Statistica Sinica **34** (2024), 27-46
doi:https://doi.org/10.5705/ss.202021.0049

SELECTION OF PROPOSAL DISTRIBUTIONS FOR MULTIPLE IMPORTANCE SAMPLING

Vivekananda Roy* and Evangelos Evangelou

Iowa State University and University of Bath

Abstract: In general, the naive importance sampling (IS) estimator does not work well in examples involving simultaneous inference on several targets, because the importance weights can take arbitrarily large values, making the estimator highly unstable. In such situations, researchers prefer alternative multiple IS estimators involving samples from multiple proposal distributions. Just like the naive IS, the success of these multiple IS estimators depends crucially on the choice of the proposal distributions, which is the focus of this study. We propose three methods: (i) a geometric space-filling approach, (ii) a minimax variance approach, and (iii) a maximum entropy approach. The first two methods apply to any IS estimator, whereas the third approach is described in the context of a two-stage IS estimator. For the first method, we propose a suitable measure of "closeness" based on the symmetric Kullback–Leibler divergence and the second and third approaches use estimates of asymptotic variances of an IS estimator and the reverse logistic regression estimator, respectively. Thus, when samples from the proposal distributions are obtained by running Markov chains, we provide consistent spectral variance estimators for these asymptotic variances. Lastly, we demonstrate the proposed methods for selecting proposal densities using various detailed examples.

Key words and phrases: Bayes factor, central limit theorem, marginal likelihood, Markov chain, polynomial ergodicity, reverse logistic regression.

1. Introduction

Importance sampling (IS) is a popular Monte Carlo procedure in which samples from one distribution are weighted in order to estimate features of other distributions. Here, we consider IS in the context of the following problem. Let Π be the family of target densities on the space X with respect to a measure μ, where $\pi(x) = \nu(x)/\theta \in \Pi$. Here, $\nu(x)$ is known, but the normalizing constant $\theta = \int_{\mathsf{X}} \nu(x)\mu(dx)$ is unknown. Let f be a π-integrable, real-valued function defined on X, for all $\pi \in \Pi$. There are two goals. The first goal is to estimate the normalizing constants θ up to a constant of proportionality for all $\pi \in \Pi$. The second goal is to estimate the integrals $E_\pi f := \int_{\mathsf{X}} f(x)\pi(x)\mu(dx)$ for all $\pi \in \Pi$. Estimations of normalizing constants play important roles in frequentist and Bayesian inference, as well as in other areas, such as statistical physics. In Bayesian statistics, the ratio of normalizing constants for two different posteriors

*Corresponding author.

is the Bayes factor, which is at the core of Bayesian hypothesis testing and model selection (Doss (2010)). The empirical Bayes estimate corresponds to the value of a hyper parameter where the normalizing constant (marginal likelihood) attains its maximum (Doss (2010); Roy, Evangelou and Zhu (2016)). In latent variable models, for example, generalized linear mixed models, the ratio of the normalizing constants is the likelihood ratio used for hypothesis testing (Christensen (2004)). The normalizing constants also need to be estimated in problems involving intractable likelihoods, such as exponential random graph models and autologistic models (Geyer and Thompson (1992)). Similarly, in statistical physics, an important problem is the estimation of some normalizing constants, known as the partition function. On the other hand, estimating the (posterior) means of certain functions f as the posterior density varies is the key issue in Bayesian sensitivity analysis (Buta and Doss (2011)). In Bayesian penalized regression methods, plotting regularization paths is equivalent to estimating the means of the regression coefficients as the penalty parameters vary (Roy and Chakraborty (2017)).

Our two objectives can be accomplished using naive importance sampling. Let $q_1(x) = \varphi_1(x)/c_1$ be another density on X with respect to μ, such that we are able to generate samples from q_1, and $\nu(x) = 0$ whenever $\varphi_1(x) = 0$. Indeed, if $\{X_i\}_{i=1}^n$ is either an independent and identically distributed (i.i.d.) samples from q_1 or a positive Harris recurrent Markov chain with invariant density q_1, then the naive IS estimator is consistent; that is,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\nu(X_i)}{\varphi_1(X_i)} \xrightarrow{\text{a.s.}} \int_{\mathsf{X}}\frac{\nu(x)}{\varphi_1(x)}q_1(x)\,\mu(dx) = \frac{\theta}{c_1}\int_{\mathsf{X}}\frac{\nu(x)/\theta}{\varphi_1(x)/c_1}q_1(x)\,\mu(dx) = \frac{\theta}{c_1}. \quad (1.1)$$

Similarly, we can estimate $E_\pi f$ using the ratio of $(1/n)\sum_{i=1}^n[f(X_i)\nu(X_i)/\varphi_1(X_i)]$ and the estimator in (1.1). These naive IS estimators suffer from high variance when the target probability density function (pdf) π is not "close" to the proposal pdf q_1 (Geyer (2011)) because, in that case, the ratio $\nu(X_i)/\varphi_1(X_i)$ takes arbitrarily large values for some X_i.

To alleviate this issue, we can use samples from multiple proposals, properly weighted, as in the variants of multiple importance sampling (Veach and Guibas (1995); Owen and Zhou (2000); Elvira et al. (2019)), umbrella sampling (Geyer (2011); Doss (2010)), and parallel, serial, or simulated tempering (George and Doss (2018); Geyer and Thompson (1995); Marinari and Parisi (1992)). In an IS estimation based on multiple proposal densities, we usually replace the single density q_1 with a linear combination of k densities (Geyer (2011)). In particular, let $q_i(x) = \varphi_i(x)/c_i$, for $i = 1, \ldots, k$, be k densities from the set of potential proposal densities $Q \equiv \{q(x) = \varphi(x)/c\}$, where φ_i are known but c_i may be unknown. Let $\boldsymbol{a} = (a_1, \ldots, a_k)$ be a vector of k positive constants such that $\sum_{i=1}^k a_i = 1$, $\overline{q} \equiv \sum_{i=1}^k a_i q_i$, $d_i = c_i/c_1$, for $i = 1, 2, \ldots, k$, with $d_1 = 1$, and

$\boldsymbol{d} \equiv (c_2/c_1, \ldots, c_k/c_1)$. For $l = 1, \ldots, k$, let $\{X_i^{(l)}\}_{i=1}^{n_l}$ be either i.i.d. samples from q_l or a positive Harris recurrent Markov chain with invariant density q_l. Then, as $n_l \to \infty, \forall\, l$,

$$\begin{aligned}\hat{u} &\equiv \sum_{l=1}^{k} \frac{a_l}{n_l} \sum_{i=1}^{n_l} \frac{\nu(X_i^{(l)})}{\sum_{j=1}^{k} a_j \varphi_j(X_i^{(l)})/d_j} \xrightarrow{\text{a.s.}} \sum_{l=1}^{k} a_l \int_{\mathsf{X}} \frac{\nu(x)}{\sum_{j=1}^{k} a_j \varphi_j(x)/d_j} q_l(x)\, \mu(dx) \\ &= \frac{1}{c_1} \int_{\mathsf{X}} \frac{\nu(x)}{\overline{q}(x)} \overline{q}(x)\, \mu(dx) = \frac{\theta}{c_1}. \end{aligned} \tag{1.2}$$

Similarly, we estimate $E_\pi f$ using $\hat{\eta}^{[f]} \equiv \hat{v}^{[f]}/\hat{u}$, where

$$\hat{v}^{[f]} := \sum_{l=1}^{k} \frac{a_l}{n_l} \sum_{i=1}^{n_l} \frac{f(X_i^{(l)})\nu(X_i^{(l)})}{\sum_{j=1}^{k} a_j \varphi_j(X_i^{(l)})/d_j}.$$

Estimations using (1.2) have been considered in several works (see, e.g., Gill, Vardi and Wellner (1988); Kong et al. (2003); Meng and Wong (1996); Tan (2004); Vardi (1985); Buta and Doss (2011); Geyer (1994); Tan, Doss and Hobert (2015)). Although alternative weighting schemes have been proposed, such as the population Monte Carlo of Cappé et al. (2004), none is as widely applicable as (1.2). If the normalizing constants c_i are known, the estimator (1.2) resembles the balance heuristic estimator of Veach and Guibas (1995), which is discussed in Owen and Zhou (2000) as a deterministic mixture. On the other hand, in several applications of IS methods, $\boldsymbol{d}$ in (1.2) is unknown. This occurs when $Q = \Pi$, that is, when samples from a subset of densities of Π are used to estimate the normalizing constants for the entire family using (1.2). Routine applications of IS estimation with $Q = \Pi$ can be found in Monte Carlo maximum likelihood estimation, Bayesian sensitivity analysis, and model selection (Geyer and Thompson (1992); Buta and Doss (2011); Doss (2010)). For unknown $\boldsymbol{d}$, Doss (2010) proposed a two-stage method. In the first step, using samples from q_i, for $i = 1, \ldots, k$, $\boldsymbol{d}$ is estimated by $\hat{\boldsymbol{d}}$ using Geyer's (1994) reverse logistic regression estimator or Meng and Wong's (1996) bridge sampling method. Then, independent of step one, new samples are used to calculate (1.2), with $\boldsymbol{d}$ replaced by $\hat{\boldsymbol{d}}$.

The effectiveness of (1.2) depends on the choices of k, $\boldsymbol{a}$, n_l, and the importance densities $\boldsymbol{q} = \{q_1, \ldots, q_k\}$. Here, we focus on choosing the importance densities, because they are the most crucial. Furthermore, the multiple IS estimator (1.2), just like the naive IS estimator (1.1), is useless if these densities are not "close" to some target densities. Although increasing k or n_l may lead to estimators with less variance, doing so results in a higher computational cost. Therefore, these are often determined based on available computational resources. On the other hand, for fixed k, $\boldsymbol{a}$, and n_l, we can improve the efficiency and stability of the estimator (1.2) by choosing the k importance densities $\boldsymbol{q}$ from the

set Q appropriately.

This study is the first to develop and test systematic methods for selecting proposal distributions for IS. We propose three approaches. (i) Our first approach is based on a geometric spatial design method, called the space-filling (SF) method. In particular, among all subsets $\boldsymbol{q} \subset Q$ with $|\boldsymbol{q}| = k$, we choose the one that minimizes the gaps between the elements of $\boldsymbol{q}$ and those of Π. The choice of the distance between the elements of $\boldsymbol{q}$ and Π is crucial, and here we propose using the symmetric Kullback–Leibler divergence. (ii) The second approach, called the minimax (MNX) method, chooses $\boldsymbol{q}$ that minimizes the maximum standard error, or the maximum relative standard error of the estimator $\hat{u}$ (or $\hat{\eta}^{[f]}$). (iii) Finally, the third approach is applicable when $\boldsymbol{d}$ in (1.2) is unknown. Here, we use Doss's (2010) two-stage IS method. In this approach, called the maximum entropy (ENT) method, following the maximum entropy criterion of experimental design, $\boldsymbol{q}$ is chosen by maximizing the determinant of the asymptotic covariance matrix of $\hat{\boldsymbol{d}}$. We describe and compare these three methods in Section 3. The methods apply to different situations. MNX is applicable to any IS estimator for which valid standard errors are available. Implementing MNX and ENT requires that we estimate the asymptotic variances in a central limit theorem. In the absence of such variance estimates, SF can be used. SF does not depend on the form of the IS estimator (1.2). Thus, the same SF distributions can be used for any IS estimator. However, a successful implementation of the SF, as shown later, depends crucially on the choice of the metric. Unlike the MNX design, which depends on the choice of the function f, the same SF and ENT proposals work regardless of whether we need to estimate the normalizing constants or the means. Overall, SF is the most straightforward to implement. However, it may not always be ideal, because it is independent of the form of the estimator and the estimand of interest. In our experience, with a properly chosen metric, it consistently provides desirable results. The three methods are implemented using the R package geoBayes (Evangelou and Roy (2022)). We illustrate these methods using several examples involving autologistic models, Bayesian regression models, and spatial generalized linear mixed models.

Unfortunately, few studies have examined the choice of the importance densities in multiple IS methods. However, given $\boldsymbol{q}$, in the special case when $\boldsymbol{d}$ is known and i.i.d. samples are available from the proposal densities, several methods exist for selecting the weights $\boldsymbol{a}$ (see, e.g., Li, Tan and Chen (2013)). One exception is Buta and Doss (2011), who describe an ad-hoc method in the important special case of $Q = \Pi$. Buta and Doss (2011) state that solving the minimax variance design problem, that is, the one that minimizes $\phi(\boldsymbol{q}) = \max_{\pi\in\Pi}\sigma_u^2(\pi;\boldsymbol{q})$ exactly, where $\sigma_u^2(\pi;\boldsymbol{q})$ is the asymptotic variance of $\hat{u}$ in (1.2), is "hopeless". Assuming that a consistent estimator $\hat{\sigma}_u^2(\pi;\boldsymbol{q})$ of $\sigma_u^2(\pi;\boldsymbol{q})$ is available, Buta and Doss (2011) propose a procedure in which they compute $\hat{\sigma}_u^2(\pi;\boldsymbol{q})$, for all $\pi \in \Pi$ starting from some "trial" proposal pdfs. Then, proposal

densities are either moved to regions of Π where $\hat{\sigma}_u^2(\pi; \boldsymbol{q})$ is large, or new proposal densities from these high variance regions are added by increasing k. Here, we develop a principled approach, called the sequential method (SEQ), that formalizes this procedure, and compare its performance with that of the three proposed methods.

As mentioned above, the MNX, ENT, and SEQ methods use asymptotic standard errors of $\hat{\boldsymbol{d}}$ and/or $\hat{u}$. A further contribution of this study is the development of spectral variance (SV) estimators of the asymptotic variances for $\hat{\boldsymbol{d}}$ and $\hat{u}$. The availability of consistent estimators is important in its own right, because it allows us to calculate asymptotically valid standard errors of the IS estimators. Recently, Roy, Tan and Flegal (2018) provided standard error estimators of $\hat{\boldsymbol{d}}$ and $\hat{u}$ using the batch means method. In other numerical examples (not shown here), we observe that the proposed SV estimators are, in general, less variable than the batch means estimators. This observation is in line with the findings of Flegal and Jones (2010), who show that, when estimating the means of scalar-valued functions, certain SV estimators are less variable than the batch means estimators by a factor of 1.5.

The rest of this paper is organized as follows. In Section 2, we describe the multiple IS estimation and the reverse logistic regression estimation. In Section 3, we describe the proposed methods for selecting the proposal densities for the IS estimators. Several illustrative examples are given in Section 4. Section 5 concludes the paper. The proofs of the theorems and several other examples are relegated to the Supplementary Material.

2. Multiple IS Estimation of Normalizing Constants and Expectations

Recall that $\Pi = \{\pi : \pi(x) = \nu(x)/\theta\}$ is a family of target densities on X, and that $f : \mathsf{X} \to \mathbb{R}$ is a function of interest. Given samples $\Phi_l \equiv \{X_i^{(l)}\}_{i=1}^{n_l}$, for $l = 1 \ldots, k$, from a small number of proposal densities $\{q_l = \varphi_l(x)/c_l, l = 1, \ldots, k\}$, we want to estimate θ (or, rather θ/c_1) and $E_\pi f$, for all $\pi \in \Pi$. Recall that we estimate $u(\pi, q_1) \equiv \theta/c_1$ and $E_\pi f$ by $\hat{u}(\boldsymbol{d}) \equiv \hat{u}(\pi; \boldsymbol{d})$ defined in (1.2) and $\hat{\eta}^{[f]} \equiv \hat{\eta}^{[f]}(\pi; \boldsymbol{d})$, respectively. We also consider the more general setting when $\boldsymbol{d}$ is unknown, which is the case if $Q = \Pi$. In such situations, we use the two-stage IS procedure of Doss (2010), where, $\boldsymbol{d}$ is first estimated using Geyer's (1994) reverse logistic regression method (described in Section 2.1) based on Markov chain samples $\tilde{\Phi}_l \equiv \{\tilde{X}_i^{(l)}\}_{i=1}^{N_l}$ with stationary density q_l, for $l = 1, \ldots, k$. Once we have $\hat{\boldsymbol{d}}$, independent of stage 1, we obtain new samples $\Phi_l \equiv \{X_i^{(l)}\}_{i=1}^{n_l}$, for $l = 1 \ldots, k$, to estimate $u(\pi, q_1)$ and $E_\pi f$ by $\hat{u}(\hat{\boldsymbol{d}})$ and $\hat{\eta}^{[f]}(\pi; \hat{\boldsymbol{d}})$, respectively. Buta and Doss (2011) quantify the benefits of using the two-stage scheme, rather than using the same samples to estimate both $\boldsymbol{d}$ and $u(\pi, q_1)$.

2.1. Reverse logistic regression estimator of $\boldsymbol{d}$

Let $N = \sum_{l=1}^{k} N_l$ and $a_l \in [0,1]$, for $l = 1, \ldots, k$, such that $\sum_{l=1}^{k} a_l = 1$. Define

$$\zeta_l = -\log(c_l) + \log(a_l),\ l = 1, \ldots, k, \tag{2.1}$$

and

$$p_l(x, \boldsymbol{\zeta}) = \frac{\varphi_l(x) e^{\zeta_l}}{\sum_{s=1}^{k} \varphi_s(x) e^{\zeta_s}},\ l = 1, \ldots, k, \tag{2.2}$$

where $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_k)$. (Note that if $a_l = N_l/N$, given that x belongs to the pooled sample $\{\tilde{X}_i^{(l)},\ i = 1, \ldots, N_l,\ l = 1, \ldots, k\}$, $p_l(x, \boldsymbol{\zeta})$ is the probability that x comes from the lth distribution.) Following Doss and Tan (2014), consider the log quasi-likelihood function

$$\ell_N(\boldsymbol{\zeta}) = \sum_{l=1}^{k} a_l \frac{N}{N_l} \sum_{i=1}^{N_l} \log\{p_l(\tilde{X}_i^{(l)}, \boldsymbol{\zeta})\}. \tag{2.3}$$

Note that adding the same constant to all ζ_l leaves (2.3) invariant. Let $\boldsymbol{\zeta}^0 \in \mathbb{R}^k$ denote the true $\boldsymbol{\zeta}$ normalized to add to zero, that is, $\boldsymbol{\zeta}_l^0 = \boldsymbol{\zeta}_l - \big(\sum_{j=1}^{k} \boldsymbol{\zeta}_j\big)/k$. Here, $\boldsymbol{\zeta}_l$ denotes the lth element of $\boldsymbol{\zeta}$. Note that the function $g\colon \mathbb{R}^k \to \mathbb{R}^{k-1}$ that maps $\boldsymbol{\zeta}^0$ into $\boldsymbol{d}$ is given by $g(\boldsymbol{\zeta}) = (e^{\zeta_1 - \zeta_2} a_2/a_1, e^{\zeta_1 - \zeta_3} a_3/a_1, \ldots, e^{\zeta_1 - \zeta_k} a_k/a_1)^\top$. We estimate $\boldsymbol{\zeta}^0$ by $\hat{\boldsymbol{\zeta}}$, where

$$\hat{\boldsymbol{\zeta}} = \operatorname{argmax} \ell_N(\boldsymbol{\zeta}) \text{ subject to } \sum_{j=1}^{k} \zeta_j = 0,$$

and thus obtain $\hat{\boldsymbol{d}} = g(\hat{\boldsymbol{\zeta}})$.

3. Selection of Proposal Distributions

In this section, we propose three criteria for selecting the proposal distributions $\boldsymbol{q} = \{q_1, \ldots, q_k\} \subset Q$ for efficient use of the multiple IS estimators. For $\boldsymbol{q} \subset Q$, the proposed criterion is generally denoted by $\phi(\boldsymbol{q})$ and the optimal set is obtained as follows:

$$\text{Minimize} \phi(\boldsymbol{q}) \text{over} \boldsymbol{q} \subset Q.$$

We consider the case where the set Q corresponds to a family of densities parameterized by $\xi \in \Xi$; thus, searching over Q is equivalent to searching over Ξ. The variable ξ can be multi-dimensional and the range of ξ can be infinite in every direction. Thus, for computational purposes, we may need to narrow down the potential region of the search, depending on the application. Evangelou and Roy (2019) considered the problem of maximizing (1.2) with respect to ξ. As mentioned in the Introduction, this is the situation in empirical Bayes methods, so they used Laplace approximations to identify the region in which the maximizer

may lie. Thus, using Laplace approximations, as in Evangelou and Roy (2019), we can narrow Ξ down to a search set $\tilde{\Xi}$. In Section S10 of the Supplementary Material, we demonstrate an alternative approach to choosing $\tilde{\Xi}$ using preliminary samples.

Solving the minimization problem is a research problem in its own right. We implemented two algorithms for searching over $\tilde{\Xi}$, namely, the point-swapping algorithm of Royle and Nychka (1998), and a simulated annealing algorithm. Details about these algorithms are given in Section S7 of the Supplementary Material. The point-swapping algorithm requires more iterations, in general, and so is better suited to cases in which the design criterion ϕ can be computed quickly after a swap, as is often the case for the SF method.

3.1. Space-filling approach

In this method, from among all subsets $\boldsymbol{q} = \{q_1, \ldots, q_k\}$ of Q, we choose the one that minimizes the gaps between the elements of $\boldsymbol{q}$ and the elements of Π. For $\pi \in \Pi$, and $q \in Q$, let $\Upsilon(\pi, q)$ be a suitably chosen metric. Define

$$\psi_p(\boldsymbol{q}, \pi) = \left\{ \sum_{q \in \boldsymbol{q}} \Upsilon(\pi, q)^p \right\}^{1/p}$$

as a measure of "closeness" of $\boldsymbol{q}$ to π. Note that, for $p < 0$, $\psi_p(\boldsymbol{q}, \pi) \to 0$ if π is allowed to converge to a point in $\boldsymbol{q}$. The design criterion is to choose $\boldsymbol{q}$ to minimize

$$\phi_{\mathrm{SF}}(\boldsymbol{q}) = \Psi_{p,\tilde{p}}(\boldsymbol{q}) = \left\{ \sum_{\pi \in \Pi} \psi_p(\boldsymbol{q}, \pi)^{\tilde{p}} \right\}^{1/\tilde{p}},$$

over all subsets $\boldsymbol{q}$, with $|\boldsymbol{q}| = k$. In the limit ($p \to -\infty, \tilde{p} \to \infty$), $\Psi_{p,\tilde{p}}$ is related to the minimax design. However, as Royle and Nychka (1998) show, keeping p and $\tilde{p}$ finite allows us to quickly evaluate ϕ after a swap of the point-swapping algorithm. We use $p = -30$, and $\tilde{p} = 30$ in our examples, which allows us to obtain a near-minimax SF design.

The choice of the metric $\Upsilon(\pi, q)$ is crucial. For instance, in the binomial robit model with degrees of freedom parameter ξ (see the example in Section S8 of the Supplementary Material), the family of target densities $\Pi \equiv \{\pi_\xi(x) = \nu_\xi(x)/\theta_\xi : \xi \in \Xi\}$ is indexed by the Student's t degrees of freedom parameter ξ. Here, the relevant geometry (with respect to ξ) in $\mathbb{R}$ is not Euclidean. Indeed, the degrees of freedom $\xi = 10^2$ and 10^3 are close, but $\xi = 0.5$ and $\xi = 1$ are not. Thus, the SF based on the Euclidean distance metric (SFE) may not be appropriate unless the indexing variable is a location parameter. The Euclidean distance is also sensitive to reparameterizations of the family of proposal distributions. Another choice is the information metric (Kass (1989); Rao (1982)), which measures the distance between two parametric distributions using asymptotic standard deviation units

of the best estimator. The Kullback–Leibler divergence uses the information metric (Ghosh, Delampady and Samanta (2007)), which may be difficult to implement in practice, but seems appropriate for this context. Here, we use the symmetric Kullback–Leibler divergence (SKLD), although it is not a metric, and denote the corresponding method by SFS. Thus,

$$\Upsilon(\pi, q) = \int_{\mathsf{X}} \pi(x) \log \frac{\nu(x)}{\varphi(x)} \mu(dx) - \int_{\mathsf{X}} q(x) \log \frac{\nu(x)}{\varphi(x)} \mu(dx). \tag{3.1}$$

In the special case when $\Pi \equiv \{\pi_\xi(x) = \nu_\xi(x)/c_\xi : \xi \in \Xi\}$, that is, the target family is indexed by some variable ξ and $Q = \Pi$, the SKLD between $\pi_{\xi_1}(x)$ and $\pi_{\xi_2}(x)$ is

$$\Upsilon(\xi_1, \xi_2) = \int_{\mathsf{X}} \pi_{\xi_1}(x) \log \frac{\nu_{\xi_1}(x)}{\nu_{\xi_2}(x)} \mu(dx) - \int_{\mathsf{X}} \pi_{\xi_2}(x) \log \frac{\nu_{\xi_1}(x)}{\nu_{\xi_2}(x)} \mu(dx) \tag{3.2}$$

$$= \frac{\int_{\mathsf{X}} \nu_{\xi_1}(x) \log\{\nu_{\xi_1}(x)/\nu_{\xi_2}(x)\} \mu(dx)}{\int_{\mathsf{X}} \nu_{\xi_1}(x) \mu(dx)} - \frac{\int_{\mathsf{X}} \nu_{\xi_2}(x) \log\{\nu_{\xi_1}(x)/\nu_{\xi_2}(x)\} \mu(dx)}{\int_{\mathsf{X}} \nu_{\xi_2}(x) \mu(dx)}. \tag{3.3}$$

In general, the SKLD (3.1) is not available in a closed form. We use a modified Laplace method (Evangelou, Zhu and Smith (2011)) to approximate (3.3), and describe the method in Section S1. The second-order approximation described in the Supplementary Material is exact when π_{ξ_1} and π_{ξ_2} are any two Gaussian densities. If X is discrete, or if the target distributions are far from Gaussian, we can use a Monte Carlo estimate of (3.2) with samples from π_{ξ_1} and π_{ξ_2}. Indeed, for some examples considered here, we use the Monte Carlo estimate of (3.2) to implement SFS.

The SF method does not require any particular form of IS estimator. When $Q = \Pi$, the uniform (with respect to the chosen metric) selection of the proposal distributions attempts to guarantee that each target density is close to at least one proposal distribution. In addition, the SF method is attractive, because an IS estimator is usually used to simultaneously estimate several quantities of interest, resulting in different optimal design criteria.

3.2. Minimax approach

Our second method is the minimax (MNX) design, based on minimizing the maximum SE or relative SE of $\hat{u}(\pi, \hat{\boldsymbol{d}})$ or $\hat{\eta}^{[f]}(\pi; \hat{\boldsymbol{d}})$ over $\pi \in \Pi$. The consistency and asymptotic normality of $\hat{\boldsymbol{d}}$, $\hat{u}(\pi; \hat{\boldsymbol{d}})$, and $\hat{\eta}^{[f]}(\pi; \hat{\boldsymbol{d}})$ are described in Theorems 1, 2, and 3, respectively, of Roy, Tan and Flegal (2018). Let $\sigma_u^2(\pi, \boldsymbol{q})$ denote the asymptotic variance of $\hat{u}(\pi, \hat{\boldsymbol{d}})$ when the set of proposal densities is $\boldsymbol{q} \subset Q$. Then, the standard error is $\sigma_u(\pi, \boldsymbol{q})/\sqrt{n}$, where $n = \sum_{l=1}^{k} n_l$. The minimax approach chooses $\boldsymbol{q}$ that minimizes the largest standard error or the

relative standard error, given by

$$\phi_{\mathrm{MNX}}(\boldsymbol{q}) = \max_{\pi\in\Pi} \frac{\sigma_u(\pi,\boldsymbol{q})}{\sqrt{n}}, \text{ and } \phi_{\mathrm{MNX}}(\boldsymbol{q}) = \max_{\pi\in\Pi} \frac{\sigma_u(\pi,\boldsymbol{q})}{\sqrt{n}\hat{u}(\pi,\hat{\boldsymbol{d}})},$$

respectively. Similar measures can be derived for $\hat{\eta}^{[f]}(\pi;\hat{\boldsymbol{d}})$, with variance $\sigma^2_\eta(\pi,\boldsymbol{q})$. In the following, we discuss estimating the asymptotic variances $\sigma^2_u(\pi,\boldsymbol{q})$ and $\sigma^2_\eta(\pi,\boldsymbol{q})$ of these estimators. Note that the ratios of the normalizing constants (θ/c_1) can take large values as π varies in Π, especially when X is multi-dimensional. The standard errors corresponding to distributions with large ratios tend to be larger, whereas these standard errors for distributions with small (relative) normalizing constants can potentially be large relative to the value of the estimates. Thus, if the goal is to estimate the parameters corresponding to the largest normalizing constants (as in the empirical Bayes methods; see, e.g., Roy, Evangelou and Zhu (2016)), then the first criterion can be used. On the other hand, if one wants to estimate θ for all $\pi \in \Pi$, then the second criterion (relative standard error) may be preferred.

Spectral variance estimation in reverse logistic regression and multiple IS methods: First, we provide an SV estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{d}}$, because we need it for the asymptotic variances of $\hat{u}(\pi;\hat{\boldsymbol{d}})$ and $\hat{\eta}^{[f]}(\pi;\hat{\boldsymbol{d}})$. The SV estimator of $\mathrm{Var}(\hat{\boldsymbol{d}})$ is also important in its own right, and is used in Section 3.3 in our third approach to selecting proposal distributions.

As in Roy, Tan and Flegal (2018), we assume that the Markov chains Φ_l, and $\tilde{\Phi}_l$ are *polynomially ergodic* for $l = 1,\ldots,k$. (The definition of the polynomial ergodicity of Markov chains can be found in Roy, Tan and Flegal (2018).) They showed that if the Markov chain $\tilde{\Phi}_l$ is polynomially ergodic of order $t > 1$, for $l = 1,\ldots,k$, then $\hat{\boldsymbol{\zeta}}$ and $\hat{\boldsymbol{d}}$ defined in section 2.1 are consistent and asymptotically normal as $N_1,\ldots,N_k \to \infty$; that is, there exist matrices $B,\Omega \in \mathbb{R}^{k,k}$ and $D \in \mathbb{R}^{k,k-1}$ such that

$$\sqrt{N}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \xrightarrow{d} \mathcal{N}(0,U) \quad \text{and} \quad \sqrt{N}(\hat{\boldsymbol{d}} - \boldsymbol{d}) \xrightarrow{d} \mathcal{N}(0,V),$$

where $U = B^\dagger \Omega B^\dagger$ and $V = D^\top U D$. Here, for a square matrix C, $C^\dagger$ denotes its Moore–Penrose inverse. The matrices B, Ω, and D are defined in (2.7), (2.8), and (2.5), respectively, in Roy, Tan and Flegal (2018). Theorem 1 below provides consistent SV estimators of the asymptotic variances of $\hat{\boldsymbol{\zeta}}$ and $\hat{\boldsymbol{d}}$.

We now introduce some notation. Assume $N_l \to \infty$, such that $\lim N_l/N \in (0,1)$, for $l = 1,\ldots,k$. Recall that $\hat{\boldsymbol{d}} = g(\hat{\boldsymbol{\zeta}})$, and its gradient at $\hat{\boldsymbol{\zeta}}$ (in terms of $\hat{\boldsymbol{d}}$) is

$$\widehat{D} = \begin{pmatrix} \hat{d}_2 & \hat{d}_3 & \dots & \hat{d}_k \\ -\hat{d}_2 & 0 & \dots & 0 \\ 0 & -\hat{d}_3 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -\hat{d}_k \end{pmatrix}. \tag{3.4}$$

As in Roy, Tan and Flegal (2018), the $k \times k$ matrix $\widehat{B}$ is defined by

$$\begin{aligned} \widehat{B}_{rr} &= \sum_{l=1}^{k} a_l \left[\frac{1}{N_l} \sum_{i=1}^{N_l} p_r(\tilde{X}_i^{(l)}, \hat{\boldsymbol{\zeta}}) \Big\{ 1 - p_r(\tilde{X}_i^{(l)}, \hat{\boldsymbol{\zeta}}) \Big\} \right] \text{ and} \\ \widehat{B}_{rs} &= -\sum_{l=1}^{k} a_l \left\{ \frac{1}{N_l} \sum_{i=1}^{N_l} p_r(\tilde{X}_i^{(l)}, \hat{\boldsymbol{\zeta}}) p_s(\tilde{X}_i^{(l)}, \hat{\boldsymbol{\zeta}}) \right\}, \text{ for } r \neq s; \end{aligned} \tag{3.5}$$

that is, $\widehat{B}$ denotes the matrix of the second derivatives of $-\ell_N(\boldsymbol{\zeta})/N$ evaluated at $\hat{\boldsymbol{\zeta}}$, where $\ell_N(\boldsymbol{\zeta})$ is defined in (2.3). Set $Z_i^{(l)} = (p_1(\tilde{X}_i^{(l)}, \hat{\boldsymbol{\zeta}}), \dots, p_k(\tilde{X}_i^{(l)}, \hat{\boldsymbol{\zeta}}))^\top$, for $i = 1, \dots, N_l$, and $\bar{Z}^{(l)} = \sum_{i=1}^{N_l} Z_i^{(l)}/N_l$. Define the lag j sample autocovariance as

$$\gamma_N^{(l)}(j) = \frac{1}{N_l} \sum_{i \in S_{j,N}} \left(Z_i^{(l)} - \bar{Z}^{(l)} \right) \left(Z_{i+j}^{(l)} - \bar{Z}^{(l)} \right)^\top, \text{ for } l = 1, \dots, k, \tag{3.6}$$

where $S_{j,N} = \{1, \dots, N-j\}$ for $j \geq 0$, and $S_{j,N} = \{(1-j), \dots, N\}$ for $j < 0$. Let

$$\widehat{\Sigma}^{(l)} = \sum_{j=-(b_{N_l}-1)}^{b_{N_l}-1} w_{N_l}(j) \gamma_N^{(l)}(j), \tag{3.7}$$

where $w_{N_l}(\cdot)$ is the lag window, and b_{N_l} are the truncation points for $l = 1, \dots, k$. Finally, define

$$\widehat{\Omega} = \sum_{l=1}^{k} \frac{N}{N_l} a_l^2 \widehat{\Sigma}^{(l)}. \tag{3.8}$$

Theorem 1. *Assume that the Markov chains $\tilde{\Phi}_1, \dots, \tilde{\Phi}_k$ are polynomially ergodic of order $t > 1$, and for all $l = 1, \dots, k$, w_{N_l} and b_{N_l} satisfy conditions* 1–4 *in Vats, Flegal and Jones (*2018*, Thm.* 2*). Let $\widehat{D}$, $\widehat{B}$, and $\widehat{\Omega}$ be the matrices defined by* (3.4)*,* (3.5)*, and* (3.8)*, respectively. Then, as $N_l \to \infty$, for all $l = 1, \dots, k$, $\widehat{U} := \widehat{B}^\dagger \widehat{\Omega} \widehat{B}^\dagger$ and $\widehat{V} := \widehat{D}^\top \widehat{U} \widehat{D}$ converge almost surely to U and V, respectively.*

Next, we estimate the asymptotic variances of $\hat{u}(\pi; \hat{\boldsymbol{d}})$ and $\hat{\eta}^{[f]}(\pi; \hat{\boldsymbol{d}})$. Roy, Tan and Flegal (2018) showed that, under certain conditions, there exist $\sigma_u^2, \sigma_\eta^2 > 0$ such that, as $n_1, \dots, n_k \to \infty$,

$$\sqrt{n}\{\hat{u}(\pi; \hat{\boldsymbol{d}}) - u(\pi, q_1)\} \xrightarrow{d} N(0, \sigma_u^2) \quad \text{and} \quad \sqrt{n}(\hat{\eta}^{[f]}(\pi; \hat{\boldsymbol{d}}) - E_\pi f) \xrightarrow{d} N(0, \sigma_\eta^2). \tag{3.9}$$

In Theorem 2, we provide consistent SV estimators of σ_u^2 and σ_η^2. We first introduce some notation. Let

$$u^\pi(x;\boldsymbol{d}) := \frac{\nu(x)}{\sum_{s=1}^k a_s\varphi_s(x)/d_s} \quad \text{and} \quad v^{[f],\pi}(x;\boldsymbol{d}) := f(x)u^\pi(x;\boldsymbol{d}). \tag{3.10}$$

Define the vectors $c(\pi;\boldsymbol{d})$ and $e(\pi;\boldsymbol{d})$ of length $k-1$ and with $(j-1)$th coordinate as

$$[c(\pi;\boldsymbol{d})]_{j-1} = \frac{u(\pi,q_1)}{d_j^2}\int_{\mathsf{X}} \frac{a_j\varphi_j(x)}{\sum_{s=1}^k a_s\varphi_s(x)/d_s}\pi(x)\mu(dx) \tag{3.11}$$

$$[e(\pi;\boldsymbol{d})]_{j-1} = \frac{a_j}{d_j^2}\int_{\mathsf{X}} \frac{[f(x)-E_\pi f]\varphi_j(x)}{\sum_{s=1}^k a_s\varphi_s(x)/d_s}\pi(x)\mu(dx), \tag{3.12}$$

respectively, for $j=2,\ldots,k$, and their estimators $\hat{c}(\pi;\boldsymbol{d})$ and $\hat{e}(\pi;\boldsymbol{d})$ as

$$[\hat{c}(\pi;\boldsymbol{d})]_{j-1} = \sum_{l=1}^k \frac{1}{n_l}\sum_{i=1}^{n_l} \frac{a_j a_l \nu(X_i^{(l)})\varphi_j(X_i^{(l)})}{\{\sum_{s=1}^k a_s\varphi_s(X_i^{(l)})/d_s\}^2 d_j^2}, \tag{3.13}$$

$$\begin{aligned}&[\hat{e}(\pi;\boldsymbol{d})]_{j-1}\\ &= \frac{\sum_{l=1}^k (a_l/n_l)\sum_{i=1}^{n_l} a_j f(X_i^{(l)})\nu(X_i^{(l)})\varphi_j(X_i^{(l)})/[d_j^2\{\sum_{s=1}^k a_s\varphi_s(X_i^{(l)})/d_s\}^2]}{\hat{u}(\pi;\boldsymbol{d})}\\ &\quad - \frac{[\hat{c}(\pi;\boldsymbol{d})]_{j-1}\hat{\eta}^{[f]}(\pi;\boldsymbol{d})}{\hat{u}(\pi;\boldsymbol{d})},\end{aligned} \tag{3.14}$$

respectively. Suppose b_{n_l} are the truncation points, $w_{n_l}(j)$ is the lag window, $u_i \equiv u_i(\boldsymbol{d}) \equiv u^\pi(X_i^{(l)};\boldsymbol{d})$, $v_i^{[f]} \equiv v_i^{[f]}(\boldsymbol{d}) \equiv v^{[f],\pi}(X_i^{(l)};\boldsymbol{d})$, and $\bar{u} \equiv \bar{u}(\boldsymbol{d})$, and $\bar{v}^{[f]} \equiv \bar{v}^{[f]}(\boldsymbol{d})$ are the averages of $\{u^\pi(X_1^{(l)};\boldsymbol{d}),\ldots,u^\pi(X_{n_l}^{(l)};\boldsymbol{d})\}$ and $\{v^{[f],\pi}(X_1^{(l)};\boldsymbol{d}),\ldots,v^{[f],\pi}(X_{n_l}^{(l)};\boldsymbol{d})\}$, respectively. (Note that, with a slight abuse of the notation, the dependence on l is ignored in $u_i, v_i^{[f]}, \bar{u}$, and $\bar{v}^{[f]}$.) Let

$$\hat{\tau}_l^2(\pi;\boldsymbol{d}) = \frac{1}{n_l}\sum_{j=-(b_{n_l}-1)}^{b_{n_l}-1} w_{n_l}(j)\sum_{i\in S_{j,n}}(u_i-\bar{u})(u_{i+j}-\bar{u}), \text{ and} \tag{3.15}$$

$$\widehat{\Gamma}_l(\pi;\boldsymbol{d}) = \frac{1}{n_l}\sum_{j=-(b_{n_l}-1)}^{b_{n_l}-1} w_{n_l}(j)\sum_{i\in S_{j,n}}\left\{\begin{pmatrix} v_i^{[f]}\\ u_i\end{pmatrix}-\begin{pmatrix}\bar{v}^{[f]}\\ \bar{u}\end{pmatrix}\right\}\left\{\begin{pmatrix} v_{i+j}^{[f]}\\ u_{i+j}\end{pmatrix}-\begin{pmatrix}\bar{v}^{[f]}\\ \bar{u}\end{pmatrix}\right\}^\top.$$

Finally, let $\hat{\tau}^2(\pi;\boldsymbol{d}) = \sum_{l=1}^k (a_l^2 n/n_l)\hat{\tau}_l^2(\pi;\boldsymbol{d})$, $\widehat{\Gamma}(\pi;\boldsymbol{d}) = \sum_{l=1}^k (a_l^2 n/n_l)\widehat{\Gamma}_l(\pi;\boldsymbol{d})$, and

$$\hat{\rho}(\pi;\hat{\boldsymbol{d}}) = \nabla h\{\hat{v}^{[f]}(\pi;\hat{\boldsymbol{d}}),\hat{u}(\hat{\boldsymbol{d}})\}^\top \widehat{\Gamma}(\pi;\hat{\boldsymbol{d}})\nabla h\{\hat{v}^{[f]}(\pi;\hat{\boldsymbol{d}}),\hat{u}(\hat{\boldsymbol{d}})\},$$

where $\nabla h(x,y) = (1/y, -x/y^2)^\top$.

Theorem 2. *Suppose that for $\tilde{\Phi}_l$, for $l=1,\ldots,k$, the conditions of Theorem 1 hold and $\widehat{V}$ is the consistent SV estimator of V. Suppose that $N_l, n_l \to \infty$, for*

all $l = 1, \ldots, k$, and there exists $\varpi \in [0, \infty)$ such that $n/N \to \varpi$. In addition, let $n_l/n \to s_l \in (0, 1)$, for $l = 1, \ldots, k$. Assume that the Markov chains $\Phi_1, \ldots, \Phi_k$ are polynomially ergodic of order $t \geq (1+\epsilon)(1+2/\delta)$, for some $\epsilon, \delta > 0$, such that $E_{q_l}|u^\pi(X; \boldsymbol{d})|^{4+\delta} < \infty$, and for each $l = 1, \ldots, k$, w_{n_l} and b_{n_l} satisfy conditions 1–4 *in Vats, Flegal and Jones (*2018*, Thm.* 2*).*

(a) Then, $\hat{\sigma}_u^2 = (n/N)\hat{c}(\pi; \hat{\boldsymbol{d}})^\top \widehat{V} \hat{c}(\pi; \hat{\boldsymbol{d}}) + \hat{\tau}^2(\pi; \hat{\boldsymbol{d}})$ converges almost surely to σ_u^2.

(b) In addition, suppose that $E_{q_l}|v^{[f],\pi}(X; \boldsymbol{d})|^{4+\delta} < \infty$. Then, $\hat{\sigma}_\eta^2 = (n/N)$ $\hat{e}(\pi; \hat{\boldsymbol{d}})^\top \widehat{V} \hat{e}(\pi; \hat{\boldsymbol{d}}) + \hat{\rho}(\pi; \hat{\boldsymbol{d}})$ converges almost surely to σ_η^2.

The estimators $\widehat{V}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_\eta^2$ are implemented using the R package geoBayes (Evangelou and Roy (2022)). Because we obtain the samples by running the Markov chains with stationary densities in $\boldsymbol{q}$, we denote the corresponding reverse logistic regression estimator of $\boldsymbol{d} \equiv \boldsymbol{d}_{\boldsymbol{q}}$ by $\hat{\boldsymbol{d}}_{\boldsymbol{q}}$ and its asymptotic variance as $V_{\boldsymbol{q}}$. Similarly, in this case, we denote the SV estimators of the asymptotic variances (3.9) of $\hat{u}(\pi; \hat{\boldsymbol{d}}_{\boldsymbol{q}})$ and $\hat{\eta}^{[f]}(\pi; \hat{\boldsymbol{d}}_{\boldsymbol{q}})$ as $\hat{\sigma}_u^2(\pi; \boldsymbol{q})$ and $\hat{\sigma}_\eta^2(\pi; \boldsymbol{q})$, respectively.

When $Q = \Pi$, a less computationally demanding approach is the SEQ method, in which the densities are chosen sequentially from Π, where $\hat{\sigma}_u^2(\pi; \boldsymbol{q})$ is the largest. Specifically, starting with an initial density $\boldsymbol{q}_1 = \{\tilde{q}\}$, suppose that we have completed the ith step, with the set $\boldsymbol{q}_i$ chosen, along with (Markov chain) samples from each density in $\boldsymbol{q}_i$. If $\boldsymbol{d}$ is unknown, part of this sample (stage 1) is used to calculate the estimator $\hat{\boldsymbol{d}}$, and the remaining sample is used to compute $\hat{\sigma}_u^2(\pi; \boldsymbol{q}_i)$ for the remaining densities $\pi \in \Pi \setminus \boldsymbol{q}_i$. Then, $\boldsymbol{q}_{i+1} = \boldsymbol{q}_i \cup \{\pi_j\}$, where $\pi_j = \operatorname{argmax}_{\pi \in \Pi \setminus \boldsymbol{q}_i} \hat{\sigma}_u^2(\pi; \boldsymbol{q}_i)$, and the existing (Markov chain) sample is augmented with samples from π_j. Thus, at each step, we choose the density corresponding to the largest (estimated) asymptotic variance. The process is repeated until we have selected k densities. The initial $\tilde{q}$ can be the density that maximizes the multiple IS estimator (1.2) or any other interesting quantity based on samples from a preliminary SF set (see Section S10 of the Supplementary Material for an example).

3.3. Maximum entropy approach

The third method uses maximum entropy sampling (Shewry and Wynn (1987)) to select $\boldsymbol{q}$. This method is applicable when $\boldsymbol{d}$ is unknown, and is developed in the context of Doss's (2010) two-stage IS estimation scheme. We use the notation $\mathrm{Ent}(\cdot)$ to denote the Boltzmann–Shannon entropy of the random variable inside the brackets. The maximum entropy (ENT) approach chooses $\boldsymbol{q}$ that minimizes

$$\phi_{\mathrm{ENT}}(\boldsymbol{q}) = -\mathrm{Ent}(\hat{\boldsymbol{d}}_{\boldsymbol{q}}).$$

This is interpreted as sampling those elements of Q that carry the most uncertainty in $\hat{\boldsymbol{d}}_{\boldsymbol{q}}$. As shown below, because we use $\hat{\boldsymbol{d}}_{\boldsymbol{q}}$ to calculate both $\hat{u}$ and $\hat{\eta}^{[f]}$, the optimal $\boldsymbol{q}$ will result in (asymptotically) lower uncertainty in those

estimators. Note that because $\hat{\boldsymbol{d}}_{\boldsymbol{q}}$ depends on the reference density q_1, we assume that q_1 remains fixed, which can be the density $\tilde{q}$ discussed in Section 3.2. In the following, we assume that the objective is to estimate ratios of normalizing constants. In the Supplementary Material, we derive similar results under the objective of estimating the means $E_\pi f$.

To derive a formula for $\mathrm{Ent}(\hat{\boldsymbol{d}}_{\boldsymbol{q}})$, we require the asymptotic joint distribution of $\hat{\boldsymbol{d}}_{\boldsymbol{q}}$ with $\hat{u}$ over Π. Let $\hat{\mathbf{u}}(\boldsymbol{\pi}; \hat{\boldsymbol{d}}_{\boldsymbol{q}})$ be the vector of length $|\Pi|$ consisting of $\hat{u}(\pi; \hat{\boldsymbol{d}}_{\boldsymbol{q}})$, for $\pi \in \Pi$ in any fixed order. Indeed, we refer to this fixed ordering whenever we write Π in this section. Similarly, define the vector of true (ratios of) normalizing constants $\mathbf{u}(\boldsymbol{\pi}, q_1)$. Let $C(\boldsymbol{\pi}; \boldsymbol{d}_{\boldsymbol{q}})$ be the $|\Pi| \times (k-1)$ matrix with rows $c(\pi; \boldsymbol{d}_{\boldsymbol{q}})$ (defined in (3.11)), for $\pi \in \Pi$. Similarly, define $\widehat{C}(\boldsymbol{\pi}; \boldsymbol{d}_{\boldsymbol{q}})$ with rows $\hat{c}(\pi; \boldsymbol{d}_{\boldsymbol{q}})$ (defined in (3.13)), for $\pi \in \Pi$. Let $\mathbf{u}^{\boldsymbol{\pi}}(x; \boldsymbol{d}_{\boldsymbol{q}})$ be the $|\Pi|$-dimensional vector consisting of $u^\pi(x; \boldsymbol{d}_{\boldsymbol{q}})$ defined in (3.10). Let $T_l(\boldsymbol{d}_{\boldsymbol{q}})$ be the $|\Pi| \times |\Pi|$ matrix with elements

$$\begin{aligned} \tau_l^2(\pi, \pi'; \boldsymbol{d}_{\boldsymbol{q}}) = {} & \mathrm{Cov}_{q_l}\{u^\pi(X_1^{(l)}; \boldsymbol{d}_{\boldsymbol{q}}), u^{\pi'}(X_1^{(l)}; \boldsymbol{d}_{\boldsymbol{q}})\} \\ & + \sum_{g=1}^{\infty} \mathrm{Cov}_{q_l}\{u^\pi(X_1^{(l)}; \boldsymbol{d}_{\boldsymbol{q}}), u^{\pi'}(X_{1+g}^{(l)}; \boldsymbol{d}_{\boldsymbol{q}})\} \\ & + \sum_{g=1}^{\infty} \mathrm{Cov}_{q_l}\{u^\pi(X_{1+g}^{(l)}; \boldsymbol{d}), u^{\pi'}(X_1^{(l)}; \boldsymbol{d}_{\boldsymbol{q}})\}. \end{aligned} \tag{3.16}$$

Finally, let

$$\widehat{T}_l(\boldsymbol{d}_{\boldsymbol{q}}) = \frac{1}{n_l} \sum_{j=-(b_{n_l}-1)}^{b_{n_l}-1} w_{n_l}(j) \sum_{i \in S_{j,n}} \left\{ \mathbf{u}^{\boldsymbol{\pi}}(X_i^{(l)}; \boldsymbol{d}_{\boldsymbol{q}}) - \bar{\mathbf{u}}(\boldsymbol{d}_{\boldsymbol{q}}) \right\} \left\{ \mathbf{u}^{\boldsymbol{\pi}}(X_{i+j}^{(l)}; \boldsymbol{d}_{\boldsymbol{q}}) - \bar{\mathbf{u}}(\boldsymbol{d}_{\boldsymbol{q}}) \right\}^\top, \tag{3.17}$$

where b_{n_l} are the truncation points, $w_{n_l}(j)$ are the lag windows, and $\bar{\mathbf{u}}(\boldsymbol{d}_{\boldsymbol{q}}) = \sum_{i=1}^{n_l} \mathbf{u}^{\boldsymbol{\pi}}(X_i^{(l)}; \boldsymbol{d}_{\boldsymbol{q}})/n_l$.

Theorem 3. *Suppose that $N_l, n_l \to \infty$, for all $l = 1, \ldots, k$, and there exists $\varpi \in [0, \infty)$ such that $n/N \to \varpi$. In addition, let $n_l/n \to s_l \in (0,1)$, for $l = 1, \ldots, k$.*

(a) Assume that the stage-1 Markov chains $\tilde{\Phi}_l$, for $l = 1, \ldots, k$, are polynomially ergodic of order $t > 1$. Furthermore, assume that the stage-2 Markov chains Φ_l, for $l = 1, \ldots, k$, are polynomially ergodic of order t, and for some $\delta > 0$, $E_{q_l}|u^\pi(X; \boldsymbol{d}_{\boldsymbol{q}})|^{2+\delta} < \infty$, for each $\pi \in \Pi$ and $l = 1, \ldots, k$, where $t > 1 + 2/\delta$. Then, as $n_1, \ldots, n_k \to \infty$,

$$\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{d}}_{\boldsymbol{q}} - \boldsymbol{d}_{\boldsymbol{q}} \\ \hat{\mathbf{u}}(\boldsymbol{\pi}; \hat{\boldsymbol{d}}_{\boldsymbol{q}}) - \mathbf{u}(\boldsymbol{\pi}, q_1) \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \varpi V_{\boldsymbol{q}} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right), \tag{3.18}$$

where $\Sigma_{21} = \varpi C(\boldsymbol{\pi}; \boldsymbol{d}_{\boldsymbol{q}}) V_{\boldsymbol{q}}$, $\Sigma_{12} = \Sigma_{21}^\top$, and $\Sigma_{22} = \varpi C(\boldsymbol{\pi}; \boldsymbol{d}_{\boldsymbol{q}}) V_{\boldsymbol{q}} C(\boldsymbol{\pi}; \boldsymbol{d}_{\boldsymbol{q}})^\top + \sum_{l=1}^k (a_l^2/s_l) T_l(\boldsymbol{d}_{\boldsymbol{q}})$.

(b) Suppose that the conditions of Theorem 1 *hold for the stage-*1 *Markov chains. Let* $\widehat{V}_{\boldsymbol{q}}$ *be the consistent estimator of* $V_{\boldsymbol{q}}$ *given in Theorem* 1*. Assume that the Markov chains* Φ_l*, for* $l = 1, \ldots, k$ *are polynomially ergodic of order* $t \geq (1+\epsilon)(1+2/\delta)$*, for some* $\epsilon, \delta > 0$*, such that* $E_{q_l}\|\mathbf{u}^{\boldsymbol{\pi}}(X; \boldsymbol{d}_{\boldsymbol{q}})\|^{4+\delta} < \infty$*, (*$\|\cdot\|$ *denotes the Euclidean norm) for all* $l = 1, \ldots, k$*, and* w_{n_l} *and* b_{n_l} *satisfy conditions* 1–4 *in Vats, Flegal and Jones (*2018*, Thm.* 2*). Then,* $(n/N)\widehat{C}(\boldsymbol{\pi}; \hat{\boldsymbol{d}}_{\boldsymbol{q}})\widehat{V}_{\boldsymbol{q}}\widehat{C}(\boldsymbol{\pi}; \hat{\boldsymbol{d}}_{\boldsymbol{q}})^\top + \sum_{l=1}^{k}(a_l^2/s_l)\widehat{T}_l(\hat{\boldsymbol{d}}_{\boldsymbol{q}})$ *and* $(n/N)\widehat{C}(\boldsymbol{\pi}; \hat{\boldsymbol{d}}_{\boldsymbol{q}})\widehat{V}_{\boldsymbol{q}}$ *converge almost surely to* Σ_{22} *and* Σ_{21}*, respectively.*

Let $Y \equiv (Y_{\boldsymbol{q}}^T, Y_\Pi^T)^T$ be a random vector following the normal distribution in (3.18). The Boltzmann–Shannon entropy of Y is $\mathrm{Ent}(Y) = \text{constant} + (1/2)\log\det(\Sigma)$, where Σ is the covariance matrix of Y. Note that

$$\log\det(\Sigma) = \log\det(\varpi V_{\boldsymbol{q}}) + \log\det\{\Sigma_{22} - \varpi C(\boldsymbol{\pi}; \boldsymbol{d}_{\boldsymbol{q}})V_{\boldsymbol{q}}C(\boldsymbol{\pi}; \boldsymbol{d}_{\boldsymbol{q}})^\top\},$$

where the second matrix on the right side is the covariance matrix of the conditional distribution of $Y_\Pi|Y_{\boldsymbol{q}}$. Because Theorem 3 (b) provides a consistent estimator of this conditional covariance matrix, we can minimize the determinant of this estimator matrix to choose $\boldsymbol{q}$.

As mentioned in Shewry and Wynn (1987), converting this conditional problem to an unconditional problem yields a significant computational benefit. In particular, as noted in Shewry and Wynn (1987), minimizing the second term is equivalent to maximizing $\log\det(V_{\boldsymbol{q}})$. In practice, we replace $V_{\boldsymbol{q}}$ with its estimator given in Theorem 1, namely, $\widehat{V}_{\boldsymbol{q}}$, using Markov chain samples from densities in $\boldsymbol{q}$. In this case, the ENT criterion simplifies to

$$\phi_{\mathrm{ENT}}(\boldsymbol{q}) = -\log\det(\widehat{V}_{\boldsymbol{q}}).$$

Unlike the SF, MNX, and SEQ methods, the ENT approach is applicable only in the context of Doss's (2010) two-stage IS estimation scheme. In contrast, if we use the multiple IS estimator (1.2), because ENT avoids the second-stage IS estimation, it needs fewer samples than the MNX and SEQ methods do, which require enough samples to be used for both stages. In addition, ENT avoids having to compute the target un-normalized densities ν for $\pi \in \Pi$. However, one advantage of the MNX and SEQ methods is that, at the end of the procedure, we have samples from densities in $\boldsymbol{q}$ that can be used in the two-stage IS estimation scheme.

4. Examples

Autologistic model: Consider the popular autologistic models (Besag (1974)), which are Markov random field models for binary observations. Let s_i denote the ith spatial location, and let $\mathrm{nb}_i \equiv \{s_j : s_j \text{ is a neighbor of } s_i\}$ denote the neighborhood set of s_i, for $i = 1, \ldots, m$. Markov random field models for $\boldsymbol{x} =$

$\{x(s_i), i = 1, \ldots, m\}$ are formulated by specifying the conditional probabilities $p_i = P(x(s_i) = 1|\{x(s_j) : j \neq i\}) = P(x(s_i) = 1|\{x(s_j) : s_j \in \text{nb}_i\})$, for $i = 1, \ldots, m$. For simplicity, we impose that all neighborhoods have the same size $w = |\text{nb}_i|$, for $i = 1, \ldots, m$. We consider a centered parameterization (Kaiser, Caragea and Furukawa (2012)) given by $\text{logit}(p_i) = \text{logit}(\kappa) + (\gamma/w)\sum_{s_j \in \text{nb}_i}\{x(s_j) - \kappa\}$, where $\text{logit}(z) = \log\{z/(1-z)\}$, γ is a dependence parameter, and κ is the probability of observing one in the absence of statistical dependence. Jointly, the probability mass function (pmf) $\pi(\boldsymbol{x}|\gamma, \kappa)$ of $\boldsymbol{x}$ is given by (see Section S9.1 of the Supplementary Material)

$$\pi(\boldsymbol{x}|\gamma, \kappa) \propto \exp\left[\{\text{logit}(\kappa) - \gamma\kappa\}\sum_{i=1}^{m} x(s_i) + \frac{\gamma}{2w}\sum_{i=1}^{m}\sum_{s_j \in \text{nb}_i} x(s_i)x(s_j)\right]. \quad (4.1)$$

The normalizing constant $\theta \equiv \theta(\gamma, \kappa)$ in $\pi(\boldsymbol{x}|\gamma, \kappa)$ is intractable when $\gamma \neq 0$. Sherman, Apanasovich and Carroll (2006) mention that "there is no known simple way to approximate this normalizing constant". Here, we use multiple IS to estimate θ, and then estimate $\xi = (\gamma, \kappa)$ using the maximum likelihood method.

We consider a 10×10 square lattice on a torus, with a four-nearest (east–west, north–south) neighborhood structure, with the family of autologistic pmfs $\Pi = \{\pi(x|\gamma, \kappa) : \gamma = -4, -3.2, \ldots, 4, \kappa = 0.1, 0.2, \ldots, 0.9\}$. In this case, the family of importance densities $Q = \Pi$. Therefore, choosing the importance densities amounts to choosing the parameters ξ. We choose $k = 5$ densities from Q, that is, k different ξ values, one of which must be $\xi_1 = (0, 0.5)$. We apply the multiple IS using the proposal densities from the five methods, namely, SFE, SFS, MNX, SEQ, and ENT, as well as the naive IS method, NIS. MNX and SEQ are based on the relative standard error criterion. The computations of the SFS, MNX, SEQ, and ENT criteria are based on 20,000 stage-1 and 20,000 stage-2 samples, produced from each candidate density using Gibbs sampling (except for $\gamma = 0$, where independent sampling was used), after a burn-in of 4,000 samples each time. We use the Tukey-Hanning window to compute the SV estimator (see Section S9.2). We observe that SEQ chooses a skeleton set on the boundary of the search space for (γ, κ), and that SFS, MNX, and ENT choose some points close to the boundary (see Section S9.2).

To test the performance of the different methods when used to estimate the parameters ξ, we simulate from the model for different choices of ξ, as shown in Table 1, and then estimate these parameters using the maximum likelihood method. Because the likelihood is intractable, $\theta(\gamma, \kappa)/\theta(0, 0.5)$ is estimated using (1.2) with the proposal densities derived from each method. To that end, we took 10,000 samples from each density after a burn-in of 1,000 samples. For NIS, we took 50,000 samples. We generated 125 realizations (data) for each choice of (γ, κ) parameters. Some realized data resulted in an unbounded likelihood for some methods. NIS was most affected, with 39% of the realized values resulting in

Table 1. Root mean squared error for estimating γ in the autologistic example.

κ	γ	NIS	SFE	SFS	MNX	SEQ	ENT
0.2	−1	7.55	3.68	4.14	4.62	5.19	3.65
0.2	1	10.91	3.63	1.67	1.67	1.74	1.69
0.3	−2	8.75	1.38	1.61	1.60	9.42	1.37
0.3	2	5.13	1.17	1.19	1.18	1.21	1.18
0.4	−3	4.51	5.36	1.59	1.52	9.55	1.63
0.4	3	3.76	1.11	1.12	1.11	1.18	1.11
0.5	−4	10.69	5.65	1.20	1.15	10.13	3.61
0.5	4	4.83	1.04	1.04	1.03	1.06	1.02
0.6	−3	6.71	1.33	1.21	1.22	6.16	7.59
0.6	3	3.65	1.12	1.12	1.12	1.22	1.12
0.7	−2	9.62	1.62	1.93	1.79	1.80	5.97
0.7	2	6.09	1.27	1.27	1.26	1.35	1.48
0.8	−1	14.84	5.37	4.52	3.64	4.38	5.88
0.8	1	11.88	2.14	1.96	1.94	2.06	2.40

an unbounded likelihood, followed by SEQ with 11%, and ENT with 8%. Table 1 shows the root mean squared error for estimating γ, excluding the cases with unbounded likelihoods for each method. The results show that the multiple IS methods perform significantly better than NIS. For the multiple IS methods, SEQ performs worse than MNX, in general, and SFE performs worse than SFS. The root mean squared error for estimating κ does not show significant differences between the multiple IS methods, and so is not reported here, however NIS performed worse than the multiple IS methods did. Further comparisons and computational details are given in Section S9.2 of the Supplementary Material.

Bayesian negative binomial regression: We consider a Bayesian negative binomial regression model with response variable y_i, for $i = 1, \ldots, 21$, generated independently from a negative binomial distribution with size parameter ξ and mean for y_i, $\mu_i = \exp(\beta_0 + \beta_1 \times w_i)$, where $w_i = -1 + 0.1 \times (i-1)$. Here, $x = (\beta_0, \beta_1)$ are unknown parameters, assigned a bivariate normal prior with mean zero and covariance matrix $10(W^\top W)^{-1}$, where W denotes the design matrix. As $\xi \to \infty$, the negative binomial distribution converges to the Poisson distribution. Let the family of target densities Π be the posterior densities for x for $\xi \in (0, \infty]$. Here, $\xi = \infty$ corresponds to the Poisson model. We wish to compute the logarithm of the Bayes factor $b_\xi = \log(\theta_\xi / \theta_\infty)$, where θ_ξ denotes the unknown normalizing constant of the posterior density. The Bayes factor can be used to decide between the models for given data. We estimate b_ξ by multiple IS using (1.2), with the proposal densities chosen from Π, that is, $Q = \Pi$, one of which must correspond to $\xi = \infty$, and two more densities chosen from $\tilde{\Xi} = \{1, 2, \ldots, 40\}$; that is, $k = 3$. The choice of the proposal densities for MNX and SEQ are based on the relative

Table 2. Average root mean squared difference between the estimates obtained by IS and the values obtained using numerical integration for b_ξ. The table shows the original values multiplied by 100.

	0.5	1	2	∞
NIS	1,214.640	716.045	383.153	129.079
SFE	2.916	2.698	2.080	2.172
SFS	2.337	2.343	1.712	1.850
MNX	2.222	2.161	1.594	1.806
SEQ	2.293	2.307	1.745	1.810
ENT	2.266	2.140	1.626	1.774

standard error of the multiple IS estimator of $\exp(b_\xi)$. For comparison, we also consider the NIS method with a proposal at $\xi = \infty$.

We generate data from four models, with $\xi = 0.5, 1, 2, \infty$ and $(\beta_0, \beta_1) = (1, 0.5)$, 400 times from each model. For each data set, we compute the skeleton set for the five criteria: SFE, SFS, MNX, SEQ, and ENT. We use $N_l = n_l = 3{,}600$ Monte Carlo samples from the lth proposal, after a burn-in of 1,000 samples, for $l = 1, 2, 3$, to compute the spectral variance estimates, and compute the SKLD using the same samples. The Monte Carlo algorithm is implemented using the R package rstan (Stan Development Team (2020)). After determining the skeleton set for each method and data set, we generate an additional 5,000 Monte Carlo samples from each proposal, discard the first 1,000, and use the remaining 4,000 to compute the estimator of b_ξ, for all $\xi \in \tilde{\Xi}$, using (1.2). For NIS, we use 12,000 samples in total from the proposal density. Alternatively, θ_ξ can be computed by numerical integration. For this, we use the Gauss–Kronrod method, as implemented in the R package pracma (Borchers (2021)), with relative error set to 10^{-6}, from which we compute b_ξ. We treat the estimates obtained by numerical integration as the golden standard, against which we compare each IS estimate. Because the models are very similar for large values of ξ, our comparison concentrates on the range $\xi = 1, \ldots, 10$. Table 2 shows the average root mean squared difference between the IS estimate of b_ξ for each method and that obtained using numerical integration for the 400 simulations and over $\xi = 1, \ldots, 10$. The results show that MNX and ENT outperform SEQ, in general, when estimating both the Bayes factor and the regression coefficient, and that SFS outperforms SFE. NIS performs significantly worse than the multiple IS methods.

5. Conclusion

We have considered situations in which we are simultaneously interested in a large number of target distributions, for example, in model selection and sensitivity analyses. Multiple IS estimators are particularly useful in this context.

However, few studies have examined how to choose proposal distributions for these estimators. We provide three systematic techniques to address this issue. The first method, based on a geometric space-filling criterion, and the second method, based on the minimax asymptotic standard error, can be used for any multiple IS estimators. The third method, the maximum entropy method, is designed for the two-stage multiple IS estimators of Doss (2010). We compare the performance of these three methods in several examples. Our results show that a careful choice of the proposal densities, as produced by our methods, results in estimates that are more accurate.

The proposed minimax and entropy methods use asymptotic standard errors for the multiple IS and reverse logistic regression estimators, respectively. We construct consistent SV estimators for these standard errors. These estimators are important in their own right, because they are valuable for assessing the quality of the multiple IS estimators and the reverse logistic regression estimator.

Supplementary Material

The online Supplementary Material contains proofs of Theorems 1–3, and a theorem (and its proof) on entropy decomposition for multiple IS estimators of means. Here, we also describe the point swapping algorithm and the simulated annealing algorithm used to find the optimal skeleton sets. We also provide details on the computation and derivation of the pmf for the autologistic model and the modified Laplace approximation for the SKLD. Furthermore, we present two real-data examples, one involving a binomial robit model, and one involving a spatial generalized linear mixed model. For the binomial robit model, we also demonstrate the case where the family of proposals Q corresponds to a multivariate normal family.

Acknowledgments

The authors thank an editor, an anonymous associate editor, and two anonymous reviewers for several helpful comments. The authors are particularly grateful to the associate editor for suggestions to improve the exposition of the article.

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Royal Statist. Soc.,* Ser. B **36**, 192–225.

Borchers, H. W. (2021). *pracma: Practical numerical math functions.* (R package version 2.3.3).

Buta, E. and Doss, H. (2011). Computational approaches for empirical Bayes methods and Bayesian sensitivity analysis. *Ann. Statist.* **39**, 2658–2685.

Cappé, O., Guillin, A., Marin, J. M. and Robert, C. P. (2004). Population Monte Carlo. *J. Comp. and Graph. Statist.* **13**, 907–929.

Christensen, O. F. (2004). Monte Carlo maximum likelihood in model-based geostatistics. *J. Comp. and Graph. Statist.* **13**, 702–718.

Doss, H. (2010). Estimation of large families of Bayes factors from Markov chain output. *Statist. Sinica* **20**, 537–560.

Doss, H. and Tan, A. (2014). Estimates and standard errors for ratios of normalizing constants. *J. Royal Statist. Soc.,* Ser. B **76**, 683–712.

Elvira, V., Martino, L., Luengo, D. and Bugallo, M. F. (2019). Generalized multiple importance sampling. *Statist. Sci.* **34**, 129–155.

Evangelou, E. and Roy, V. (2019). Estimation and prediction for spatial generalized linear mixed models with parametric links via reparameterized importance sampling. *Spatial Statist.* **29**, 289–315.

Evangelou, E. and Roy, V. (2022). *geoBayes: Analysis of geostatistical data using Bayes and empirical Bayes methods.* (R package version 0.7.1).

Evangelou, E., Zhu, Z. and Smith, R. L. (2011). Estimation and prediction for spatial generalized linear mixed models using high order Laplace approximation. *J. Statist. Plan. and Infer.* **141**, 3564–3577.

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.* **38**, 1034–1070.

George, C. P. and Doss, H. (2018). Principled selection of hyperparameters in the latent Dirichlet allocation model. *J. Machine Learn. Res.* **18**, 1–38.

Geyer, C. J. (1994). *Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo.* Technical Report. School of Statistics, University of Minnesota, Minneapolis.

Geyer, C. J. (2011). *Handbook of Markov Chain Monte Carlo.* Chapter Importance Sampling, Simulated Tempering, and Umbrella Sampling, 295–311. CRC Press, Boca Raton.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Royal 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.

Ghosh, J. K., Delampady, M. and Samanta, T. (2007). *An Introduction to Bayesian Analysis: Theory and Methods.* Springer Science & Business Media.

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16**, 1069–1112.

Kaiser, M. S., Caragea, P. C. and Furukawa, K. (2012). Centered parameterizations and dependence limitations in Markov random field models. *J. Statist. Plan. and Infer.* **142**, 1855–1863.

Kass, R. E. (1989). The geometry of asymptotic inference. *Statist. Sci.* **4**, 188–219.

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D. and Tan, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussion). *J. Royal Statist. Soc.,* Ser. B **65**, 585–618.

Li, W., Tan, Z. and Chen, R. (2013). Two-stage importance sampling with mixture proposals. *J. Amer. Statist. Assoc.* **108**, 1350–1365.

Marinari, E. and Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhy. Lett.* **19**, 451–458.

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.

Owen, A. and Zhou, Y. (2000). Safe and effective importance sampling. *J. Amer. Statist. Assoc.* **95**, 135–143.

Rao, C. R. (1982). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhyā: The Ind. J. Statist., Ser. A*, 1–22.

Roy, V. and Chakraborty, S. (2017). Selection of tuning parameters, solution paths and standard errors for Bayesian Lassos. *Bayes. Anal.* **12**, 753–778.

Roy, V., Evangelou, E. and Zhu, Z. (2016). Efficient estimation and prediction for the Bayesian binary spatial model with flexible link functions. *Biometrics* **72**, 289–298.

Roy, V., Tan, A. and Flegal, J. (2018). Estimating standard errors for importance sampling estimators with multiple Markov chains. *Statist. Sinica* **28**, 1079–1101.

Royle, J. A. and Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Computers & Geosciences* **24**, 479–488.

Sherman, M., Apanasovich, T. V. and Carroll, R. J. (2006). On estimation in binary autologistic spatial models. *J. Statist. Comp. and Simu.* **76**, 167–179.

Shewry, M. C. and Wynn, H. P. (1987). Maximum entropy sampling. *J. Appl. Statist.* **14**, 165–170.

Stan Development Team (2020). *RStan: The R interface to Stan.* R package version 2.21.2.

Tan, A., Doss, H. and Hobert, J. P. (2015). Honest importance sampling with multiple Markov chains. *J. Comp. and Graph. Statist.* **24**, 792–826.

Tan, Z. (2004). On a likelihood approach for Monte Carlo integration. *J. Amer. Statist. Assoc.* **99**, 1027–1036.

Vardi, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13**, 178–203.

Vats, D., Flegal, J. M. and Jones, G. L. (2018). Strong consistency of the multivariate spectral variance estimator in Markov chain Monte Carlo. *Bernoulli* **24**, 1860–1909.

Veach, E. and Guibas, L. (1995). Optimally combining sampling techniques for Monte Carlo rendering. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, 419–428. Association for Computing Machinery, New York.

Vivekananda Roy

Department of Statistics, Iowa State University, Ames, IA 50011, USA.

E-mail: vroy@iastate.edu

Evangelos Evangelou

Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK.

E-mail: ee224@bath.ac.uk

(Received February 2021; accepted April 2022)