nu_{ij}}[(\nu_{ij}^1(x))^2] + n^{i+1}j\hat{E}_{\nu_{ij+1}}[(\nu_{ij}^1(x))^2]}{(n^i + n^{i+1})^2},
\]

and that of \(V\) for the asymptotic covariance of \(\xi_{ij}^1\) and \(\xi_{ij}^2\) is

\[
\frac{n^i\hat{E}_{\nu_{ij}}[\nu_{ij}^1(x)\nu_{ij}^2(x)]}{(n^i + n^{i+1})^2}.
\]
Other elements of $V$ are specified similarly. Both $V$ and $CVC^T$ are sparse as before, due to independence between samples from different distributions.

If $\Theta_i$ is coarser than $\Theta$, then it remains necessary to compute $\log(Z_{g_{i}}/Z_{b_{0}})$ ($\theta \in \Theta_{i}$), smoothly connecting the estimators of $\log(Z_{g_{i}}/Z_{b_{0}})$ ($\theta \in \Theta_{i}$). For $\theta$ in a rectangular cell $[\theta_{1i}, \theta_{1i+1}] \times [\theta_{2j}, \theta_{2j+1}]$, estimator (15) can in principle be applied to two samples on each of the four edges. Then $q(\mathbf{x}; \Theta_i)$ need to be evaluated on the corresponding pair of samples. Moreover, additional work is required to combine estimators from different edges or select one, for example, from the edge nearest to $\Theta$. To minimize the computational cost and exploit the development in Section 2.2, we propose a two-step method combining estimator (15), calibration, and interpolation. A basic idea is to realize that problem (9) is equivalent to the following problem:

Find a biquadratic function $G(\theta)$ with $c_{22} = 0$ such that $G(\theta^{ij}) = C^{ij}$
for $(i, j) = (i, j), (i, j+1), (i+1, j), (i+1, j+1)$,
and $(i+1/2, j), (i+1/2, j+1), (i+1, j+1/2),$ and $(i+1, j+1/2)$.

Here $\theta^{ij} = (\theta_{1i}, \theta_{2j})$ with $\theta_{1i+1/2} = (\theta_{1i} + \theta_{1i+1})/2$ and $\theta_{2j+1/2} = (\theta_{2j} + \theta_{2j+1})/2$, and $G^{ij}$ can be flexibly specified. The equivalence holds because if the values at the four corners are fixed for a biquadratic function, then the derivatives $(\partial G/\partial \theta_{1i})(\theta^{ij}), (\partial G/\partial \theta_{2j})(\theta^{ij}), (\partial G/\partial \theta_{1i})(\theta^{ij}+1), (\partial G/\partial \theta_{2j})(\theta^{ij}+1)$ have a one-to-one correspondence with the function values $G(\theta^{ij+1/2}), G(\theta^{ij+1/2}+1), G(\theta^{ij+1/2}+1), G(\theta^{ij+1/2})$.

First, for $i = 1, \ldots, m_1$ and $j = 1, \ldots, m_2$, let

$$z^{ij+1/2}_{1} = \log \left( \frac{Z_{g_{i}+1/2}}{Z_{g_{i}}} \right) (i \neq m_1), \quad z^{ij+1/2}_{1} = \log \left( \frac{Z_{g_{i}+1/2}}{Z_{g_{i+1}}} \right) \quad (i \neq m_1, j \neq m_2),$$

by applying estimator (15) to two samples from $P_{g_{i}}$ and $P_{g_{i+1}}$, and from $P_{g_{i}+1}$ and $P_{g_{i}+1}$, respectively. Similarly, let

$$z^{ij+1/2}_{2} = \log \left( \frac{Z_{g_{i}+1/2}}{Z_{g_{i}}} \right) (j \neq m_2), \quad z^{ij+1/2}_{2} = \log \left( \frac{Z_{g_{i}+1/2}}{Z_{g_{i+1}}} \right) \quad (i \neq m_1, j \neq m_2),$$

by applying estimator (15) to two samples from $P_{g_{i}}$ and $P_{g_{i+1}}$ and from $P_{g_{i}+1}$ and $P_{g_{i}+1}$, respectively. Arrange all $z^{ij+1/2}_{1}$, $z^{ij+1/2}_{2}$, and $z^{ij+1/2}_{2}$ into a vector $\xi^{i}$. Consider the regression estimator similar to $\xi^{\text{reg}}$

$$\xi^{i\text{reg}} = \xi^{i} - (V^C(CVC^T)^{-1})C\xi,$$

where $V$ is a working estimator of the asymptotic covariance matrix of $\xi^{i}$ and $\xi$, specified by using the estimated influence function of estimator (15). Denote by $(\xi^{i+1/2}_{1})_{\text{reg}}, (\xi^{i+1/2}_{2})_{\text{reg}}, \text{etc.}$ the corresponding elements of $\xi^{i\text{reg}}$.

Second, approximate $\log(Z_{g_{i}}/Z_{b_{0}})$ by the solution to problem (19) over each rectangular cell in $\Theta_i$, where $G_i$ is set to the calibrated estimator of $\log(Z_{g_{i}}/Z_{b_{0}})$, and $G^{i+1/2}_i, G^{i+1/2}_i, G^{i+1/2}_i, G^{i+1/2}_i$ are set to, respectively,

$$G_i + (\xi^{i+1/2}_{1})_{\text{reg}}, \quad G_i + (\xi^{i+1/2}_{2})_{\text{reg}}, \quad G_i + (\xi^{i+1/2}_{2})_{\text{reg}} + (\xi^{i+1/2}_{1})_{\text{reg}}.$$  

For every two cells sharing an edge, the corresponding solutions to problem (19) are identical on the common edge. The resulting approximation of $\log(Z_{g_{i}}/Z_{b_{0}})$ is continuous over the entire $\Theta_i$, similarly as in Section 2.2. The first-order derivatives may not be continuous on the edges including the corners. For comparison, those in Section 2.2 are continuous at the corners but may not be so elsewhere on the edges.

3.3. Computing expectations

For two samples on each edge in $\Theta_i$, estimator (16) for expectations can be applied in addition to estimator (15) for ratios of normalizing constants. For $i = 1, \ldots, m_1$ and $j = 1, \ldots, m_2$, there are multiple estimators of $E_{g_{i}}(\phi_{g_{i}})$:

$$\hat{g}^{ij}_{0} = \text{estimator (16) based on the samples from } P_{g_{i}} \text{ and } P_{g_{i}+1} \quad (i \neq m_1),$$
$$\hat{g}^{ij}_{1} = \text{estimator (16) based on the samples from } P_{g_{i}} \text{ and } P_{g_{i}+1} \quad (j \neq m_2),$$
$$\hat{g}^{ij}_{2} = \text{estimator (16) based on the samples from } P_{g_{i}} \text{ and } P_{g_{i}+1} \quad (i \neq 1),$$
$$\hat{g}^{ij}_{3} = \text{estimator (16) based on the samples from } P_{g_{i}} \text{ and } P_{g_{i}+1} \quad (j \neq 1),$$

by application of (16) on different edges attached to $\theta^{ij}$. These estimators should ideally be identical so that their differences serve as control variates, similarly as $\hat{\gamma}$ in Section 2.1. Let $\hat{\gamma}^{ij}_{0} = (\hat{g}^{ij}_{0}, \hat{g}^{ij}_{1}, \hat{g}^{ij}_{2}) - \hat{g}^{ij}_{3}$, with suitable modification when $i = 1$ or $m_1$ or $j = 1$ or $m_2$. In principle, a regression estimator of $E_{g_{i}}(\phi_{g_{i}})$ can be constructed using such control variates $\hat{\gamma}^{ij}_{0}$.
for all $i$ and $j$. However, from our simulations, this estimator of $E_{y}(\phi_{y})$ appears to perform no better than the regression estimator constructed using only control variates $\tau^{i}$. Incorporating control variates $\tau^{ij}$ for $(i,j) \neq (i,j)$ beyond $\tau^{i}$ does not yield worthwhile variance reduction. Let

$$\hat{q}_{\text{reg}}^{ij} = \hat{q}_{\text{reg}}^{ij} - \beta^{i} \tau^{i},$$

where $\beta^{i}$ is a working estimator of the asymptotic covariance matrix of $\hat{q}_{\text{reg}}^{ij}$ and $\tau^{i}$ right-multiplied by that of the inverse asymptotic covariance matrix of $\tau^{i}$, specified by using the estimated influence function of estimator (16). Equivalently, $\hat{q}_{\text{reg}}^{ij}$ is of form $\gamma_{ij}^{R} \hat{q}_{\text{reg}}^{ij} + \gamma_{ij}^{R} \hat{q}_{\text{reg}}^{ij} + \gamma_{ij}^{R} \hat{q}_{\text{reg}}^{ij} + \gamma_{ij}^{R} \hat{q}_{\text{reg}}^{ij} + \gamma_{ij}^{R} \hat{q}_{\text{reg}}^{ij} = 1$.

To compute $E_{y}(\phi_{y}) (\theta \in \Theta \Theta_{2})$, we propose a similar two-step method to that in Section 3.2. For each cell $[\theta_{1}, \theta_{2}, \theta_{3}, \ldots, \theta_{n}]$, we approximate $E_{y}(\phi_{y})$ by the solution to problem (19), where $\hat{G}_{ij}$ is set to $\hat{G}_{\text{reg}}^{ij}$ and $\hat{G}_{ij}$ is set to the following estimator $\hat{G}_{\text{reg}}^{ij}$ for $(i,j) = (i+1,j),(i+1,j+1),(i,j+1), \text{ and } (i+1,j+1/2)$. Let

$$\hat{q}_{\text{reg}}^{i+1/2,j} = \text{the regression estimator of } E_{y}(\phi_{y+1/2}) \text{ similar to } \hat{q}_{\text{reg}}^{ij},$$

$$\hat{q}_{\text{reg}}^{i+1/2,j+1} = \text{the regression estimator of } E_{y}(\phi_{y+1/2}) \text{ similar to } \hat{q}_{\text{reg}}^{ij},$$

$$\hat{q}_{\text{reg}}^{i+1/2,j} = \frac{\hat{q}_{\text{reg}}^{ij} + \hat{q}_{\text{reg}}^{i+1/2,j}}{2}.$$

The estimators $\hat{q}_{\text{reg}}^{i+1/2,j+1}$, $\hat{q}_{\text{reg}}^{i+1/2,j}$, and $\hat{q}_{\text{reg}}^{i+1/2}$ are defined similarly. These estimators of $E_{y}(\phi_{y})$ for $\theta$ on the edges are of different form than those of log(Z$_{\theta}/Z_{\theta}$) in Section 3.1. For example, $\hat{q}_{\text{reg}}^{1+1/2} = \hat{q}_{\text{reg}}^{1+1/2}$ is an average of two regression estimators depending on different control variates used in $\hat{G}_{ij}$ and $\hat{G}_{ij+1}$. In contrast, $(\Sigma^{1+1/2})_{\text{reg}}$ is a regression estimator depending on the same control variates $\Sigma^{1}$ as used in $\Sigma^{1}$.

4. Numerical examples

We provide two examples on a censored Gaussian random field and a Gaussian state-space model. The true values can be computed independently by a specialized method of Genz (1992) in the first example and obtained analytically in the second example, so that the simulation results can be properly evaluated. However, the proposed methods are applicable without relying on any special feature of these examples.

4.1. Censored Gaussian random field

Consider a Gaussian random field measured on a regular $6 \times 6$ grid in $[0, 1)^{2}$ but right-censored at 0 as in Stein (1992). Let $(u_{1}, \ldots, u_{2})$ be $K = 36$ locations of the grid, $y_{\text{com}} = (x_{1}, \ldots, x_{6})$ be the uncensored data, and $y_{\text{obs}} = (y_{1}, \ldots, y_{6})$ be the observed data such that $y_{j} = \max(x_{j}, 0)$ for $j = 1, \ldots, 36$. Assume that $y_{\text{com}}$ is multivariate Gaussian with $E(x_{j}) = \beta$ and $\text{cov}(x_{j}, x_{j}) = \sigma^{2} e^{-|w_{j}-w_{j}|}$ for $j = 1, \ldots, K$, where $\| \cdot \|$ is the Euclidean norm. Let

$$p(y_{\text{com}}; \theta) = \frac{1}{(2\pi\sigma^{2})^{K/2} \det^{-1/2}(\Sigma)} \exp\left(-\frac{1}{2\sigma^{2}}(y_{\text{com}} - \beta)^{T}\Sigma^{-1}(y_{\text{com}} - \beta)/(2\sigma^{2})\right),$$

where $\theta = (\beta, \log \sigma^{2})$ and $\Sigma$ is the correlation matrix of $y_{\text{com}}$. For the dataset in Stein (1992, Figure 1), we are interested in computing the likelihood

$$L(\theta) = \int_{0}^{1} \cdots \int_{0}^{1} p(y_{\text{com}}; \theta) \prod_{j \neq 0} dx_{j}$$

for $\theta \in \Theta$ where $\Theta$ is a $21 \times 21$ grid in $[-2.5, 2.5] \times [-2, 1]$. There are 17 censored observations in Stein's dataset and hence $L(\theta)$ is a 17-dimensional integral. We treat $p(y_{\text{com}}; \theta)$ as an unnormalized density function and $L(\theta)$ as the normalizing constant for the conditional distribution $p(y_{\text{com}} | y_{\text{obs}}, \theta)$.

First, we take $\Theta = \Theta$ and use the Gibbs sampler to simulate a Markov chain of length 500 after 50 burn-in iterations from $p(y_{\text{com}} | y_{\text{obs}}, \theta)$ for each $\theta \in \Theta$. We compare the following methods for estimating the log-likelihood ratio, $\log(L(\theta)/L(\theta_{0}))$, for $\theta \in \Theta$, where $\theta_{0}$ is set to $(0, -0.5)$, the center of $\Theta$. For reference, the true value of $L(\theta_{0})$ is $-20.023$ by the method of Genz (1992).

(a) Chib's estimator: Apply the estimator of Chib (1995) for $L(\theta)$ based on solely the sample from $p(y_{\text{com}} | y_{\text{obs}}, \theta)$.
(b) Single-path estimator: Apply path sampling along a single path, either $\beta \rightarrow \beta \log \sigma^{2} \rightarrow \beta \log \sigma^{2} \rightarrow \beta \log \sigma^{2}$, referred to as $\beta$-first or log $c$-first, respectively.
(c) Averaged path-sampling estimator: Apply the method of Gelman and Meng (1998, Section 2.3) by averaging single-path estimators along either $\beta$-first paths or log $c$-first paths, as illustrated in Fig. 1.
(d) Calibrated path-sampling estimator: Apply $D_{\text{reg}}^{c}$ from Section 2.1.
(e) Calibrated stepwise-bridge-sampling estimator: Apply $D_{\text{reg}}^{c}$ from Section 3.2.
addition to estimating that for standardization. Each cell gives the results using 50 thinned draws (upper) and 500 unthinned draws (lower) per distribution. The true values are smaller than that along the direction of log c.

The reduction in MSEs seems moderate mainly because the variation of path sampling along the direction of log c is substantially increased biases, compared with the case of 5 grid.

The stepwise bridge-sampling method improves the averaged path-sampling estimator when using 500 unthinned draws per distribution. This phenomenon may be caused jointly by several factors including finite-sample biases of sample averages from Markov chains, approximation bias of the trapezoidal rule, and propagation of those biases from \( \tilde{\xi} \) to \( \tilde{\xi}_{\text{reg}} \). For 11 \( \times \) 5 grid \( \Theta_s \), the path-sampling method yields similar variances but substantially increased biases, compared with the case of 5 \( \times \) 11 grid \( \Theta_s \). Moreover, the stepwise bridge-sampling method appears to break down, which indicates that there is a lack of overlap between nearby distributions \( p(\mathbf{y}_{\text{com}}| \mathbf{y}_{\text{obs}}; \theta) \) along the direction of log c.

**Table 1**

<table>
<thead>
<tr>
<th></th>
<th>Chib</th>
<th>Single path</th>
<th>Averaged path</th>
<th>Calibrated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \beta ) first</td>
<td>( \log c ) first</td>
<td>( \beta ) first</td>
<td>( \log c ) first</td>
</tr>
<tr>
<td>CPU</td>
<td>1.17</td>
<td>1.25</td>
<td>1.05</td>
<td>1.05</td>
</tr>
<tr>
<td>( 10^3 \text{Bias}^2 )</td>
<td>.0239</td>
<td>.0397</td>
<td>.0397</td>
<td>.217</td>
</tr>
<tr>
<td>( 10^3 \text{Var} )</td>
<td>1.45</td>
<td>1.89</td>
<td>1.89</td>
<td>20.9</td>
</tr>
<tr>
<td>( 10^3 \text{MSE} )</td>
<td>1.46</td>
<td>1.91</td>
<td>1.91</td>
<td>31.0</td>
</tr>
</tbody>
</table>

Note: Bias\(^2 = \sum_{\theta \in \Theta} \text{Bias}_h^2 / 441\), Var = \( \sum_{\theta \in \Theta} \text{Var}_h / 441\), and MSE = \( \sum_{\theta \in \Theta} \text{MSE}_h / 441\), where Bias\(_\theta\), Var\(_\theta\), and MSE\(_\theta\) are the empirical bias, variance, and MSE based on 100 repeated simulations. CPU is the CPU time, \( a + b \), with \( a \) for evaluating estimators and \( b \) for simulating Markov chains in R, both divided by \( a \) for standardization. Each cell gives the results using 50 thinned draws (upper) and 500 unthinned draws (lower) per distribution. The true values are computed by the method of Genz (1992) with estimated errors <.001 for all \( \theta \in \Theta \).

**Table 2**

<table>
<thead>
<tr>
<th></th>
<th>( \Theta_s = 11 \times 11 ) grid</th>
<th>( \Theta_s = 5 \times 11 ) grid</th>
<th>( \Theta_s = 11 \times 5 ) grid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Path</td>
<td>S-bridge</td>
<td>Path</td>
</tr>
<tr>
<td>( 10^3 \text{Bias}^2 )</td>
<td>.153</td>
<td>.416</td>
<td>.356</td>
</tr>
<tr>
<td>( 10^3 \text{Var} )</td>
<td>.646</td>
<td>.869</td>
<td>.463</td>
</tr>
<tr>
<td>( 10^3 \text{MSE} )</td>
<td>.473</td>
<td>.994</td>
<td>43.8</td>
</tr>
</tbody>
</table>

Note: Each choice of \( \Theta_s \) is regularly spaced within \( \Theta \). Bias\(^2\), Var, and MSE are averages over estimators at all \( \theta \in \Theta \) as in Table 1.
4.2. Gaussian state-space model

Consider a linear Gaussian state-space model:

\[ x_t = 0.5x_{t-1} + u_t, \quad y_t = x_t + v_t, \]

where \( x_0 \equiv 0 \) and \( u_t \sim N(0, \sigma_u^2) \) and \( v_t \sim N(0, \sigma_v^2) \) are mutually independent for \( t = 1, \ldots, T = 100 \). The vector \( y = (y_1, \ldots, y_T) \) is observed, but \( x = (x_1, \ldots, x_T) \) is unobserved. Let \( p(y|x; \theta) = \prod_{t=1}^{T} \frac{1}{(2\pi \sigma_y^2)} \exp(-\frac{(y_t - x_t)^2}{2\sigma_y^2}) \), where \( \theta = (\sigma_u, \sigma_v) \). We aim to compute the likelihood

\[ L(\theta) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x, y; \theta) \, dx_1 \cdots dx_T \]

for \( \theta \in \Theta \), where \( \Theta \) is a 21 \times 21 regular grid in \([-3, 3] \times [-3, 3]\). In the setting of Section 1, we treat \( p(y|x; \theta) \) as an unnormalized density function and \( L(\theta) \) as the normalizing constant for the conditional distribution \( p(x|y; \theta) \). Moreover, we consider computing expectations such as \( \mathbb{E}_C \left[ \log p(y|x; \theta) \right] \), which are smaller by a factor of about 2 than those of the two single-path estimators.

4.2.1. Gibbs sampling

We adopt Gibbs sampling to generate draws from \( p(x|y; \theta) \). First, we take \( \Theta_0 = \Theta \) and simulate a Markov chain of length 500 after 50 burn-in iterations from \( p(x|y; \theta) \) for each \( \theta \in \Theta \). Table 3 summarizes the results from 100 repeated simulations for the same methods as in Table 1 for estimating \( \log(L(\theta)/L(\theta_0)) \), where \( \theta_0 = (0, 0) \).

There are interesting differences between Tables 1 and 3 in the relative performances of these methods. Chib’s estimator yields MSEs greater by a factor of 3 than those of the two calibrated estimators. The two averaged path-sampling estimators have similar MSEs to each other, which are smaller by a factor of about 2 than those of the two single-path estimators. The two calibrated estimators yield even smaller MSEs by a factor of about 10 than those of the two single-path estimators.

Table 4 summarizes the results for interpolation in addition to estimation with three choices of \( \Theta_0 \), similarly as Table 2. For all the choices of \( \Theta_0 \), the methods based on path-sampling and stepwise bridge-sampling have similar variances, but the latter method yields much smaller biases. The method of stepwise bridge-sampling, however, performs similarly with 5 \times 11 and 11 \times 5 grids \( \Theta_0 \), which reflects that there are reasonable overlaps between nearby distributions \( p(x|y; \theta) \) along both directions of \( \log \sigma_u \) and \( \log \sigma_v \).

Next, we apply path sampling in Section 2.3 to estimate \( E(x_{900}|y; \theta) - E(x_{900}|y; \theta_0) \) and stepwise bridge-sampling in Section 3.3 to estimate \( E(x_{900}|y; \theta) \). Table 5 summarizes the results from 100 repeated simulations for the regression estimators of

| Table 3 | Log-likelihood ratio for state-space model (\( \Theta_0 = \Theta \); Gibbs sampling). |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                | Chib            | Single path     | Averaged path   | Calibrated      |
| CPU             |                 |                 |                 |                 |
|                 |                 |                 |                 |                 |
| \( 10^3 \text{Bias}^2 \) | .193            | .332            | .217            | .131            | .156            | .134            | .0933           |
|                 | .160            | .0348           | .0419           | .0198           | .0286           | .0166           | .00740          |
| \( 10^3 \text{Var} \) | 8.58            | 28.5            | 28.7            | 15.5            | 14.9            | 2.81            | 2.81            |
|                 | 9.73            | 3.02            | 3.07            | 1.57            | 1.55            | 2.88            | 2.87            |
| \( 10^3 \text{MSE} \) | 8.69            | 28.5            | 28.7            | 15.5            | 14.9            | 2.92            | 2.88            |
|                 | 1.12            | 3.02            | 3.08            | 1.58            | 1.57            | .302            | .292            |

Note: Each cell gives the results using 50 thinned draws (upper) and 500 unthinned draws (lower) per distribution.

| Table 4 | Log-likelihood ratio for state-space model (\( \Theta_0 \subset \Theta \); Gibbs sampling). |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| \( \Theta_0 = 11 \times 11 \text{ grid} \) | \( \Theta_0 = 5 \times 11 \text{ grid} \) | \( \Theta_0 = 11 \times 5 \text{ grid} \) |
| Path             | S-bridge        | Path             | S-bridge        | Path             | S-bridge        |
| \( 10^3 \text{Bias}^2 \) | .286            | .105            | 2.27            | .349            | 1.19            | .0655           |
|                 | .0830           | .0156           | 1.08            | .0724           | .852            | .0166           |
| \( 10^3 \text{Var} \) | 8.02            | 7.62            | 14.7            | 14.8            | 13.8            | 14.8            |
|                 | 9.38            | .897            | 1.84            | 1.93            | 1.51            | 1.55            |
| \( 10^3 \text{MSE} \) | 8.23            | 7.65            | 16.9            | 15.0            | 14.9            | 14.8            |
|                 | 1.01            | .904            | 2.91            | 1.98            | 2.34            | 1.55            |
the two methods together with the crude estimator \( \hat{E}(x_{90}) \). The path-sampling estimator of \( E(x_{90} \mid y; \theta) \) is taken as that of \( E(x_{90} \mid y; \theta) - E(x_{90} \mid y; \theta_0) \) plus \( \hat{E}_b(x_{90}) \).

To estimate \( E(x_{90} \mid y; \theta) - E(x_{90} \mid y; \theta_0) \), the path-sampling estimator is the most accurate among the three, with smaller MSEs by a factor of 13–20 than those of the sample-average estimator. On the other hand, to estimate \( E(x_{90} \mid y; \theta) \), the stepwise bridge-sampling estimator is the most accurate among the three, with smaller MSEs by a factor of about 3 than those of the sample-average estimator.

### 4.2.2. Particle filtering

To demonstrate the applicability of the proposed methods, we adopt particle filtering to generate draws from \( p(x \mid y; h) \). Particle filtering is also known as sequential importance sampling with resampling and extensively studied in connection with state-space models (e.g., Liu, 2001; Robert and Casella, 2005). A sample of draws generated by particle filtering are interdependent in a different manner than serial dependency in a Markov chain. Therefore, the simple device of thinning cannot be used to extract approximately independent draws from the original sample.

We apply particle filtering to simulate 500 or 2000 draws from \( p(x \mid y; h) \). Tables 6–8 summarize the results from 100 repeated simulations in parallel to Tables 3–5, except that Chib's estimator is replaced by a raw estimator based on importance weights (Liu, 2001). The relative performances of various methods are overall similar to those in Tables 3–5. However, there are interesting new patterns. In Table 6, the two calibrated estimators have appreciable biases relatively to their standard deviations when using 500 draws per distribution, although the biasedness decreases when the sample size

### Table 5

<table>
<thead>
<tr>
<th></th>
<th>( E(x_{90} \mid y; \theta) - E(x_{90} \mid y; \theta_0) )</th>
<th>( E(x_{90} \mid y; \theta) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Crude</td>
<td>Path</td>
</tr>
<tr>
<td>( 10^3 \text{Bias}^2 )</td>
<td>.146</td>
<td>.0882</td>
</tr>
<tr>
<td>( 10^3 \text{Var} )</td>
<td>17.0</td>
<td>1.22</td>
</tr>
<tr>
<td>( 10^3 \text{MSE} )</td>
<td>17.0</td>
<td>1.29</td>
</tr>
</tbody>
</table>

### Table 6

<table>
<thead>
<tr>
<th></th>
<th>Raw</th>
<th>Single path</th>
<th>Averaged path</th>
<th>Calibrated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>log ( \sigma_x ) first</td>
<td>log ( \sigma_y ) first</td>
<td>log ( \sigma_x ) first</td>
<td>log ( \sigma_y ) first</td>
</tr>
<tr>
<td>CPU</td>
<td>1+.001</td>
<td>1+.005</td>
<td>1+.01</td>
<td>1+.01</td>
</tr>
<tr>
<td>( 10^2 \text{Bias}^2 )</td>
<td>1.04</td>
<td>.911</td>
<td>.813</td>
<td>1.17</td>
</tr>
<tr>
<td>( 10^2 \text{Var} )</td>
<td>47.1</td>
<td>16.4</td>
<td>16.0</td>
<td>8.83</td>
</tr>
<tr>
<td>( 10^2 \text{MSE} )</td>
<td>47.6</td>
<td>17.2</td>
<td>16.7</td>
<td>9.91</td>
</tr>
<tr>
<td>11.8</td>
<td>3.91</td>
<td>4.48</td>
<td>2.24</td>
<td>2.19</td>
</tr>
</tbody>
</table>

Note: Each cell gives the results using 500 draws (upper) and 2000 draws (lower) per distribution.

### Table 7

<table>
<thead>
<tr>
<th>( \theta_i = 11 \times 11 \text{ grid} )</th>
<th>( \theta_i = 5 \times 11 \text{ grid} )</th>
<th>( \theta_i = 11 \times 5 \text{ grid} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Path</td>
<td>S-bridge</td>
<td>Path</td>
</tr>
<tr>
<td>( 10^2 \text{Bias}^2 )</td>
<td>1.70</td>
<td>1.54</td>
</tr>
<tr>
<td>( 10^2 \text{Var} )</td>
<td>5.14</td>
<td>4.95</td>
</tr>
<tr>
<td>( 10^2 \text{MSE} )</td>
<td>6.799</td>
<td>6.459</td>
</tr>
<tr>
<td>1.42</td>
<td>1.34</td>
<td>2.67</td>
</tr>
</tbody>
</table>
increases to 2000. The reduction in MSEs by using the two calibrated estimators over the single-path and averaged-path estimators stays at similar levels, but that over the raw estimator becomes more substantial than over Chib’s estimator.

5. Conclusion

Given probability distributions with unnormalized density functions indexed by parameters on a 2-dimensional grid, we have proposed new methods for computing normalizing constants and expectations using samples from distributions on a subgrid. These methods are built on either path sampling or bridge sampling, but designed to exploit linear constraints on single-path estimators for variance reduction.

The proposed methods have the advantage of being generally applicable to draws generated by various Markov chain or sequential Monte Carlo algorithms. Moreover, implementation of these methods does not depend on the details of sampling algorithms. Therefore, these methods can be used as a self-contained package for output analysis, regardless of the complexity of sampling algorithms.

The methods of path sampling and stepwise bridge sampling are related but rely on different operating conditions. Path sampling is in general biased unless the log normalizing constant is quadratic along the edge linking two nearby distributions. Bridge sampling is consistent only if the two distributions have a non-negligible overlap. The two methods can be extended to a 3 or higher dimensional grid. It is straightforward to construct single-path estimators and identify linear constraints on the faces of the grid. On the other hand, the computational cost seems prohibitive for a large, high-dimensional grid. Then it may be useful to develop a single-loop approach by randomly generating both parameters \( \theta \) and observations \( x \) as in Gelman and Meng (1998).

Appendix

Proof of Proposition 1. Result (i) is already explained in the text. To show (ii), we explicitly derive a unique solution to problem (7) if \( c_{22} = 0 \) and condition (8) holds. Write \( G(\theta) = \sum_{i=0}^{2} \sum_{j=0}^{2} c_{i,j} (\theta_{1} - \theta_{1i})^{i} (\theta_{2} - \theta_{2j})^{j} \) with \( c_{22} = 0 \). Then \( c_{00} = 0 \) because \( G(\theta^0) = 0 \). The remaining coefficients \( c_{i,j} \) satisfy the equations

\[
\begin{align*}
\beta_{1}^{i} &= c_{10}, \\
\beta_{1}^{i} \beta_{1}^{j} &= c_{10} + 2c_{20}(\theta_{1,i+1} - \theta_{1i}), \\
\beta_{1}^{i+1} \beta_{1}^{j} &= c_{10} + c_{11}(\theta_{2,j+1} - \theta_{2j}) + c_{12}(\theta_{2,j+1} - \theta_{2j})^{2}, \\
\beta_{1}^{i+1} \beta_{1}^{j+1} &= c_{10} + c_{11}((\theta_{2,j+1} - \theta_{2j})^{2}) + c_{12}((\theta_{2,j+1} - \theta_{2j})^{2} + 2c_{20}(\theta_{1,i+1} - \theta_{1i}) + 2c_{21}(\theta_{1,i+1} - \theta_{1i})(\theta_{2,j+1} - \theta_{2j}),
\end{align*}
\]

and four similar equations regarding \( \gamma_{2}, \gamma_{2}^{i+1}, \gamma_{2}^{i+1}, \) and \( \gamma_{2}^{i+1}+1 \). Then it is straightforward to show that

\[
\begin{align*}
\beta_{1}^{10} &= \beta_{1}^{1}, \hspace{2cm} \beta_{20} = \frac{\beta_{1}^{1} - \beta_{1}^{1}}{2(\theta_{1,i+1} - \theta_{1i})}, \hspace{2cm} \beta_{21} = \frac{\beta_{1}^{1} - \beta_{1}^{1} - \beta_{1}^{1} + \beta_{1}^{1}}{2(\theta_{1,i+1} - \theta_{1i})(\theta_{2,j+1} - \theta_{2j})},
\end{align*}
\]
There are two equivalent expressions for \(c_{11}\) due to condition (8).

**Proof of Proposition 2.** Write \(A_\theta = \int \phi(x; \theta)q(x; \theta)\, \text{d}\mu\). Then \(E_\theta(\phi_\theta) = A_\theta/Z_\theta\) and

\[
\partial/\partial \theta E_\theta(\phi_\theta) = \partial/\partial \theta Z_\theta - A_\theta \partial/\partial \theta Z_\theta
\]

The first item on the right side is

\[
\partial/\partial \theta A_\theta = \int \left( \partial/\partial \theta \phi(x; \theta) \right) q(x; \theta) + \phi(x; \theta) \left( \partial/\partial \theta q(x; \theta) \right) \, \text{d}\mu
\]

\[
= E_\theta \left\{ \partial/\partial \theta \phi(x; \theta) \right\} + E_\theta(\phi(x; \theta)U(x; \theta))
\]

assuming the legitimacy of interchange of differentiation and integration. The second item is \(E_\theta(\phi(x; \theta)E_\theta(U(x; \theta))\) due to (1). Therefore, Eq. (10) holds.

**Influence functions for bridge sampling.** We derive the influence functions of estimators (15) and (16), denoted by \(Z_\theta/Z_{\theta,1}\) and \(E_\theta(\phi_\theta)\), provided that the draws are independent from \(P_{\theta,1}\) and \(P_\theta\). Let \(\pi = \{(n_{i-1},n_i)\}/(n_{i-1}+n_i)\) and \(\delta(x) = (1,0)^T\) or \((0,1)^T\) if \(x\) is generated from \(P_{\theta,1}\) or \(P_\theta\), i.e., \(\delta(x)\) is the vector of stratum indicators for each \(x\) in the pooled sample. Let \(Q(x) = (q(x; \theta_j), q(x; \theta), \phi(x; \theta)q(x; \theta))^T\) and \(Z = \int Q(x) \, \text{d}\mu / \int q(x; \theta_{j-1}) \, \text{d}\mu\). By formula (8) of Tan (2004), the asymptotic variance matrix of \(\hat{Z}\) can be consistently estimated by \(N^{-2} \sum_{i=1}^N \psi(x)\psi(x)^T\) with

\[
\psi(x) = H^{-1}[W(x) - \tilde{Z} - L^T(\delta(x) - \pi)],
\]

where \(W(x) = Q(x)/(n_{i-1}q(x; \theta_{j-1}) + n_ie^{-\delta(x)q(x; \theta_j)})\), \(L = [E_{\theta_{j-1}}(W(x)), E_\theta(W(x))]^T\), and \(H = (N^{-1} \sum_{i=1}^N W(x_i)W(x_i)^T)\) diag\((e^{-\delta(x)q(x; \theta_{j-1})})\). Then an estimated influence function of \(Z_{\theta,1}/Z_{\theta,0}\) is \(\psi_2(x)\) and that of \(E_\theta(\delta_{\theta_{j-1}})\) is \(\psi_3(x) = \psi_2(x) - E_\theta(\phi_\theta)\psi_2(x)/(Z_\theta/Z_{\theta,0})\).

**References**


