Statistical Science
2019, Vol. 34, No. 1, 129–155
https://doi.org/10.1214/18-STS668

Generalized Multiple Importance Sampling

Víctor Elvira, Luca Martino, David Luengo and Mónica F. Bugallo

Abstract. Importance sampling (IS) methods are broadly used to approximate posterior distributions or their moments. In the standard IS approach, samples are drawn from a single proposal distribution and weighted adequately. However, since the performance in IS depends on the mismatch between the targeted and the proposal distributions, several proposal densities are often employed for the generation of samples. Under this multiple importance sampling (MIS) scenario, extensive literature has addressed the selection and adaptation of the proposal distributions, interpreting the sampling and weighting steps in different ways. In this paper, we establish a novel general framework with sampling and weighting procedures when more than one proposal is available. The new framework encompasses most relevant MIS schemes in the literature, and novel valid schemes appear naturally. All the MIS schemes are compared and ranked in terms of the variance of the associated estimators. Finally, we provide illustrative examples revealing that, even with a good choice of the proposal densities, a careful interpretation of the sampling and weighting procedures can make a significant difference in the performance of the method.

Key words and phrases: Monte Carlo methods, multiple importance sampling, Bayesian inference.

1. INTRODUCTION

Importance sampling (IS) is a well-known Monte Carlo technique that can be applied to compute integrals involving target probability density functions (p.d.f.'s) (Robert and Casella, 2004; Liu, 2008). The standard IS technique draws samples from a single proposal p.d.f. and assigns them weights based on the ratio between the target and the proposal p.d.f.'s, both evaluated at the sample value. The choice of a suitable proposal p.d.f. is crucial for obtaining a good approximation of the target p.d.f. using the IS method. Indeed, although the validity of this approach is guaranteed under mild assumptions, the variance of the estimator depends on the discrepancy between the shape of the proposal and the target (Robert and Casella, 2004; Liu, 2008).

Several advanced strategies have been proposed in the literature to design more robust IS schemes (Liu, 2008, Chapter 2; Owen, 2013, Chapter 9; Liang, 2002). A powerful approach is based on using a population of different proposal p.d.f.'s. This approach is often referred to as *multiple* importance sampling (MIS) and several possible implementations have been proposed depending on the specific assumptions of the problem, for example, the knowledge of the normalizing constants, prior information of the proposals, etc. (Veach and Guibas, 1995, Hesterberg, 1995; Owen and Zhou, 2000; Tan, 2004; He and Owen, 2014; Elvira et al., 2015). In general, MIS strategies provide more robust algorithms, since they avoid entrusting the performance of the method to a single proposal. Moreover,

Víctor Elvira is Associate Professor, IMT Lille Douai & CRIStAL Laboratory (UMR 9189), Rue Guglielmo Marconi, Villeneuve d'Ascq 59653, France (e-mail: victor.elvira@imt-lille-douai.fr). Luca Martino is Visiting Professor, Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganés, Madrid, Spain (e-mail: lmartino@ing.uc3m.es). David Luengo is Associate Professor, Department of Signal Theory and Communications, Universidad Politécnica de Madrid, C/ Nikola Tesla, 28031 Madrid, Spain (e-mail: david.luengo@upm.es). Mónica F. Bugallo is Professor, Department of Electrical & Computer Engineering, Stony Brook University, Stony Brook, New York 11794-2350, USA (e-mail: monica.bugallo@stonybrook.edu).

many algorithms have been proposed in order to conveniently adapt the set of proposals in MIS (Cappé et al., 2004; Martino et al., 2017; Elvira et al., 2017).

When a set of proposal p.d.f.'s is available, the way in which the samples can be drawn and weighted is not unique, unlike the case of using a single proposal. Indeed, different MIS algorithms in the literature (both adaptive and nonadaptive) have implicitly and independently interpreted the sampling and weighting procedures in different ways (Owen and Zhou, 2000; Cappé et al., 2004, 2008; Elvira et al., 2015; Martino et al., 2015a; Cornuet et al., 2012; Bugallo et al., 2017). Namely, there are several possible combinations of sampling and weighting schemes, when a set of proposal p.d.f.'s is available, which lead to valid MIS approximations of the target p.d.f. However, these different possibilities can largely differ in terms of performance of the corresponding estimators.

In this paper, we introduce a unified framework for MIS schemes, providing a general theoretical description of the possible sampling and weighting procedures when a set of proposal p.d.f.'s is used to produce an IS approximation. Within this unified context, it is possible to interpret that all the MIS algorithms draw samples from an equally-weighted mixture distribution obtained from the set of available proposal p.d.f.'s. Three different sampling approaches and five different functions to calculate the weights of the generated samples are proposed and discussed. Moreover, we state two basic rules for possibly devising new valid sampling and weighting strategies within the proposed framework. All the analyzed combinations of sampling/weighting provide consistent estimates of the parameters of interest.

The proposed generalized framework includes all of the existing MIS methodologies that we are aware of (applied within different algorithms, e.g., in Elvira et al., 2015, 2017; Cappé et al., 2004; Cornuet et al., 2012; Martino et al., 2015a, 2017) and allows the design of novel techniques (here we propose three new schemes, but more can be introduced). An exhaustive theoretical analysis is provided by introducing general expressions for sampling and weighting in this generalized MIS context, and by proving that they yield consistent estimators. Furthermore, we compare the performance of the different MIS schemes (the proposed and existing ones) in terms of the variance of the estimators.

The rest of this paper is organized as follows. In Section 2, we describe the problem and we revisit the standard IS methodology. In Section 3, we discuss the sampling procedure in MIS, propose three new sampling strategies, and analyze some distributions of interest. In Section 4, we propose five different weighting functions, some of them completely new, and show their validity. The different combinations of sampling/weighting strategies are analyzed in Section 5, establishing the connections with existent MIS schemes, and describing three novel MIS schemes. In Section 6, we analyze the performance of the different MIS schemes in terms of the variance of the estimators. Then Section 7 discusses some relevant aspects about the application of the proposed MIS schemes, including their use in adaptive settings. Finally, Section 8 presents some descriptive numerical examples where the different MIS schemes are simulated, and Section 9 contains some concluding remarks. A running example is introduced in Section 3 and continued in Section 4, Section 5, Section 6 and Section 8 in order to clarify the flow of the paper. In addition, we perform numerical simulations on the running example, where the proposal p.d.f.'s are intentionally well chosen, to evidence the significant effects produced by the different interpretations of the sampling and weighting schemes.

2. PROBLEM STATEMENT AND BACKGROUND

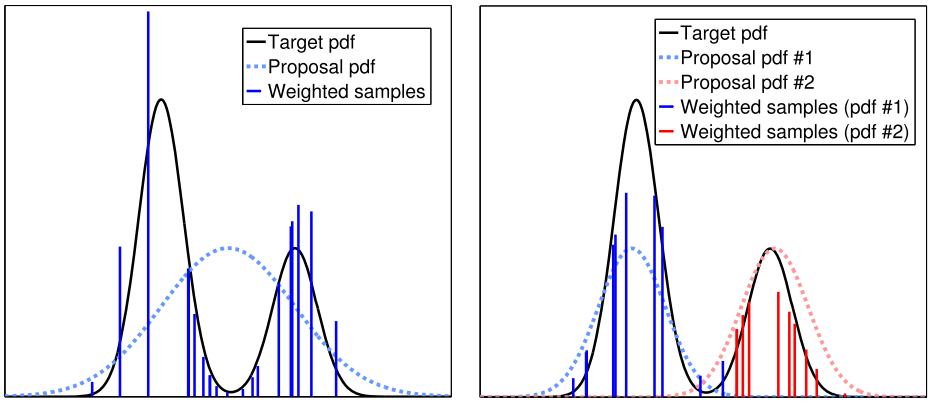
Let us consider a system characterized by a vector of d_x unknown parameters, $\mathbf{x} \in \mathbb{R}^{d_x}$ and a set of d_y observed data, $\mathbf{y} \in \mathbb{R}^{d_y}$.[1] A general objective is to extract the complete information about the latent state, $\mathbf{x}$, given the observations, $\mathbf{y}$, by means of studying the posterior distribution defined as

$$\tilde{\pi}(\mathbf{x}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\mathbf{x})h(\mathbf{x})}{Z(\mathbf{y})} \propto \pi(\mathbf{x}|\mathbf{y}) = \ell(\mathbf{y}|\mathbf{x})h(\mathbf{x}), \tag{2.1}$$

where $\ell(\mathbf{y}|\mathbf{x})$ is the likelihood function, $h(\mathbf{x})$ is the prior p.d.f. and $Z(\mathbf{y})$ is the normalization factor.[2] The objective is to approximate the p.d.f. of interest (referred to as target p.d.f.) by Monte Carlo-based sampling (Kong et al., 2003; Robert and Casella, 2004; Liu, 2008; Owen, 2013). The resulting approximation of $\tilde{\pi}(\mathbf{x})$ will be denoted as $\hat{\pi}(\mathbf{x})$ and will be attained using IS techniques.

[1] Vectors are denoted by bold-faced letters, for example, $\mathbf{x}$, while regular-faced letters are used for scalars, for example, x.

[2] In the sequel, to simplify the notation, the dependence on $\mathbf{y}$ is removed, for example, $Z \equiv Z(\mathbf{y})$.



(a) Single proposal pdf (standard IS). (b) Two proposal pdfs (MIS).

FIG. 1. *Approximation of the target p.d.f., $\pi(\mathbf{x})$, by the random measure χ.*

2.1 Standard Importance Sampling

IS is a general Monte Carlo technique for the approximation of a p.d.f. of interest by a random measure composed of samples and weights (Robert and Casella, 2004). In its original formulation, a set of N samples, $\{\mathbf{x}_n\}_{n=1}^N$, is drawn from a single proposal p.d.f., $q(\mathbf{x})$, with heavier tails than those of the target p.d.f., $\pi(\mathbf{x})$. A particular sample, $\mathbf{x}_n$, is assigned an importance weight given by

$$w_n = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}, \quad n = 1, \ldots, N, \tag{2.2}$$

which represents the ratio between the target p.d.f., π and the proposal p.d.f., q, both evaluated at $\mathbf{x}_n$. The samples and weights form the random measure $\chi = \{\mathbf{x}_n, w_n\}_{n=1}^N$ that approximates the measure of the target p.d.f. as

$$\hat{\pi}_{\mathrm{IS}}(\mathbf{x}) = \frac{1}{N\hat{Z}} \sum_{n=1}^{N} w_n \delta_{\mathbf{x}_n}(\mathbf{x}), \tag{2.3}$$

where $\delta_{\mathbf{x}_n}(\mathbf{x})$ is the unit delta measure concentrated at $\mathbf{x}_n$ and $\hat{Z} = \frac{1}{N}\sum_{j=1}^N w_j$ is an unbiased estimator of $Z = \int \pi(\mathbf{x})\, d\mathbf{x}$ (Robert and Casella, 2004). Figure 1(a) displays an example of a target p.d.f. and a proposal p.d.f., as well as the samples and weights that form a random measure approximating the posterior. Note that, unlike Markov chain Monte Carlo (MCMC) methods, all the generated samples are used to build the estimators, for example, there is no burn-in period.

2.2 Estimators in Importance Sampling

Let us consider the integral $I = \int g(\mathbf{x})\tilde{\pi}(\mathbf{x})\, d\mathbf{x}$, where g is any integrable function w.r.t. $\tilde{\pi}(\mathbf{x})$. When Z is known, an unbiased IS estimator of I is given by

$$\hat{I} = \frac{1}{NZ} \sum_{n=1}^{N} w_n g(\mathbf{x}_n). \tag{2.4}$$

Otherwise, if the target distribution is only known up to the normalizing constant, Z, one can use the self-normalized estimator

$$\tilde{I} = \frac{1}{N\hat{Z}} \sum_{n=1}^{N} w_n g(\mathbf{x}_n), \tag{2.5}$$

where Z is approximated by the estimate

$$\hat{Z} = \frac{1}{N} \sum_{n=1}^{N} w_n. \tag{2.6}$$

Under some mild assumptions regarding the tails of the proposal and target distributions, Z is an unbiased and consistent estimator of Z, and $\tilde{I}$ is a consistent estimator of I (Robert and Casella, 2004). Furthermore, the variance of $\hat{I}$ and $\tilde{I}$ is directly related to the discrepancy between $\tilde{\pi}(\mathbf{x})|g(\mathbf{x})|$ and $q(\mathbf{x})$ (Robert and Casella, 2004; Kahn and Marshall, 1953). For a general g, a common strategy is decreasing the mismatch between the proposal $q(\mathbf{x})$ and the target $\tilde{\pi}(\mathbf{x})$. A very common strategy consists in using several proposal p.d.f.'s.

3. SAMPLING IN MULTIPLE IMPORTANCE SAMPLING

MIS schemes consider a set of N proposal p.d.f.'s, $\{q_n(\mathbf{x})\}_{n=1}^N \equiv \{q_1(\mathbf{x}), \ldots, q_N(\mathbf{x})\}$, and proceed by drawing M samples, $\{\mathbf{x}_m\}_{m=1}^M$ (where $M \neq N$, in general) and properly weighting them. As a visual example, Figure 1(b) displays a target p.d.f. and two proposal p.d.f.'s, as well as the samples and weights that form a random measure approximating the posterior.

It is in the way that the sampling and the weighting are performed that different variants of MIS can be devised. In this section, we focus on the generation of samples $\{\mathbf{x}_m\}_{m=1}^M$. For clarity in the explanations and the theoretical proofs, we always consider $M = N$, that is, the number of samples to be generated coincides with the number of proposal p.d.f.'s. All the considerations can be automatically extended to the case with $M = kN$ samples, with $k \in \mathbb{N}^+$. The sampling and weighting procedures that we propose in the following can be easily applied to each block of N samples. Then the estimators described in the previous section would use all the $M = kN$ samples. In this work, we consider that we have no prior information about the adequacy of the proposals. Hence, all the proposals will be equally used for sampling purposes (see more details in Section 3.1). The use of the complete set of N proposal p.d.f.'s with no prior information about them can also represent a single equally weighted mixture proposal,

$$\psi(\mathbf{x}) \equiv \frac{1}{N}\sum_{n=1}^{N} q_n(\mathbf{x}). \tag{3.1}$$

Unequal weights could also be considered in the mixture. In He and Owen (2014), the weights can be optimized to minimize the variance for a certain integrand.

3.1 Sampling from the Set of Proposal P.d.f.'s

Let us consider a generic mechanism for the simulation of N samples from the set of N proposals. Starting with $n = 1$:

1. Choose an index $j_n \in \{1, \dots, N\}$, which corresponds to the selection of the proposal p.d.f. q_{j_n}.
2. Generate a sample $\mathbf{x}_n$ from the selected proposal p.d.f., that is, $\mathbf{x}_n \sim q_{j_n}(\mathbf{x}_n)$.
3. Set $n = n + 1$ and go to step 1.

Note that, in step 1, the probabilities associated to each possible value of j_n are not specified yet. The graphical model corresponding to this sampling scheme is shown in Figure 2.

Therefore, obtaining the set of samples $\{\mathbf{x}_n\}_{n=1}^N \equiv \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ is in general a two step sequential procedure. First, the nth index j_n is drawn according to some conditional p.d.f., $P(j_n|j_{1:n-1})$, where $j_{1:n-1} \equiv \{j_1, \dots, j_{n-1}\}$ is the sequence of the previously generated indexes.[3] Then the nth sample is drawn from the selected proposal p.d.f. as $\mathbf{x}_n \sim p(\mathbf{x}_n|j_n)$. The joint probability distribution of the current sample and all the indexes used to generate the samples from 1 to n is

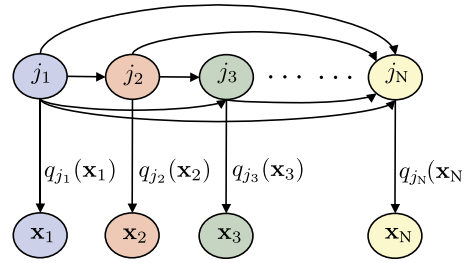


FIG. 2. *Graphical model associated to the generic sampling scheme.*

$$\begin{aligned} p(\mathbf{x}_n, j_{1:n}) &= P(j_{1:n-1})P(j_n|j_{1:n-1})p(\mathbf{x}_n|j_n) \\ &= P(j_1)\left[\prod_{i=2}^{n} P(j_i|j_{1:i-1})\right] q_{j_n}(\mathbf{x}_n), \end{aligned} \tag{3.2}$$

where $p(\mathbf{x}_n|j_n) = q_{j_n}(\mathbf{x}_n)$ is the nth selected proposal p.d.f., $q_{j_n}(\mathbf{x}_n)$.

3.2 Selection of the Proposal P.d.f.'s

In the sequel, we describe three mechanisms for obtaining the sequence of indexes, $j_{1:N}$. All the mechanisms share the property that

$$\frac{1}{N}\sum_{n=1}^{N} P(J_n = k) = \frac{1}{N} \quad \forall k \in \{1, \dots, N\}, \tag{3.3}$$

that is, all the indexes have the same (marginal) probability of being selected.

$\mathcal{S}_1$: *Random index selection with replacement*: The N indexes are independently drawn from the set $\{1, \dots, N\}$ with equal probability. Thus, we have

$$P(j_n|j_{1:n-1}) = P(j_n) = \frac{1}{N}. \tag{3.4}$$

With this type of index sampling, there may be more than one sample drawn from some proposal, and there may be proposal p.d.f.'s that are not used at all.

$\mathcal{S}_2$: *Random index selection without replacement*: The indexes are uniformly and sequentially drawn from different sets as $j_1 \in \mathcal{I}_1 = \{1, \dots, N\}, \dots, j_n \in \mathcal{I}_n =$

[3] We use a simplified argument-wise notation, where $p(\mathbf{x}_n)$ denotes the p.d.f. of the continuous random variable (r.v.) $\mathbf{X}_n$, while $P(j_n)$ denotes the probability mass function (p.m.f.) of the discrete r.v. J_n. Also, $p(\mathbf{x}_n, j_n)$ denotes the joint p.d.f. and $p(\mathbf{x}_n|j_n)$ is the conditional p.d.f. of $\mathbf{X}_n$ given $J_n = j_n$. If the argument of $p(\cdot)$ is different from $\mathbf{x}_n$, then it denotes the evaluation of the p.d.f. as a function, for example, $p(\mathbf{z}|j_n)$ denotes the p.d.f. $p(\mathbf{x}_n|j_n)$ evaluated at $\mathbf{x}_n = \mathbf{z}$.

$\{1,\dots,N\}\setminus\{j_{1:n-1}\}$, that is, removing the proposals previously used. Hence, the conditional probability mass function (p.m.f.) of the nth index given the previous ones is now

$$P(J_n=k|j_{1:n-1})=\begin{cases}\dfrac{1}{N-n+1} & \text{if } k\in\mathcal{I}_n,\\ 0 & \text{if } k\notin\mathcal{I}_n,\end{cases} \tag{3.5}$$

where $|\mathcal{I}_n|=N-n+1$. Note that the marginal p.m.f. of the jth index is still given by (3.4).[4] However, exactly one sample is drawn from each of the proposal p.d.f.'s by following this strategy.

$\mathcal{S}_3$: *Deterministic index selection without replacement*: This sampling is a particular case of sampling $\mathcal{S}_2$, where a fixed deterministic sequence of indexes is drawn. For instance, and without loss of generality: $j_1=1, j_2=2,\dots,j_n=n,\dots,j_N=N$. Therefore, $\mathbf{x}_n\sim q_{j_n}(\mathbf{x}_n)=q_n(\mathbf{x}_n)$, and the conditional p.m.f. of the nth index given the $n-1$ previous ones becomes

$$P(j_n|j_{1:n-1})=P(j_n)=\mathbb{1}_{j_n=n}, \tag{3.6}$$

where $\mathbb{1}$ denotes the indicator function. Again, each of the N proposal p.d.f.'s is used to generate exactly one sample of the set $\{\mathbf{x}_n\}_{n=1}^N$. This index selection procedure has been used by several MIS algorithms (e.g., in Cornuet et al., 2012, Elvira et al., 2017), and it is also implicitly used in some particle filters (PFs), such as the bootstrap PF (Gordon, Salmond and Smith, 1993).

The connexions of the sampling mechanisms with some resampling schemes are discussed in Appendix B.

3.3 Running Example

Let us consider $N=3$ Gaussian proposal p.d.f.'s $q_1(x)=\mathcal{N}(x;\mu_1,\sigma_1^2)$, $q_2(x)=\mathcal{N}(x;\mu_2,\sigma_2^2)$ and $q_3(x)=\mathcal{N}(x;\mu_3,\sigma_3^2)$ with predefined means and variances. In $\mathcal{S}_1$, a possible realization of the indexes is the sequence $\{j_1,j_2,j_3\}=\{3,3,1\}$. Therefore, in this situation, $\mathbf{x}_1\sim q_3$, $\mathbf{x}_2\sim q_3$, and $\mathbf{x}_3\sim q_1$. In $\mathcal{S}_2$, the realization could result from the permutation $\{j_1,j_2,j_3\}=\{3,1,2\}$. In $\mathcal{S}_3$, the sequence is deterministically obtained as $\{j_1,j_2,j_3\}=\{1,2,3\}$.

[4] There are $N!$ equiprobable configurations (permutations) of the sequence $\{j_1,\dots,j_N\}$, and in $(N-1)!$ the kth index is drawn at the nth position $\forall k,n=1,\dots,N$. Therefore, $P(J_n=k)=\frac{(N-1)!}{N!}=\frac{1}{N}$ $\forall k,n=1,\dots,N$.

3.4 Distributions of Interest of the nth Sample, $\mathbf{x}_n$

In the following, we discuss some important distributions related to the set of samples drawn. These distributions are of utmost importance to understand the different methods for weighting the samples discussed in the following section.

Note that the distribution of the nth sample given all the knowledge of the process up to that point is $p(\mathbf{x}_n|j_{1:n-1},\mathbf{x}_{1:n-1})=p(\mathbf{x}_n|j_{1:n-1})$. In $\mathcal{S}_1$, this distribution corresponds to $p(\mathbf{x}_n|j_{1:n-1})=\psi(\mathbf{x}_n)$. We recall that ψ is the mixture of proposals defined in equation (3.1). In $\mathcal{S}_2$, we have $p(\mathbf{x}_n|j_{1:n-1})=\frac{1}{|\mathcal{I}_n|}\sum_{\forall k\in\mathcal{I}_n}q_k(\mathbf{x})$. Finally, under $\mathcal{S}_3$, $p(\mathbf{x}_n|j_{1:n-1})=q_n(\mathbf{x}_n)$. Once the nth index j_n has been selected, the nth sample, $\mathbf{x}_n$, is distributed as $p(\mathbf{x}_n|j_n)=q_{j_n}(\mathbf{x}_n)$ in any sampling method within the proposed framework. The marginal distribution of this nth sample, $\mathbf{x}_n$, is then given by

$$p(\mathbf{x}_n)=\sum_{k=1}^{N}q_k(\mathbf{x}_n)P(J_n=k), \tag{3.7}$$

where we have used the fact that $p(\mathbf{x}_n|J_n=k)=q_k(\mathbf{x}_n)$, and the marginal distribution, $P(J_n=k)$, depends on the sampling method. When randomly selecting the indexes ($\mathcal{S}_1$ or $\mathcal{S}_2$), $P(J_n=k)=\frac{1}{N},\forall n,k$, and thus $p(\mathbf{x}_n)=\frac{1}{N}\sum_{k=1}^{N}q_k(\mathbf{x}_n)=\psi(\mathbf{x}_n)$. In the case of the deterministic index selection ($\mathcal{S}_3$), $P(J_n=k)=\mathbb{1}_{k=n}$, and thus $p(\mathbf{x}_n)=q_n(\mathbf{x}_n)$, that is, the distribution of the r.v. $\mathbf{X}_n$ is the nth proposal p.d.f., and not the whole mixture.

3.5 Distributions of Interest Beyond $\mathbf{x}_n$

The traditional IS approach focuses just on the distribution of the r.v. $\mathbf{X}_n$. In MIS, we are also interested in the statistical properties of the set of samples, regardless of their index n, since the N samples are used jointly in the estimators, regardless their order of appearance. Hence, we introduce a generic r.v.,

$$\mathbf{X}=\mathbf{X}_n \quad \text{with } n\sim\mathcal{U}\{1,2,\dots,N\}, \tag{3.8}$$

where $\mathcal{U}\{1,2,\dots,N\}$ is the discrete uniform distribution on the set $\{1,2,\dots,N\}$. The density of $\mathbf{X}$ is then given by

$$f(\mathbf{x})=\frac{1}{N}\sum_{n=1}^{N}p_{\mathbf{x}_n}(\mathbf{x})=\psi(\mathbf{x}), \tag{3.9}$$

where $p_{\mathbf{x}_n}(\mathbf{x})$ denotes the marginal p.d.f. of $\mathbf{X}_n$, given by equation (3.7), evaluated at $\mathbf{x}$, and $\psi(\mathbf{x})$ is the mixture p.d.f.[5] Moreover, one can also obtain the condi-

[5] For the sake of clarity, in equation (3.9) we have used the notation $p_{\mathbf{x}_n}(\mathbf{x})$, instead of $p(\mathbf{x})$ as in equation (3.7) and the rest of the paper, to denote the marginal p.d.f. of $\mathbf{X}_n$ evaluated at $\mathbf{x}$.

tional p.d.f. of $\mathbf{X}$ given the sequence of indexes as

$$f(\mathbf{x}|j_{1:N}) = \frac{1}{N}\sum_{k=1}^{N} p_{\mathbf{x}_k}(\mathbf{x}|j_{1:N}) = \frac{1}{N}\sum_{k=1}^{N} q_{j_k}(\mathbf{x}). \tag{3.10}$$

Note that, in this case, $f(\mathbf{x}|j_{1:N}) = \psi(\mathbf{x})$ for the schemes without replacement at the index selection ($\mathcal{S}_2$ and $\mathcal{S}_3$), but $f(\mathbf{x}|j_{1:N}) = \frac{1}{N}\sum_{n=1}^{N} q_{j_n}(\mathbf{x})$ for the case with replacement ($\mathcal{S}_1$), that is, some proposal p.d.f.'s may not appear while others may appear repeated.

REMARK 3.1 (Sampling). In the proposed framework, we consider valid, any sequential sampling scheme for generating the set $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ such that the p.d.f. of the r.v. $\mathbf{X}$ defined in equation (3.8) is given by $\psi(\mathbf{x})$. Further considerations about the r.v. $\mathbf{X}$ and connections with variance reduction methods (Robert and Casella, 2004; Owen, 2013) are given in Appendix A.

Table 1 summarizes all the distributions of interest. Note that, the p.d.f. of the r.v. $\mathbf{X}$ is always the mixture $\psi(\mathbf{x})$, but different sampling procedures yield different conditional and marginal distributions that will be exploited to justify different strategies for calculation of the importance weights in the next section. Finally, the last row of the table shows the joint distribution $p(\mathbf{x}_{1:N})$ of the variables $\mathbf{X}_1, \ldots, \mathbf{X}_N$, that is, $p(\mathbf{x}_{1:N}) = \prod_{n=1}^{N} \psi(\mathbf{x}_n)$ and $p(\mathbf{x}_{1:N}) = \prod_{n=1}^{N} q_n(\mathbf{x}_n)$ for $\mathcal{S}_1$ and $\mathcal{S}_3$, respectively. For $\mathcal{S}_2$,

$$p(\mathbf{x}_{1:N}) = \psi(\mathbf{x}_1)\prod_{n=2}^{N}\frac{1}{|\mathcal{I}_n|}\sum_{\ell\in\mathcal{I}_n} q_\ell(\mathbf{x}_n), \tag{3.11}$$

with $\mathcal{I}_n = \{1, \ldots, N\} \setminus \{j_{1:n-1}\}$.

4. WEIGHTING IN MULTIPLE IMPORTANCE SAMPLING

Our approach is based on analyzing which weighting functions yield *proper* MIS estimators. We consider that the set of weighting functions $\{w_n\}_{n=1}^{N}$ is proper if

$$\frac{E_{p(\mathbf{x}_{1:N}, j_{1:N})}[\frac{1}{N}\sum_{n=1}^{N} w_n g(\mathbf{x}_n)]}{E_{p(\mathbf{x}_{1:N}, j_{1:N})}[\frac{1}{N}\sum_{n=1}^{N} w_n]} = E_\pi[g(\mathbf{x})]. \tag{4.1}$$

This is equivalent to imposing the restriction

$$\frac{E_{p(\mathbf{x}_{1:N}, j_{1:N})}[Z\hat{I}]}{E_{p(\mathbf{x}_{1:N}, j_{1:N})}[\hat{Z}]} = I, \tag{4.2}$$

which is fulfilled if $\mathrm{E}[\hat{I}] = I$ and $\mathrm{E}[\hat{Z}] = Z$. In order to narrow down the set of all possible proper functions, we impose the weight function to have the (deterministic) structure $w_n = \frac{\pi(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}$, where $\varphi_{\mathcal{P}_n}$ is a generic function parametrized by a set of parameters $\mathcal{P}_n \subseteq \{j_1, \ldots, j_N\}$ (further details are given below). Note that $\varphi_{\mathcal{P}_n}$ plays the role of the *interpreted* proposal from which $\mathbf{x}_n$ is drawn. It is on this interpretation of what the proposal p.d.f. used for the generation of the sample is that different weighting strategies can be devised.[6] The expectation of the generic estimator $\hat{I}$ of equation (2.4) can be computed as

$$\mathrm{E}[\hat{I}] = \frac{1}{ZN}\sum_{n=1}^{N}\sum_{j_{1:N}}\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)} \cdot P(j_{1:N})\, p(\mathbf{x}_n|j_n)\, d\mathbf{x}_n, \tag{4.3}$$

where we use the joint distribution of indexes and samples from equation (3.2).

REMARK 4.1. (Weighting): In the proposed framework, we consider valid any weighting scheme (i.e., any function $\varphi_{\mathcal{P}_n}$ at the denominator of the weight) that yields $\mathrm{E}[\hat{I}] \equiv I$ in equation (4.3).

4.1 Weighting Functions

Here we present several possible functions $\varphi_{\mathcal{P}_n}$, that yield an unbiased estimator of I according to equation (4.3). The different choices for $\varphi_{\mathcal{P}_n}$, used in the denominator of the weight $w_n = \frac{\pi(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}$, come naturally from the sampling densities discussed in Section 3. More precisely, they correspond to the five different functions in Table 1 related to the distributions of the generated samples. From now on, $p(\cdot)$ and $f(\cdot)$, which correspond to the p.d.f.'s of $\mathbf{X}_n$ and $\mathbf{X}$, respectively, are used as functions and the argument represents a functional evaluation.

$\mathcal{W}_1$: $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi_{j_{1:n-1}}(\mathbf{x}_n) = p(\mathbf{x}_n|j_{1:n-1})$
Since the sampling process is sequential, this option is of particular interest. It interprets the proposal p.d.f. as the conditional density of $\mathbf{x}_n$ given all the previous proposal indexes of the sampling process.

$\mathcal{W}_2$: $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi_{j_n}(\mathbf{x}_n) = p(\mathbf{x}_n|j_n) = q_{j_n}(\mathbf{x}_n)$
It interprets that if the index j_n is known, $\varphi_{\mathcal{P}_n}$ is the proposal q_{j_n}.

[6] In an even more generalized framework, w_n could hypothetically depend on more than one sample of the set $\mathbf{x}_{1:N}$ if one could properly design the function φ_n that yields valid estimators.

TABLE 1
Summary of the distributions of the r.v.'s J_n, $\mathbf{X}_n$ and $\mathbf{X}$, for the three different sampling procedures

	Selection of the indexes			
		Without replacement		
Distributions	With replacement $\mathcal{S}_1$	Random selection $\mathcal{S}_2$	Deterministic selection $\mathcal{S}_3$	Text references
$J_n \sim P(j_n)$	$\frac{1}{N}$	$\frac{1}{N}$	$\mathbb{1}_{j_n=n}$	Eqs. (3.4) and (3.6)
$J_n \vert J_{1:n-1} \sim P(j_n \vert j_{1:n-1})$	$\frac{1}{N}$	$\frac{1}{\vert\mathcal{I}_n\vert}\mathbb{1}_{j_n\in\mathcal{I}_n}$	$\mathbb{1}_{j_n=n}$	Eqs. (3.4)–(3.5)–(3.6)
$\mathbf{X}_n \vert J_{1:n-1} \sim p(\mathbf{x}_n \vert j_{1:n-1})$	$\psi(\mathbf{x}_n)$	$\frac{1}{\vert\mathcal{I}_n\vert}\sum_{\forall k\in\mathcal{I}_n} q_k(\mathbf{x}_n)$	$q_n(\mathbf{x}_n)$	Section 3.4
$\mathbf{X}_n \vert J_n \sim p(\mathbf{x}_n \vert j_n)$	$q_{j_n}(\mathbf{x}_n)$	$q_{j_n}(\mathbf{x}_n)$	$q_{j_n}(\mathbf{x}_n) = q_n(\mathbf{x}_n)$	Section 3.1
$\mathbf{X}_n \sim p(\mathbf{x}_n)$	$\psi(\mathbf{x}_n)$	$\psi(\mathbf{x}_n)$	$q_n(\mathbf{x}_n)$	Eq. (3.7)
$\mathbf{X} \vert J_{1:N} \sim f(\mathbf{x} \vert j_{1:N})$	$\frac{1}{N}\sum_{n=1}^N q_{j_n}(\mathbf{x})$	$\psi(\mathbf{x})$	$\psi(\mathbf{x})$	Eq. (3.10)
$\mathbf{X} \sim f(\mathbf{x})$	$\psi(\mathbf{x})$	$\psi(\mathbf{x})$	$\psi(\mathbf{x})$	Eq. (3.9)
$\mathbf{X}_{1:N} \sim p(\mathbf{x}_{1:N})$	$\prod_{n=1}^N \psi(\mathbf{x}_n)$	$\psi(\mathbf{x}_1)\prod_{n=2}^N \frac{1}{\vert\mathcal{I}_n\vert}\sum_{\ell\in I_n} q_\ell(\mathbf{x}_n)$	$\prod_{n=1}^N q_n(\mathbf{x}_n)$	Section 3.5; Eq. (3.11)

$\mathcal{W}_3$: $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = p(\mathbf{x}_n)$
It interprets that $\mathbf{x}_n$ is a realization of the marginal $p(\mathbf{x}_n)$. This is probably the most "natural" option (as it does not assume any further knowledge in the generation of $\mathbf{x}_n$) and is a usual choice for the calculation of the weights in some of the existing MIS schemes (see Section 5).

$\mathcal{W}_4$: $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi_{j_{1:N}}(\mathbf{x}_n) = f(\mathbf{x}_n \vert j_{1:N}) = \frac{1}{N}\sum_{k=1}^N q_{j_k}(\mathbf{x}_n)$
This interpretation makes use of the distribution of the r.v. $\mathbf{X}$ conditioned on the whole set of indexes (defined in Section 3.5).

$\mathcal{W}_5$: $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi(\mathbf{x}_n) = f(\mathbf{x}_n) = \frac{1}{N}\sum_{k=1}^N q_k(\mathbf{x}_n)$
This option considers that all the $\mathbf{x}_n$ are realizations of the r.v. $\mathbf{X}$ defined in Section 3.5 (see Appendix A for a thorough discussion of this interpretation).

Table 2 summarizes the discussed functions $\varphi_{\mathcal{P}_n}$. Although some of the selected functions $\varphi_{\mathcal{P}_n}$ may seem more natural than others, all of them yield valid estimators. The proofs can be found in Appendix C. Other proper weighting functions are described in Section 7.2.

4.2 Connection with Liu-Properness of Single IS

We consider the definition of properness by Liu (2008), Section 2.5 and we extend (or relax) it to the MIS scenario. Namely, Liu-properness in standard IS states that a weighted sample $\{\mathbf{x}_n, w_n\}$ drawn from a single proposal q is proper if, for any square integrable function g,

$$\frac{E_q[g(\mathbf{x})w(\mathbf{x})]}{E_q[\pi(\mathbf{x})]} = E_\pi\big[g(\mathbf{x})\big], \tag{4.4}$$

that is, w can be in any form as long as the condition of equation (4.4) is fulfilled. Note that, for a deterministic weight assignment, the only proper weights are the ones considered by the standard IS approach. Note also that the MIS properness is a relaxation of the one proposed by Liu, that is, any Liu-proper weighting scheme is also proper a according to our definition, but not vice versa.

4.3 Running Example

Here we follow the running example of Section 3.3. For instance, let us consider the sampling method $\mathcal{S}_1$ and let the realization of the indexes be the sequence

TABLE 2
Summary of the different generic functions $\varphi_{\mathcal{P}_n}$. The distributions depend on the specific sampling scheme used for drawing the samples as shown in Table 3

$\varphi_{\mathcal{P}_n}$	$\mathcal{W}_1$ $p(\mathbf{x}_n \vert j_{1:n-1})$	$\mathcal{W}_2$ $p(\mathbf{x}_n \vert j_n)$	$\mathcal{W}_3$ $p(\mathbf{x}_n)$	$\mathcal{W}_4$ $f(\mathbf{x} \vert j_{1:N})$	$\mathcal{W}_5$ $f(\mathbf{x})$
$w_n = \frac{\pi(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}$	$\frac{\pi(\mathbf{x}_n)}{p(\mathbf{x}_n \vert j_{1:n-1})}$	$\frac{\pi(\mathbf{x}_n)}{p(\mathbf{x}_n \vert j_n)}$	$\frac{\pi(\mathbf{x}_n)}{p(\mathbf{x}_n)}$	$\frac{\pi(\mathbf{x}_n)}{f(\mathbf{x}_n \vert j_{1:N})}$	$\frac{\pi(\mathbf{x}_n)}{f(\mathbf{x}_n)}$

TABLE 3
Specific function, $\varphi_{\mathcal{P}_n}$, at the denominator of weight, $w_n = \frac{\pi(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}$, resulting from the combination of the different sampling schemes (Section 3.5) and weighting functions (Section 4.1)

$\varphi_{\mathcal{P}_n}$	$\mathcal{W}_1$ $p(\mathbf{x}_n \mid j_{1:n-1})$	$\mathcal{W}_2$ $p(\mathbf{x}_n \mid j_n)$	$\mathcal{W}_3$ $p(\mathbf{x}_n)$	$\mathcal{W}_4$ $f(\mathbf{x} \mid j_{1:N})$	$\mathcal{W}_5$ $f(\mathbf{x})$
$\mathcal{S}_1$: with replacement	$\psi(\mathbf{x}_n)$ [R3]	$q_{j_n}(\mathbf{x}_n)$ [R1]	$\psi(\mathbf{x}_n)$ [R3]	$\frac{1}{N}\sum_{k=1}^{N} q_{j_k}(\mathbf{x}_n)$ [R2]	$\psi(\mathbf{x}_n)$ [R3]
$\mathcal{S}_2$: w/o (random)	$\frac{1}{\lvert\mathcal{I}_n\rvert}\sum_{\forall k\in\mathcal{I}_n} q_k(\mathbf{x}_n)$ [N2]	$q_{j_n}(\mathbf{x}_n)$ [N1]	$\psi(\mathbf{x}_n)$ [N3]	$\psi(\mathbf{x}_n)$ [N3]	$\psi(\mathbf{x}_n)$ [N3]
$\mathcal{S}_3$: w/o (deterministic)	$q_n(\mathbf{x}_n)$ [N1]	$q_n(\mathbf{x}_n)$ [N1]	$q_n(\mathbf{x}_n)$ [N1]	$\psi(\mathbf{x}_n)$ [N3]	$\psi(\mathbf{x}_n)$ [N3]

$\{j_1, j_2, j_3\} = \{3, 3, 1\}$. Under the weighting scheme $\mathcal{W}_2$, the weights would be computed as $w_1 = \frac{\pi(\mathbf{x}_1)}{q_3(\mathbf{x}_1)}$, $w_2 = \frac{\pi(\mathbf{x}_2)}{q_3(\mathbf{x}_2)}$, and $w_3 = \frac{\pi(\mathbf{x}_3)}{q_1(\mathbf{x}_3)}$. However, under $\mathcal{W}_4$, $w_1 = \frac{\pi(\mathbf{x}_1)}{\frac{1}{3}(q_1(\mathbf{x}_1)+2q_3(\mathbf{x}_1))}$, $w_2 = \frac{\pi(\mathbf{x}_2)}{\frac{1}{3}(q_1(\mathbf{x}_2)+2q_3(\mathbf{x}_2)}$ and $w_3 = \frac{\pi(\mathbf{x}_3)}{\frac{1}{3}(q_1(\mathbf{x}_3)+2q_3(\mathbf{x}_3))}$. Note that all weighing schemes require the same number of target evaluations (which are usually more expensive) but different numbers of proposal evaluations.

5. MULTIPLE IMPORTANCE SAMPLING SCHEMES

In this section, we describe the different possible combinations of the three sampling strategies considered in Section 3 and the five weighting functions devised in Section 4. Once combined, the fifteen possibilities only lead to six unique MIS methods. Three of the methods are associated to the sampling scheme with replacement ($\mathcal{S}_1$), while the other three methods correspond to the sampling schemes without replacement ($\mathcal{S}_2$ and $\mathcal{S}_3$). Table 3 summarizes the possible combinations of sampling/weighting and indicates the resulting MIS method within brackets. The six MIS methods are labeled either by an `R` (sampling with *replacement*) or with an `N` (sampling with *no* replacement). We remark that these schemes are examples of proper MIS techniques fulfilling Remarks 3.1 and 4.1.

5.1 MIS Schemes with Replacement

In all `R` schemes, the nth sample is drawn with replacement (i.e., $\mathcal{S}_1$) from the whole mixture ψ:

`[R1]`**:** *Sampling with replacement, $\mathcal{S}_1$, and weight denominator $\mathcal{W}_2$*:
For the weight calculation of the nth sample, only the proposal selected for generating the sample is evaluated in the denominator.

`[R2]`**:** *Sampling with replacement, $\mathcal{S}_1$, and weight denominator $\mathcal{W}_4$*:
With the N selected indexes j_n, for $n = 1, \ldots, N$, one forms a mixture comprising all the corresponding proposal p.d.f.'s. The weight calculation of the nth sample considers this *a posteriori* mixture evaluated at the nth sample in the denominator, that is, some proposals might be used more than once while other proposals might not be used.

`[R3]`**:** *Sampling with replacement, $\mathcal{S}_1$, and weight denominator $\mathcal{W}_1$, $\mathcal{W}_3$ or $\mathcal{W}_5$*:
For the weight calculation of the nth sample, the denominator applies the value of the nth sample to the whole mixture ψ composed of the set of initial proposal p.d.f.'s (i.e., the function in the denominator of the weight does not depend on the sampling process). This is the approach followed by the so-called mixture PMC method (Cappé et al., 2008).

5.2 MIS Schemes Without Replacement

In all `N` schemes, exactly one sample is generated from each proposal p.d.f. This corresponds to having a sampling strategy without replacement.

`[N1]`**:** *Sampling without replacement (random or deterministic), $\mathcal{S}_2$ or $\mathcal{S}_3$, and weight denominator $\mathcal{W}_2$ (for $\mathcal{S}_2$) or $\mathcal{W}_1$, $\mathcal{W}_2$ or $\mathcal{W}_3$ (for $\mathcal{S}_3$)*:
For calculating the denominator of the nth weight, the specific proposal used for the generation of the sample is used. This is the approach frequently used in particle filtering (Gordon, Salmond and Smith, 1993) and in the standard PMC method (Cappé et al., 2004).

`[N2]`**:** *Sampling without replacement (random), $\mathcal{S}_2$, and weight denominator $\mathcal{W}_1$*:
This MIS implementation draws one sample from each proposal, but the order matters (it must be random) since the calculation of the nth weight uses for the evaluation of the denominator the mixture p.d.f. formed by the proposal p.d.f.'s that were still available at the generation of the nth sample.

TABLE 4
Summary of the sampling procedure and the weighting function of each MIS scheme

MIS scheme	Sampling	$w(\mathbf{x}_n)$	Used in
`R1`	$\mathcal{S}_1$	$\frac{\pi(\mathbf{x}_n)}{q_{j_n}(\mathbf{x}_n)}$	Novel scheme
`R2`	$\mathcal{S}_1$	$\frac{\pi(\mathbf{x}_n)}{\frac{1}{N}\sum_{k=1}^{N} q_{j_k}(\mathbf{x}_n)}$	Novel scheme
`R3`	$\mathcal{S}_1$	$\frac{\pi(\mathbf{x}_n)}{\psi(\mathbf{x}_n)}$	Cappé et al. (2008)
`N1`	$\mathcal{S}_3$	$\frac{\pi(\mathbf{x}_n)}{q_n(\mathbf{x}_n)}$	Cappé et al. (2004)
`N2`	$\mathcal{S}_2$	$\frac{\pi(\mathbf{x}_n)}{\frac{1}{\lvert\mathcal{I}_n\rvert}\sum_{\forall k\in\mathcal{I}_n} q_k(\mathbf{x}_n)}$	Novel scheme
`N3`	$\mathcal{S}_3$	$\frac{\pi(\mathbf{x}_n)}{\psi(\mathbf{x}_n)}$	Martino et al. (2015a); Cornuet et al. (2012)

`[N3]`**:** *Sampling without replacement (random or deterministic), $\mathcal{S}_2$ or $\mathcal{S}_3$, and weight denominator $\mathcal{W}_3$, $\mathcal{W}_4$ or $\mathcal{W}_5$ (for $\mathcal{S}_2$), or $\mathcal{W}_4$ or $\mathcal{W}_5$ (for $\mathcal{S}_3$)*:
In the calculation of the nth weight, one uses for the denominator the whole mixture. This is the approach, for instance, of Martino et al. (2015a), Cornuet et al. (2012). As shown in Section 6, this scheme has several benefits over the others.

Table 4 summarizes the six resulting MIS schemes and their references in literature, indicating the sampling procedure and weighting function that are applied to obtain the nth weighted sample $\mathbf{x}_n$. We consider `N1` and `N3` associated to $\mathcal{S}_3$ (they can also be obtained with $\mathcal{S}_2$) since it is simpler than $\mathcal{S}_2$. All the different algorithms in the literature (as far as we know) correspond to one of the MIS schemes described above (see Section 7.3). Moreover, several new valid schemes have also appeared naturally (`R1`, `R2` and `N2`), and new ones can be proposed within this framework.

5.3 Running Example

Let us consider the example from Section 3.3 where the realizations of the sequence of indexes for the sampling schemes $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ are respectively $\{j_1, j_2, j_3\} = \{3, 3, 1\}$, $\{j_1, j_2, j_3\} = \{3, 1, 2\}$ and $\{j_1, j_2, j_3\} = \{1, 2, 3\}$. Figure 3(a) shows the three first schemes of Table 4 related to the sampling with replacement, $\mathcal{S}_1$. The figure shows a possible realization of all MIS schemes with $M = N = 3$ samples and p.d.f.'s. For the nth sample, we show the set of available proposals, the index j_n of the proposal p.d.f. that was actually selected to draw the sample, the function φ_n, and the importance weight. Similarly, Figures 3(b)–(c) depict the three schemes of Table 4 related to the sampling without replacement, $\mathcal{S}_2$ and $\mathcal{S}_3$, where exactly one sample is drawn from each available proposal.

6. VARIANCE ANALYSIS OF THE SCHEMES

Although the six different MIS schemes that appear in Section 5 yield the estimator $\hat{I}$ of equation (2.4) unbiased (see Appendix C), the performance of each of the possible obtained estimators can be dramatically different. In this section, we provide an exhaustive variance analysis of the MIS schemes presented in the previous section. The details of the derivations are in Appendix D.2. The estimators of the three methods with *replacement* present the following variances:

$$\begin{aligned}\operatorname{Var}(\hat{I}_{\mathtt{R1}}) &= \frac{1}{Z^2N^2}\sum_{k=1}^{N}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_k(\mathbf{x})}\,d\mathbf{x} \\ &\quad - \frac{I^2}{N},\end{aligned} \tag{6.1}$$

$$\begin{aligned}\operatorname{Var}(\hat{I}_{\mathtt{R2}}) &= \frac{1}{Z^2N}\frac{1}{N^N}\sum_{j_{1:N}}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{f(\mathbf{x}|j_{1:N})}\,d\mathbf{x} \\ &\quad - \frac{1}{Z^2N^2}\frac{1}{N^N} \\ &\quad \cdot \sum_{j_{1:N}}\sum_{n=1}^{N}\left(\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{f(\mathbf{x}_n|j_{1:N})}q_{j_n}(\mathbf{x}_n)\,d\mathbf{x}_n\right)^2\end{aligned} \tag{6.2}$$

and

$$\operatorname{Var}(\hat{I}_{\mathtt{R3}}) = \frac{1}{Z^2N}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{\psi(\mathbf{x})}\,d\mathbf{x} - \frac{I^2}{N}. \tag{6.3}$$

On the other hand, the variances associated to the estimators of the three methods with *no* replacement

are

$$\operatorname{Var}(\hat{I}_{\mathrm{N1}}) = \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{q_n(\mathbf{x}_n)} \, d\mathbf{x}_n - \frac{I^2}{N}, \tag{6.4}$$

$$\begin{aligned} &\operatorname{Var}(\hat{I}_{\mathrm{N2}}) \\ &= \left[\frac{1}{Z^2 N^2} \sum_{n=1}^{N} \sum_{j_{1:n-1}} \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{p(\mathbf{x}_n | j_{1:n-1})} \right. \\ &\qquad \left. \cdot P(j_{1:n-1}) \, d\mathbf{x}_n \right] \\ &\quad - \left[\frac{1}{Z^2 N^2} \sum_{n=1}^{N} \sum_{j_{1:n}} \left(\int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{p(\mathbf{x}_n | j_{1:n-1})} q_{j_n} \, d\mathbf{x}_n \right)^2 \right] \\ &\quad \cdot P(j_{1:n}) \end{aligned} \tag{6.5}$$

and

$$\begin{aligned} &\operatorname{Var}(\hat{I}_{\mathrm{N3}}) \\ &= \frac{1}{Z^2 N} \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{\psi(\mathbf{x})} \, d\mathbf{x} \\ &\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \left(\int \frac{\pi(\mathbf{x}) g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x}) \, d\mathbf{x} \right)^2. \end{aligned} \tag{6.6}$$

One of the goals of this paper is to provide the practitioner with solid theoretical results about the superiority of some specific MIS schemes. In the following, we state two theorems that relate the variance of the estimator with these six methods, establishing a hierarchy among them. Note that obtaining an IS estimator with finite variance essentially amounts to having a proposal with heavier tails than the target. See Robert and Casella (2004), Geweke (1989) for sufficient conditions that guarantee this finite variance.

THEOREM 6.1. *For any target distribution $\pi(\mathbf{x})$, any square integrable function g, and any set of proposal densities $\{q_n(\mathbf{x})\}_{n=1}^{N}$ such that the variance of the corresponding MIS estimators is finite,*

$$\operatorname{Var}(\hat{I}_{\mathrm{R1}}) = \operatorname{Var}(\hat{I}_{\mathrm{N1}}) \geq \operatorname{Var}(\hat{I}_{\mathrm{R3}}) \geq \operatorname{Var}(\hat{I}_{\mathrm{N3}}).$$

PROOF. See Appendix D.2. □

THEOREM 6.2. *For any target distribution $\pi(\mathbf{x})$, any square integrable function g, and any set of proposal densities $\{q_n(\mathbf{x})\}_{n=1}^{2}$ such that the variance of the corresponding MIS estimators is finite,*

$$\begin{aligned} &\operatorname{Var}(\hat{I}_{\mathrm{R1}}) = \operatorname{Var}(\hat{I}_{\mathrm{N1}}) \geq \operatorname{Var}(\hat{I}_{\mathrm{R2}}) = \operatorname{Var}(\hat{I}_{\mathrm{N2}}) \\ &\qquad \geq \operatorname{Var}(\hat{I}_{\mathrm{N3}}). \end{aligned} \tag{6.7}$$

PROOF. See Appendix D.3. □

First let us note that the scheme N3 outperforms (in terms of the variance) any other MIS scheme in the literature that we are aware of. Moreover, for $N = 2$, it also outperforms the other novel schemes R2 and N2. While the MIS schemes R2 and N2 do not appear in Theorem 6.1, we hypothesize that the conclusions of Theorem 6.2 might be extended to $N > 2$. The intuitive reason is that, regardless of N, both methods partially reduce the variance of the estimators by placing more than one proposal at the denominator of some or all the weights. A possible interpretation of the superiority of N3 is that it uses the whole mixture at the denominator of each weight, thus providing an exchange of information between all the proposals. This exchange of information is essential in multimodal scenarios, where the whole set of proposals, seen as a mixture, should mimic the whole target, but each proposal should adapt locally to the target. Since the variance of the IS weight depends on the mismatch of the target (numerator) w.r.t. the proposal (denominator), the use of the whole mixture in the denominator reduces the variance of the weight in general and, therefore, also the variance of the estimator (see the variance analysis in Appendix D). The scheme N3 goes a step further w.r.t. R3, drawing deterministically one sample from each component of $\psi(\mathbf{x})$, which can be seen as drawing N samples from the mixture $\psi(\mathbf{x})$ with a modified version of stratified sampling, a well-known variance reduction technique (see Appendix A and Owen, 2013, Section 9.12), which is also related to the residual resampling.

The variance analysis of the self-normalized estimator $\tilde{I}$ in equation (2.5) implies a ratio of dependent r.v.'s and, therefore, it cannot be performed without resorting to an approximation, for example, by means of a Taylor expansion as it is performed in Kong (1992), Kong, Liu and Wong (1994), Owen (2013). In this case, the bias of $\tilde{I}$ is usually considered negligible compared to the variance for large N. With this approximation, the variance depends on the variances of the numerator (which is a scaled version of $\hat{I}$), the variance of $\hat{Z}$, and the covariance of both. Therefore, the variance results that we have proved above for $\hat{I}$ and $\hat{Z}$, cannot be directly extrapolated for $\tilde{I}$. However, it is reasonable to assume that methods that reduce the variance of $\hat{I}$ and $\hat{Z}$, in general will also reduce the variance of $\tilde{I}$. In Section 8, this hypothesis is reinforced by means of numerical simulations. Therefore, N3 should always be used whenever possible (it requires extra proposal evaluations). See a detailed discussion in Section 7.

Available proposals		(1 2 3)	(1 2 3)	(1 2 3)
Sampling	j_n	3	3	1
	$\mathbf{x}_n$	$\mathbf{x}_1 \sim q_3$	$\mathbf{x}_2 \sim q_3$	$\mathbf{x}_3 \sim q_1$
Weighting options $w_n = \frac{\pi(\mathbf{x})}{\varphi_n(\mathbf{x})}$	R1	$\frac{\pi(\mathbf{x}_1)}{q_3(\mathbf{x}_1)}$	$\frac{\pi(\mathbf{x}_2)}{q_3(\mathbf{x}_2)}$	$\frac{\pi(\mathbf{x}_3)}{q_1(\mathbf{x}_3)}$
	R2	$\frac{\pi(\mathbf{x}_1)}{\frac{1}{3}(q_3(\mathbf{x}_1)+q_3(\mathbf{x}_1)+q_1(\mathbf{x}_1))}$	$\frac{\pi(\mathbf{x}_2)}{\frac{1}{3}(q_3(\mathbf{x}_2)+q_3(\mathbf{x}_2)+q_1(\mathbf{x}_2))}$	$\frac{\pi(\mathbf{x}_3)}{\frac{1}{3}(q_3(\mathbf{x}_3)+q_3(\mathbf{x}_3)+q_1(\mathbf{x}_3))}$
	R3	$\frac{\pi(\mathbf{x}_1)}{\frac{1}{3}(q_1(\mathbf{x}_1)+q_2(\mathbf{x}_1)+q_3(\mathbf{x}_1))}$	$\frac{\pi(\mathbf{x}_2)}{\frac{1}{3}(q_1(\mathbf{x}_2)+q_2(\mathbf{x}_2)+q_3(\mathbf{x}_2))}$	$\frac{\pi(\mathbf{x}_3)}{\frac{1}{3}(q_1(\mathbf{x}_3)+q_2(\mathbf{x}_3)+q_3(\mathbf{x}_3))}$

(a) Schemes R1, R2, and R3

Available proposals		(1 2 3)	(1 2)	(2)
Sampling	j_n	3	1	2
	$\mathbf{x}_n$	$\mathbf{x}_1 \sim q_3$	$\mathbf{x}_2 \sim q_1$	$\mathbf{x}_3 \sim q_2$
Weighting $w_n = \frac{\pi(\mathbf{x})}{\varphi_n(\mathbf{x})}$	N2	$\frac{\pi(\mathbf{x}_1)}{\frac{1}{3}(q_1(\mathbf{x}_1)+q_2(\mathbf{x}_1)+q_3(\mathbf{x}_1))}$	$\frac{\pi(\mathbf{x}_2)}{\frac{1}{2}(q_1(\mathbf{x}_2)+q_2(\mathbf{x}_2))}$	$\frac{\pi(\mathbf{x}_3)}{q_2(\mathbf{x}_3)}$

(b) Scheme N2

Available proposals		(1)	(2)	(3)
Sampling	j_n	1	2	3
	$\mathbf{x}_n$	$\mathbf{x}_1 \sim q_1$	$\mathbf{x}_2 \sim q_2$	$\mathbf{x}_3 \sim q_3$
Weighting options $w_n = \frac{\pi(\mathbf{x})}{\varphi_n(\mathbf{x})}$	N1	$\frac{\pi(\mathbf{x}_1)}{q_1(\mathbf{x}_1)}$	$\frac{\pi(\mathbf{x}_2)}{q_2(\mathbf{x}_2)}$	$\frac{\pi(\mathbf{x}_3)}{q_1(\mathbf{x}_3)}$
	N3	$\frac{\pi(\mathbf{x}_1)}{\frac{1}{3}(q_1(\mathbf{x}_1)+q_2(\mathbf{x}_1)+q_3(\mathbf{x}_1))}$	$\frac{\pi(\mathbf{x}_2)}{\frac{1}{3}(q_1(\mathbf{x}_2)+q_2(\mathbf{x}_2)+q_3(\mathbf{x}_2))}$	$\frac{\pi(\mathbf{x}_3)}{\frac{1}{3}(q_1(\mathbf{x}_3)+q_2(\mathbf{x}_3)+q_3(\mathbf{x}_3))}$

(c) Schemes N1 and N3

FIG. 3. (a) *Example of a realization of the indexes selection* ($N = 3$) *with the sampling procedure* $\mathcal{S}_1$ (*with replacement*), *and all weighting possibilities*, *yielding the MIS schemes* R1, R2 *and* R3. (b) *Example of a realization of the indexes selection* ($N = 3$) *with the sampling procedure* $\mathcal{S}_2$ (*without replacement*) *yielding the MIS scheme* N2. (c) *Sampling procedure* $\mathcal{S}_3$ (*deterministic index selection*), *and all weighting possibilities*, *yielding the MIS schemes* N1 *and* N3.

6.1 Running Example: Exact Variances of the MIS Estimators of Z

Here we focus on computing the exact variances of estimators related to the running example. We simplify the case study to $N = 2$ proposals, for the sake for conciseness in the proofs. The proposal p.d.f.'s are then $q_1(x) = \mathcal{N}(x; \mu_1, \sigma^2)$ and $q_2(x) = \mathcal{N}(x; \mu, \sigma^2)$ with means $\mu_1 = -3$ and $\mu_2 = 3$, and variance $\sigma^2 = 1$. We consider a normalized bimodal target p.d.f. $\pi(x) = \frac{1}{2}\mathcal{N}(x; \nu_1, c_1^2) + \frac{1}{2}\mathcal{N}(x; \nu_2, c_2^2)$ and set $\nu_1 = \mu_1$, $\nu_2 = \mu_2$ and $c_1^2 = c_2^2 = \sigma^2$. Then both proposal p.d.f.'s can

be seen as a whole mixture that exactly replicates the target, that is, $\pi(x) = \frac{1}{2}q_1(x) + \frac{1}{2}q_2(x)$. This is the desired situation pursued by an AIS algorithm: Each proposal is centered at each target mode, and the scale parameters perfectly match the scale of the modes. The goal consists in estimating the normalizing constant with the six schemes described in Section 5. We use the $\hat{Z}$ estimator of equation (2.6) and the estimator $\hat{I}$ of (2.4) when $g = x$, both with $N = 2$ samples. The closed-form variance expressions of the six schemes are presented in the following:

The variances of the estimators of the normalizing constant (true value $Z = \int \pi(x)\,dx = 1$) are given by

$$\begin{aligned}\mathrm{Var}(\hat{Z}_{\mathtt{R1}}) = \mathrm{Var}(\hat{Z}_{\mathtt{N1}}) &= \frac{3 + \exp\left(\frac{4\mu^2}{\sigma^2}\right)}{8} - \frac{1}{2} \\ &= \frac{\exp(36) - 1}{8} \approx 5.4 \cdot 10^{14},\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(\hat{Z}_{\mathtt{R2}}) = \mathrm{Var}(\hat{Z}_{\mathtt{N2}}) &= \frac{3 + \exp\left(\frac{4\mu^2}{\sigma^2}\right)}{16} - \frac{1}{4} \\ &= \frac{\exp(36) - 1}{16} \approx 2.7 \cdot 10^{14},\end{aligned}$$

and

$$\mathrm{Var}(\hat{Z}_{\mathtt{R3}}) = \mathrm{Var}(\hat{Z}_{\mathtt{N3}}) = 0.$$

The variances of the estimators of the target mean (true value $I = \int x\pi(x)\,dx = 0$) are given by

$$\begin{aligned}\mathrm{Var}(\hat{I}_{\mathtt{R1}}) = \mathrm{Var}(\hat{I}_{\mathtt{N1}}) &= \frac{3(\sigma^2 + \mu^2)}{8} \\ &\quad + \frac{\sigma^2 + 9\mu^2}{8} \exp\left(\frac{4\mu^2}{\sigma^2}\right) \\ &= \frac{30}{8} + \frac{82}{8}\exp(36) \approx 4.42 \cdot 10^{16},\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(\hat{I}_{\mathtt{R2}}) = \mathrm{Var}(\hat{I}_{\mathtt{N2}}) &= \frac{3(\sigma^2 + \mu^2)}{16} \\ &\quad + \frac{\sigma^2 + 9\mu^2}{16} \exp\left(\frac{4\mu^2}{\sigma^2}\right) + \frac{\sigma^2}{4} \\ &= \frac{30}{16} + \frac{82}{16}\exp(36) + \frac{1}{4} \approx 2.21 \cdot 10^{16},\end{aligned}$$

$$\mathrm{Var}(\hat{I}_{\mathtt{R3}}) = \frac{\sigma^2 + \mu^2}{2} = 5$$

and

$$\mathrm{Var}(\hat{I}_{\mathtt{N3}}) = \frac{\sigma^2}{2} = \frac{1}{2}.$$

The derivations can be found in Appendix D.4. We observe that, for a very simple bimodal scenario where the proposals are perfectly placed in the target modes, the schemes `R3` and `N3` present a good performance while the other schemes do not work.

7. APPLYING THE MIS SCHEMES

7.1 Computational Complexity

Table 5 compares the total number of target and proposal evaluations in each MIS scheme. First note that the estimators of any MIS scheme within the proposed general framework perform N target evaluations in total. However, depending on the function $\varphi_{\mathcal{P}_n}$ used by each specific scheme at the weight denominator, a different number of proposal evaluations is performed. We see that `R3` and `N3` always require the largest number of proposal evaluations. In `R2`, the number of proposal evaluations is variable: although each weight evaluates N proposals, some proposals may be repeated, whereas others may not be used.

In many relevant scenarios, the cost of evaluating the proposal densities is negligible compared to the cost of evaluating the target function. In this scenario, the MIS scheme `N3` should always be chosen, since it yields a lower variance with a negligible increase in computational cost. For instance, this is the case in the *Big Data* Bayesian framework, where the target function is a posterior distribution with a large amount of data in the likelihood function. However, in some other scenarios, for example, when the number of proposals N is too large and/or the target evaluations are not very expensive, limiting the number of proposal evaluations can result in a better cost-performance trade off.

Unlike most MCMC methods, several strategies of parallelization can be applied in IS-based techniques. In the adaptive context, the adaptation of all proposals usually depends on the performance of all previous proposals and, therefore, the adaptivity is the bottleneck of the parallelization. The six schemes proposed in this paper can be parallelized to some extent. Once all the proposals are available, the schemes `R1`, `N1`, `R3` and `N3` can draw and weight the N samples in parallel, which represents a large advantage w.r.t. MCMC methods. In the schemes `R2` and `N2`, the samples can be drawn independently, but the denominator of the weight cannot be computed in a parallel way. However, since the target evaluation in the numerator of the weights is fully parallelizable, the drawback of these schemes can be considered negligible for a small/medium number of proposals.

TABLE 5
Number of target and proposal evaluations for the different MIS schemes. Note that the number of proposal evaluations for R2 is a random variable with a range from N to N^2

MIS Scheme	R1	N1	R2	N2	R3	N3
Target Evaluations	N	N	N	N	N	N
Proposal Evaluations	N	N	$\leq N^2$	$N(N+1)/2$	N^2	N^2

7.2 A Priori Partition Approach

The extra computational cost of some MIS schemes occurs because each sample must be evaluated in more than one proposal q_n, or even in all of the available proposals (e.g., the MIS scheme N3). In order to limit the number of proposal evaluations, let us first define a partition of the set of the indexes of all proposals, $\{1, \dots, N\}$, into P disjoint subsets of L elements (indexes), $\mathcal{J}_p$ with $p = 1, \dots, P$, s.t.

$$\mathcal{J}_1 \cup \mathcal{J}_2 \cup \cdots \cup \mathcal{J}_P = \{1, \dots, N\}, \tag{7.1}$$

where $\mathcal{J}_k \cap \mathcal{J}_q = \varnothing$ for all $k, q = 1, \dots, P$ and $k \neq q$.[7] Therefore, each subset, $\mathcal{J}_p = \{j_{p,1}, j_{p,2}, \dots, j_{p,L}\}$, contains L indexes, $j_{p,\ell} \in \{1, \dots, N\}$ for $\ell = 1, \dots, L$ and $p = 1, \dots, P$.

After this *a priori* partition, one could apply any MIS scheme in each (partial) subset of proposals, and then perform a suitable convex combination of the partial estimators. This general strategy is inspired by a specific scheme, partial deterministic mixture MIS (p-DM-MIS), which was recently proposed in Elvira et al. (2015). That work applies the idea of the partitions just for the MIS scheme N3, denoted there as full deterministic mixture MIS (f-DM-MIS). The sampling procedure is then $\mathcal{S}_3$, that is, exactly one sample is drawn from each proposal. The weight of each sample in p-DM-MIS, instead of evaluating the whole set of proposals (as in N3), evaluates only the proposals within the subset that the generating proposal belongs to. Mathematically, the weights of the samples corresponding to the pth mixture are computed as

$$w_n = \frac{\pi(\mathbf{x}_n)}{\psi_p(\mathbf{x}_n)} = \frac{\pi(\mathbf{x}_n)}{\frac{1}{L}\sum_{j \in \mathcal{J}_p} q_j(\mathbf{x}_n)}, \quad n \in \mathcal{J}_p. \tag{7.2}$$

Note that the number of proposal evaluations is $N \leq \frac{N^2}{P} \leq N^2$. Specifically, we have the particular cases $P = 1$ and $P = N$ corresponding to the MIS schemes N3 (best performance) and N1 (worst performance), respectively. In Elvira et al. (2015), it is proved that for a specific partition with P subsets of proposals, merging any pair of subsets decreases the variance of the estimator $\hat{I}$ of equation (2.4).

The previous idea can be applied to the other MIS schemes presented in Section 5 (not only N3). In particular, one can make an *a priori* partition of the proposals as in equation (7.1), and apply independently any different MIS scheme in each set. For instance, and based on some knowledge about the performance of the different proposals, one could make two disjoint sets of proposals, applying the MIS scheme N1 in the first set, and the MIS scheme N3 in the second set. Recently, a novel partition approach has been proposed in Elvira et al. (2016). In this case, the sets of proposals are performed *a posteriori*, once the samples have been drawn. The variance of the estimators is reduced at the price of introducing a bias.

7.3 Generalized Adaptive Multiple Importance Sampling

Adaptive importance sampling (AIS) methods iteratively update the parameters of the proposal p.d.f.'s using the information of the past samples (see a survey in Bugallo et al., 2017). The sampling and weighting options, described in this work within a static framework for the sake of simplicity, can be straightforwardly applied in the adaptive context. Let us consider a set of proposal p.d.f.'s $\{q_{j,t}(\mathbf{x})\}$, with $j = 1, \dots, J$ and $t = 1, \dots, T$, where the subscript t indicates the iteration index of the adaptive algorithm, T is the total number of adaptation steps, J is the number of proposals per iteration and $N = JT$ is the total number of proposal p.d.f.'s. A general adaptation procedure takes into account, at the tth iteration, statistical information about the target p.d.f. gathered in all of the previous iterations. Several algorithms have been proposed in the last decade (Cappé et al., 2008; Cornuet et al., 2012; Martino et al., 2015a, 2017; Elvira et al., 2017).

[7] Note that, for the sake of simplifying the notation, we assume that all P subsets have the same number of elements. However, this is not necessary, and it is straightforward to extend the conclusions of this section to the case where each subset has different number of elements.

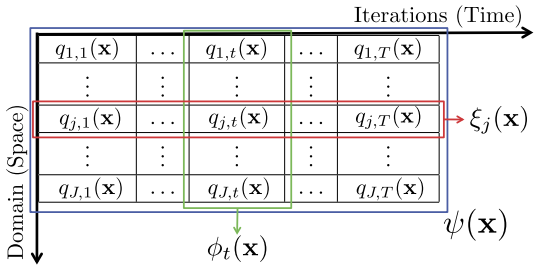


FIG. 4. *Graphical representation of the $N = JT$ proposal p.d.f.'s used in the generalized adaptive MIS scheme, spread through the state space $\mathbb{R}^{d_x}$ ($j = 1, \ldots, J$) and adapted over time ($t = 1, \ldots, T$).*

The MIS schemes considered in Section 5 can be directly applied to the adaptive context. Moreover, the *a priori* partition approach of Section 7.2 can be very useful to limit the computational cost of the different MIS schemes when the number of iterations grows (and therefore also the total number of proposals).

Let us assume that, at the tth iteration, one sample is drawn from each proposal $q_{j,t}$ (sampling $\mathcal{S}_3$), that is,

$$\mathbf{x}_{j,t} \sim q_{j,t}(\mathbf{x}_{j,t}),$$

$j = 1, \ldots, J$ and $t = 1, \ldots, T$. Then an importance weight $w_{j,t}$ is assigned to each sample $\mathbf{x}_{j,t}$. As described exhaustively in Section 4, several strategies can be applied to build $w_{j,t}$ considering the different MIS approaches. Figure 4 provides a graphical representation of this scenario, by showing both the spatial and temporal evolution of the $J = NT$ proposal p.d.f.'s. In a generic AIS algorithm, one weight

$$w_{j,t} = \frac{\pi(\mathbf{x}_{j,t})}{\varphi_{j,t}(\mathbf{x}_{j,t})}, \tag{7.3}$$

is associated to each sample $\mathbf{x}_{j,t}$. In the MIS scheme `N1`, the function employed in the denominator is

$$\varphi_{j,t}(\mathbf{x}) = q_{j,t}(\mathbf{x}). \tag{7.4}$$

In the following we focus on the MIS scheme `N3` in the adaptive framework, considering several choices of the partitioning of the set of proposals, since this scheme attains the best performance, as shown in Section 6. In the *full* `N3` scheme, the function $\varphi_{j,t}$ is

$$\varphi_{j,t}(\mathbf{x}) = \psi(\mathbf{x}) = \frac{1}{JT}\sum_{k=1}^{J}\sum_{r=1}^{T} q_{k,r}(\mathbf{x}), \tag{7.5}$$

where $\psi(\mathbf{x})$ is now the mixture of all the spatial and temporal proposal p.d.f.'s. This case corresponds to the blue rectangle in Figure 4. However, note that the computational complexity can become prohibitive as the product JT increases. Furthermore, two natural alternatives of partial `N3` schemes appear in this scenario. The first one uses the following partial mixture:

$$\varphi_{j,t}(\mathbf{x}) = \xi_j(\mathbf{x}) = \frac{1}{T}\sum_{r=1}^{T} q_{j,r}(\mathbf{x}), \tag{7.6}$$

with $j = 1, \ldots, J$, as mixture-proposal p.d.f. in the IS weight denominator, that is, using the temporal evolution of the jth single proposal $q_{j,t}$ at the weight denominator. In this case, there are $P = J$ mixtures, each one formed by $L = T$ components (red rectangle in Figure 4). Another possibility is considering the mixture of all the $q_{j,t}$'s at the tth iteration, that is,

$$\varphi_{j,t}(\mathbf{x}) = \phi_t(\mathbf{x}) = \frac{1}{J}\sum_{k=1}^{J} q_{k,t}(\mathbf{x}), \tag{7.7}$$

with $t = 1, \ldots, T$, so that we have $P = T$ mixtures, each one formed by $L = J$ components (green rectangle in Figure 4). The function $\varphi_{j,t}$ in equation (7.4) is used in the standard PMC scheme (Cappé et al., 2004), equation (7.6), in the particular case of $J = 1$, has been considered in the *adaptive multiple importance sampling* (AMIS) algorithm (Cornuet et al., 2012). Note that the schemes that consider at the denominator of the weight the temporal sequence of adapted proposals can introduce a bias in the IS estimators (see Cornuet et al., 2012, Section 5 for more details). The choice in equation (7.7) has been applied in the *adaptive population importance sampling* (APIS) (Martino et al., 2015a,) the *layered adaptive importance sampling* (LAIS) (Martino et al., 2017), and the *deterministic mixture population Monte Carlo* (DM-PMC) (Elvira et al., 2017) algorithms. In other techniques, such as mixture PMC (M-PMC) (Douc et al., 2007a, 2007b, Cappé et al., 2008), a similar strategy is employed, but using sampling $\mathcal{S}_1$ in the mixture $\phi_t(\mathbf{x})$, that is, with the MIS scheme `R3`.

Table 6 summarizes all the possible cases discussed above. The last row corresponds to a generic grouping strategy of the proposal p.d.f.'s $q_{j,t}$. As previously described, we can also divide the $N = JT$ proposals into $P = \frac{JT}{L}$ disjoint groups of P mixtures with L components. Namely, we denote the set of L pairs of indexes corresponding to the pth mixture ($p = 1, \ldots, P$) as $\mathcal{J}_p = \{(k_{p,1}, r_{p,1}), \ldots, (k_{p,L}, r_{p,L})\}$, where $k_{p,\ell} \in \{1, \ldots, J\}$, $r_{\ell,p} \in \{1, \ldots, T\}$ (i.e., $|\mathcal{J}_p| = L$, with each element being a pair of indexes), and $\mathcal{J}_p \cap \mathcal{J}_q = \varnothing$ for any pair $p, q = 1, \ldots, P$, and $p \neq q$. In this scenario

TABLE 6
Summary of possible MIS strategies in an adaptive framework

MIS scheme	Function $\varphi_{j,t}(\mathbf{x})$	N	$LP=N$ P	$LP=N$ L	Corresponding algorithm
N1	$q_{j,t}(\mathbf{x})$	JT	JT	1	PMC (Cappé et al., 2004)
Full N3	$\psi(\mathbf{x}) = \frac{1}{JT}\sum_{j=1}^{J}\sum_{t=1}^{T} q_{j,t}(\mathbf{x})$	JT	1	JT	suggested in Elvira et al. (2015)
Partial (temporal) N3	$\xi_j(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} q_{j,t}(\mathbf{x})$	JT	J	T	AMIS (Cornuet et al., 2012), with $J=1$
Partial (spatial) N3	$\phi_t(\mathbf{x}) = \frac{1}{J}\sum_{j=1}^{J} q_{j,t}(\mathbf{x})$	JT	T	J	APIS (Martino et al., 2015a)
Partial (spatial) R3	$\phi_t(\mathbf{x}) = \frac{1}{J}\sum_{j=1}^{J} q_{j,t}(\mathbf{x})$	JT	T	J	Cappé et al. (2008), Douc et al. (2007a, 2007b)
Partial (generic) N3	generic $\varphi_{j,t}(\mathbf{x})$ in equation (7.8)	JT	P	L	suggested in Elvira et al. (2015)

we have

$$\varphi_{j,t}(\mathbf{x}) = \frac{1}{L}\sum_{(k,r)\in\mathcal{J}_p} q_{k,r}(\mathbf{x}) \quad \text{with } (j,t)\in\mathcal{J}_p. \tag{7.8}$$

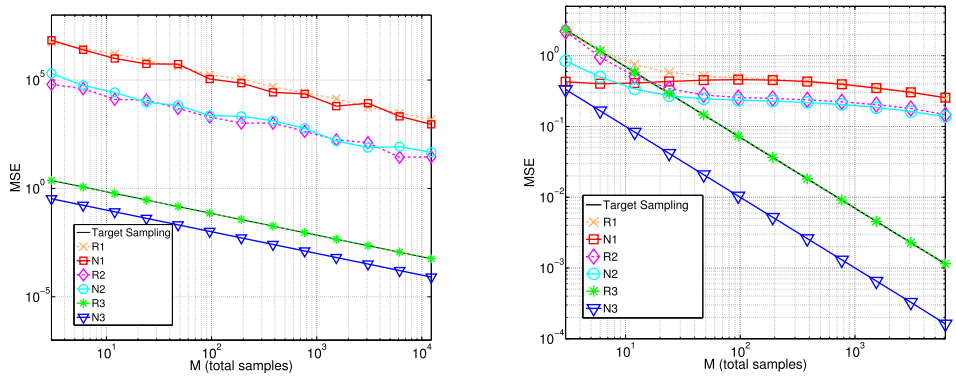
Note that using $\psi(\mathbf{x})$ and $\xi_j(\mathbf{x})$ the computational cost per iteration increases as the total number of iterations T grows. Indeed, at the tth iteration all the previous proposals $q_{j,1},\ldots,q_{j,t-1}$ (for all j) must be evaluated at all the new samples $\mathbf{x}_{j,t}$. Hence, algorithms based on these proposals quickly become unfeasible as the number of iterations grows. On the other hand, using $\phi_t(\mathbf{x})$ the computational cost per iteration is controlled by J, remaining constant regardless of the number of adaptive steps performed.

7.4 Guidelines for Applying MIS

The superiority of N3 is theoretically proved for the unnormalized estimator in Theorems 6.1 and 6.2, and practically shown by means of several numerical simulations for the self-normalized estimator (see next section). However, the associated computational complexity is also increased w.r.t. the other MIS schemes in terms of proposal evaluations. If N is small or the target evaluations are expensive (w.r.t. the cost of the proposal evaluations), N3 should be used. However, when the target evaluation is cheap and/or the number of proposals is large, the use of N3 increases notably the computational complexity. In this case, the novel schemes R2 or N2 seem to provide very good results, and their theoretical properties are superior to those of N1 and R1. However, future studies will be required to characterize these novel schemes and investigate efficient parallelization techniques. We also recommend to combine adaptive schemes with the partition approach proposed in Elvira et al. (2015, 2016), and summarized in Section 7.2. Note that further investigation is also needed for efficiently constructing the partitions of the proposals that allow to reduce the computational complexity while retaining most of the variance reduction associated to the N3 scheme.

In the adaptive context, there is a big potential for the MIS schemes where all spatial and temporal proposals are used at the denominator of all weights (blue square in Figure 4). However, the computational complexity for large number of proposals is prohibitive, and further theoretical analysis about the bias of the estimators is needed (see Cornuet et al., 2012, Section 5). The adaptivity of MIS algorithms is essential in challenging high-dimensional setups. The N3 scheme has exhibited a very good performance when used within adaptive MIS algorithms due to two main reasons. First, the variance of the estimators at each iteration is reduced as proved in Theorems 6.1 and 6.2, which explains part of the variance reduction attained in AMIS (Cornuet et al., 2012), LAIS (Martino et al., 2017) or GAPIS (Elvira et al., 2015b). Second, when the IS weights are used for adaptive purposes (e.g., in APIS (Martino et al., 2015a) or DM-PMC (Elvira et al., 2017)), the use of the whole mixture of proposals in the denominator of the weights can be seen as a cooperative adaptive procedure (see Elvira et al., 2017 for further details).

Finally, one of the strengths of the N3 scheme is its performance in multimodal scenarios, where N1 should always be avoided. If N is comparable to the number of modes, an adaptive N3 scheme should be employed; the aforementioned cooperation in the proposals adaptation has an implicit repulsive behavior that promotes the adaptation to different modes. However, if N is much larger, the adaptive algorithm may use R2 or N2 with potentially similar performance but less computational complexity.



(a) MSE of $\hat{I}$ (unnormalized weights). (b) MSE of $\tilde{I}$ (normalized weights).

FIG. 5. (Ex. of Section. 8.1) *Performance of the estimators of the target mean for the different MIS schemes.*

8. NUMERICAL EXAMPLES

In the previous sections we have provided several theoretical results for comparing different MIS schemes according to different quality measures, for example, ranking them in terms of the variance of the corresponding estimators. In this section, we provide different numerical results in order to quantify numerically the gap among these methods. In the following, we show that even in the case where the different proposals are well tuned (in the sense of a small or no mismatch with a multimodal target), the choices of the sampling and weighting procedures dramatically affect the performance of the MIS estimator.

8.1 Running Exmple: Estimation of the Target Mean

Let us consider again the target p.d.f. of the running example

$$\pi(\mathbf{x}) = \frac{1}{3}\mathcal{N}(x; \nu_1, c_1^2) + \frac{1}{3}\mathcal{N}(x; \nu_2, c_2^2) + \frac{1}{3}\mathcal{N}(x; \nu_3, c_3^2),$$

with means $\nu_1 = -3$, $\nu_2 = 0$, and $\nu_3 = 3$, and variances $c_1^2 = c_2^2 = c_3^2 = 1$. As proposal functions, we use $q_i(x) = \mathcal{N}(x; \mu_i, \sigma)$, with $\mu_i = \nu_i$ and $i = 1, 2, 3$ and $\sigma^2 = 1$, that is, the proposal p.d.f.'s can be seen as a whole mixture that exactly replicates the target, that is, $\pi(x) = \psi(x) = \frac{1}{3}q_1(x) + \frac{1}{3}q_2(x) + \frac{1}{3}q_3(x)$.

The goal is to estimate the mean of the target p.d.f. with the six MIS schemes. Figure 5(a) shows the MSE of the estimator $\hat{I}$ for all the methods w.r.t. the number of total samples (note that some schemes require that the total number of samples is multiple of $M = 3$). The results have been averaged over $5 \cdot 10^6$ runs. The solid black line shows the variance of the natural estimator, that is, sampling directly from the target p.d.f. (since this is possible in this easy example). Note that the method $\hat{I}_{\mathtt{R3}}$ exactly replicates the performance of $\bar{I}$: this method samples from the mixture of Gaussians in the traditional way and the weights, due to the perfect match, are always $w = 1$, that is, $\hat{I}_{\mathtt{R3}}$ and $\bar{I}$ are equivalent. We can see that $\hat{I}_{\mathtt{N3}}$ is the best estimator in terms of variance, while $\hat{I}_{\mathtt{R1}}$ and $\hat{I}_{\mathtt{N1}}$ present a high variance. Note that, surprisingly, $\hat{I}_{\mathtt{N3}}$ has better performance than sampling from the target, that is, estimator $\bar{I}$. This is because the sampling $\mathcal{S}_3$ can be seen as a sampling from the mixture of proposals $\psi(\mathbf{x})$ (which coincides with the target in this example) with a variance reduction technique, as we discuss in Appendix A. Note also that the inequality proved in Theorem 6.1 holds since all methods are unbiased and, therefore, the MSE is due only to the variance. We can see that $\hat{I}_{\mathtt{R2}}$ and $\hat{I}_{\mathtt{N2}}$ also behave badly in terms of variance.

Figure 5(b) shows the variance of the estimator $\tilde{I}$ of equation (2.5) for all methods. First note that the MSE of `R3` and `N3` is the same as in Figure 5(b), since the estimators $\hat{I}$ and $\tilde{I}$ are equivalent in this scenario (since they perfectly estimate the normalizing constant, that is, $\hat{Z} = Z$). Note that the relations observed and proved for the different MIS schemes in terms of the variance of the estimator $\hat{I}$, are also kept here when we increase the number of samples. The MSE curves are compared with the same number of samples M, that is, the same number of target evaluations. Note that each MIS scheme requires a different number of proposal evaluations per sample (see Table 5). However, a fair comparison is fully target dependent, and with few number of proposals we can consider that the computational complexity is similar in all schemes.

TABLE 7
(Ex. of Section 8.2.1) *MSE of the LAIS and PMC algorithms with the different MIS schemes at the lower layer. $J = 100$ proposals and $T = 200$ iterations*

Alg.	R1-LAIS	N1-LAIS	R2-LAIS	N2-LAIS	R3-LAIS	N3-LAIS	N1-PMC	N3-PMC
$\text{Var}(\hat{Z})$	0.6471	0.6380	0.0004	0.0024	0.0005	**0.0001**	0.1528	0.0006
$\text{Var}(\tilde{I})$	1.4509	2.0466	0.0335	0.0295	0.0423	**0.0088**	0.3847	0.0363

8.2 Applying the MIS Schemes in Adaptive IS (AIS)

We apply the different MIS schemes within an AIS context. In particular, we focus on the LAIS algorithm, recently proposed in Martino et al. (2017). The method consists of an upper layer with a MCMC that draws samples from the target, while a lower layer uses those samples as location parameters (means) of some proposal p.d.f.'s for applying IS. In its basic version, J Metropolis–Hastings chains independently run at the upper layer, and hence MIS is applied in the lower layer with J proposals at each iteration. In the following, we implement the six adaptive MIS schemes in a spatial manner for two different target p.d.f.'s. For instance, the N3 scheme is implemented by sampling exactly one sample from each of the J proposals at the tth iteration, and applying at the denominator of the IS weight the whole mixture of J proposals as in equation (7.7) (see the green square of Figure 4).

8.2.1 *Mixture of bivariate Gaussians*. Let us first consider a mixture of five bivariate Gaussians,

$$\pi(\mathbf{x}) = \frac{1}{5}\sum_{i=1}^{5}\mathcal{N}(\mathbf{x}; \boldsymbol{\nu}_i, \boldsymbol{\Sigma}_i), \quad \mathbf{x} \in \mathbb{R}^2, \tag{8.1}$$

where $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{C})$ denotes a Gaussian p.d.f. with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$, $\boldsymbol{\nu}_1 = [-10, -10]^\top$, $\boldsymbol{\nu}_2 = [0, 16]^\top$, $\boldsymbol{\nu}_3 = [13, 8]^\top$, $\boldsymbol{\nu}_4 = [-9, 7]^\top$, $\boldsymbol{\nu}_5 = [14, -14]^\top$, $\boldsymbol{\Sigma}_1 = [2, 0.6; 0.6, 1]$, $\boldsymbol{\Sigma}_2 = [2, -0.4; -0.4, 2]$, $\boldsymbol{\Sigma}_3 = [2, 0.8; 0.8, 2]$, $\boldsymbol{\Sigma}_4 = [3, 0; 0, 0.5]$ and $\boldsymbol{\Sigma}_5 = [2, -0.1; -0.1, 2]$. We run the LAIS algorithm with $J = 100$ spatial proposals that are adapted over $T = 200$ iterations. The proposals of the upper and lower layers are isotropic Gaussians with $\sigma_{\text{upper}} = 5$ and $\sigma_{\text{lower}} = 2$, respectively. We also run the standard PMC algorithm of Cappé et al. (2004), computing at each iteration the weights according to N1, which represents the standard PMC, and N3 which corresponds to the DM-PMC algorithm recently proposed in Elvira et al. (2017). The means of the proposals are randomly and uniformly initialized within the $[-4, 4] \times [-4, 4]$ square. Table 7 shows the MSE of the self-normalized estimator of the target mean, $\tilde{I}$, and the estimator of the normalizing constant (the true values are $\text{E}[\mathbf{X}] = [1.6, 1.4]^\top$ and $Z = 1$, resp.). The scheme N3 presents again the best performance in the adaptive setup, both in LAIS and PMC. Note that the novel schemes R2 and N2 show again a satisfactory performance.

8.2.2 *Multidimensional banana-shaped distribution*. We consider the banana shape target example used in Haario, Saksman and Tamminen (1999, 2001) which "can be be calibrated to become extremely challenging" (Cornuet et al., 2012). The target is based on a d_x-dimensional multivariate Gaussian $\mathbf{x} \sim \mathcal{N}(\mathbf{x}; \mathbf{0}_{d_x}, \boldsymbol{\Sigma})$ with $\Sigma = \text{diag}(\sigma^2, 1, \ldots, 1)$, where the second variable is nonlinearly transformed from x_2 to $x_2 - b(x_1^2 - \sigma^2)$. This transformation leads to a banana-shaped distribution with zero mean and uncorrelated components (note that the target dimension $d_x \geq 2$).

We implement the MIS schemes within the LAIS algorithm as described in the previous example. We set $J = 200$ proposals that are adapted over $T = 1000$ iterations, and isotropic Gaussian proposals with $\sigma_{\text{upper}} = 0.2$ and $\sigma_{\text{lower}} = 0.5$. The means of the proposals are randomly and uniformly initialized within the $[-4, 4] \times [-5, 5]$ square. In Figure 6, we vary the dimension of the state space d_x with, $2 \leq d_x \leq 40$, and we show the MSE of the self-normalized estimator $\tilde{I}$ of the target mean. The results have been averaged over 300 runs. We observe that N3 and R3 schemes provide a similar good performance as in previous examples, although if N were smaller, N3 would clearly outperform R3. When the dimension increases, the performance of all schemes decays, but the same hierarchy in performance holds for all schemes. N2 presents a similar performance than N3 in high dimensions.

8.3 Discussion on the Experimental Results

First of all, note that the numerical simulations provided in this section corroborate the variance analysis of Section 6. More specifically, the hierarchy shown in

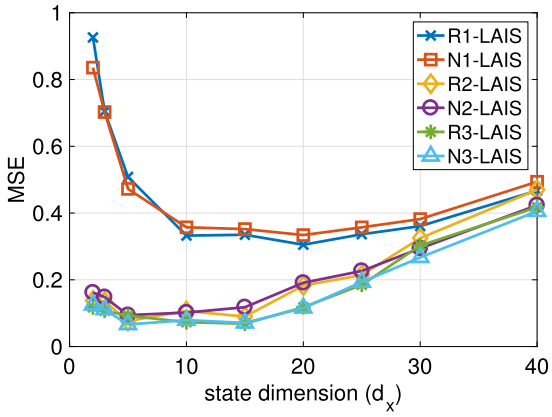


FIG. 6. (Ex. of Section 8.2.2) *LAIS algorithm with different MIS schemes in a multidimensional banana-shaped target.* $J = 100$ *proposals and* $T = 200$ *iterations.*

Figure 5, based on MSE of $\hat{I}$, corresponds to the hierarchy in terms of variance of $\hat{I}$ given in Theorems 6.1 and 6.2. The same hierarchy is represented graphically in Figure 7. Furthermore, Figure 5(b) depicts the MSE of the self-normalized estimator $\tilde{I}$: for large enough values of M (so that a good approximation of Z is attained), the MIS schemes are ordered exactly as in Figure 5 (as discussed in Section 6).

The numerical experiments confirm that N3 provides the best performance. The scheme R3 also presents a good performance in most cases. The performance of R1 and N1 is, in general, much worse than the performance of the other schemes. Both schemes account at the weight denominator only for the proposal from which the sample is drawn, which in a multimodal scenario can be problematic. While R1 is a novel scheme that has naturally arisen in this work, and it probably has little interest from a practical point of view, N1 has been applied in different adaptive MIS algorithms, such as the original version of PMC (Cappé et al., 2004).

The novel schemes R2 and N2 have appeared in this new framework and deserve a further analysis. The hierarchy theoretically proved for $N = 2$ proposals in Theorem 6.2 still holds in the numerical examples for $N > 2$, for example, in Figures 5(a) and 5(b). In some scenarios, for instance where there is a big number of proposals compared to the modes of the target, these schemes can attain most of the variance reduction of N1 and N3 while reducing the number of proposal evaluations w.r.t. N3. In the example with AIS methods, both R2 and N2 present a very competitive performance w.r.t. to N3.

Finally, observe that in Figure 5, when a small number of samples M is employed, the schemes N1, N2 and N3, that is, those with index selection without replacement ($\mathcal{S}_2$ and $\mathcal{S}_3$), behave better. This occurs because the variance associated to the index selection is reduced by guaranteeing that all proposal p.d.f.'s are always used.

9. CONCLUSIONS

In this work we have introduced a unified framework for sampling and weighting in the context of multiple importance sampling (MIS). This approach extends the concept of a proper weighted sample, enabling the design of a wide range of sampling/weighting combinations. In particular, we have considered three specific sampling procedures and we have proposed five types of generic weighting functions (related to different conditional and marginal distributions which depend on the sampling scheme). As a result of the combinations of sampling and weighting procedures, we have analyzed the six unique resulting schemes (three of them are not present in the literature to the best of our knowledge). We have provided a theoretical comparison of these schemes in terms of variance, establishing a ranking of the different methods in terms of performance and computational complexity. Moreover, we have discussed the application of the MIS schemes within adaptive procedures. In addition, we have provided the practitioner with several useful and easy-to-follow guidelines for applying the MIS schemes in different scenarios. We have analyzed the behavior of the MIS schemes in three different numerical examples which corroborate the previous theoretical analysis.

APPENDIX A: FURTHER OBSERVATIONS ABOUT THE SAMPLING $\mathcal{S}_3$

In the sampling procedure $\mathcal{S}_3$, $\mathbf{X}_n \sim q_n(\mathbf{x})$ for $n = 1, \dots, N$, that is, the selection of the index is deterministic. Note that the set of samples $\{\mathbf{x}_n\}_{n=1}^N$ is used in the IS estimators regardless of the order they are drawn. It can be interpreted that the N samples are drawn from the mixture $\psi(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^N q_n(\mathbf{x})$ via Rao–Blackwellization (see Owen, 2013, Section 9.12, for more details). More formally, if we define the r.v. $\mathbf{X} = \mathbf{X}_n$ with $n \sim \mathcal{U}\{1, 2, \dots, N\}$, then $\mathbf{X} \sim \psi(\mathbf{x})$. The procedure $\mathcal{S}_3$ follows a similar principle as a well-known variance reduction method, known as the stratified sampling (Robert and Casella, 2004, Liu, 2008), where the domain of $\mathbf{X}$ is divided into different regions that, in the case of sampling $\mathcal{S}_3$, are unbounded and overlapped (Owen, 2013, Section 9.12). Finally, note

that the approach $\mathcal{S}_3$ can also be seen as the application of a quasi-Monte Carlo technique (Niederreiter, 1992) for generating the deterministic sequence of indexes $j_1 = 1, j_2 = 2, \dots, j_N = N$ (uniform, in the sense of low-discrepancy sequence) and then drawing $\mathbf{x}_n \sim q_{j_n}(\mathbf{x}) = q_n(\mathbf{x})$ for $n = 1, \dots, N$. Note also, that $\mathcal{S}_3$ can be seen as a residual resampling step of the indexes of the proposals. Since all weights of the proposals are the same, the resampling is fully deterministic, which explains part the variance reduction of the MIS schemes with sampling $\mathcal{S}_3$.

APPENDIX B: CONNECTIONS WITH RESAMPLING METHODS

Resampling methods are used in PFs to replace a set of weighted particles with another set of equally weighted particles. The way we address the sampling process in MIS has clear connections with the resampling step in PFs (e.g., see Douc and Cappé, 2005). An important difference of the proposed framework is that the MIS proposals are equally weighted in the mixture. The sampling method $\mathcal{S}_1$ is then equivalent to the multinomial resampling, whereas the sampling methods $\mathcal{S}_2$ and $\mathcal{S}_3$ correspond to residual resampling (note that, since $M = N$ and all the proposals are equally weighted, exactly one sample per proposal is drawn). In future works, it would be interesting to analyze sampling schemes related to residual, stratified and systematic resamplings, which can be incorporated quite naturally in MIS schemes, when the weights of the proposals are different (see, for instance, He and Owen, 2014).

APPENDIX C: PROOFS OF UNBIASEDNESS OF THE MIS ESTIMATORS

In this Appendix, we prove the unbiasedness of the estimator $\hat{I}$ of equation (2.4) for the five weighting options described in Section 4. We recall that the general expression for the expectation of $\hat{I}$ within the proposed framework is

$$\mathrm{E}[\hat{I}] = \frac{1}{ZN} \sum_{n=1}^{N} \sum_{j_{1:N}} \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)} \cdot P(j_{1:N}) p(\mathbf{x}_n | j_n) \, d\mathbf{x}_n. \tag{C.1}$$

OPTION 1 ($\mathcal{W}_1$): $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi_{j_{1:n-1}}(\mathbf{x}_n) = p(\mathbf{x}_n | j_{1:n-1})$. We first marginalize in equation (C.1) over all indexes that do not affect the nth weight ($j_{n:N}$):

$$\begin{aligned} \mathrm{E}[\hat{I}] &= \frac{1}{ZN} \sum_{n=1}^{N} \sum_{j_{1:n-1}} \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_{j_{1:n-1}}(\mathbf{x}_n)} \cdot p(\mathbf{x}_n, j_{1:n-1}) \, d\mathbf{x}_n \\ &= \frac{1}{ZN} \sum_{n=1}^{N} \sum_{j_{1:n-1}} \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_{j_{1:n-1}}(\mathbf{x}_n)} \cdot p(\mathbf{x}_n | j_{1:n-1}) P(j_{1:n-1}) \, d\mathbf{x}_n. \end{aligned} \tag{C.2}$$

Then, substituting $\varphi_{j_{1:n-1}}(\mathbf{x}_n) = p(\mathbf{x}_n | j_{1:n-1})$ into equation (C.2), cancelling terms and marginalizing $j_{1:n-1}$, we have

$$\begin{aligned} \mathrm{E}[\hat{I}] &= \frac{1}{ZN} \sum_{n=1}^{N} \int \pi(\mathbf{x}_n) g(\mathbf{x}_n) \, d\mathbf{x}_n \\ &= \frac{1}{Z} \int \pi(\mathbf{x}) g(\mathbf{x}) \, d\mathbf{x} = I. \end{aligned}$$

OPTION 2 ($\mathcal{W}_2$): $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi_{j_n}(\mathbf{x}_n) = p(\mathbf{x}_n | j_n)$. We substitute $\varphi_{j_n}(\mathbf{x}_n) = p(\mathbf{x}_n | j_n)$ into equation (C.1), which cancels the denominator:

$$\begin{aligned} \mathrm{E}[\hat{I}] &= \frac{1}{ZN} \sum_{n=1}^{N} \sum_{j_{1:N}} \int \pi(\mathbf{x}_n) g(\mathbf{x}_n) P(j_{1:N}) \, d\mathbf{x}_n \\ &= \frac{1}{ZN} \sum_{n=1}^{N} \int \pi(\mathbf{x}_n) g(\mathbf{x}_n) \, d\mathbf{x}_n \\ &= \frac{1}{Z} \int \pi(\mathbf{x}) g(\mathbf{x}) \, d\mathbf{x} = I. \end{aligned}$$

OPTION 3 ($\mathcal{W}_3$): $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi_n(\mathbf{x}_n) = p(\mathbf{x}_n)$. Since φ_n does not depend on any index, we can first marginalize over the whole set of indexes $j_{1:N}$ in equation (C.1):

$$\mathrm{E}[\hat{I}] = \frac{1}{ZN} \sum_{n=1}^{N} \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_n(\mathbf{x}_n)} p(\mathbf{x}_n) \, d\mathbf{x}_n. \tag{C.3}$$

Then, substituting $\varphi_n = p(\mathbf{x}_n)$ in equation (C.3):

$$\begin{aligned} \mathrm{E}[\hat{I}] &= \frac{1}{ZN} \sum_{n=1}^{N} \int \pi(\mathbf{x}_n) g(\mathbf{x}_n) \, d\mathbf{x}_n \\ &= \frac{1}{Z} \int \pi(\mathbf{x}) g(\mathbf{x}) \, d\mathbf{x} = I. \end{aligned}$$

OPTION 4 ($\mathcal{W}_4$): $\varphi_{\mathcal{P}_n}(\mathbf{x}) = \varphi_{j_{1:N}}(\mathbf{x}) = f(\mathbf{x} | j_{1:N}) = \frac{1}{N} \sum_{n=1}^{N} q_{j_n}(\mathbf{x})$. In this case, the expectation of $\hat{I}$ can be expressed as

$$\mathrm{E}[\hat{I}] = \frac{1}{ZN} \sum_{j_{1:N}} P(j_{1:N}) \cdot \int \frac{\pi(\mathbf{x}) g(\mathbf{x})}{\varphi_{j_{1:N}}(\mathbf{x})} \sum_{n=1}^{N} q_{j_n}(\mathbf{x}) \, d\mathbf{x}. \tag{C.4}$$

Substituting $\varphi_{j_{1:N}}(\mathbf{x}) = f(\mathbf{x}|j_{1:N}) = \frac{1}{N}\sum_{n=1}^{N} q_{j_n}(\mathbf{x})$ in equation (C.4), and cancelling the denominator:

$$\begin{aligned}\mathrm{E}[\hat{I}] &= \frac{1}{Z}\sum_{j_{1:N}}\int \pi(\mathbf{x})g(\mathbf{x})P(j_{1:N})\,d\mathbf{x}\\ &= \frac{1}{Z}\int \pi(\mathbf{x})g(\mathbf{x})\,d\mathbf{x} = I.\end{aligned}$$

OPTION 5 ($\mathcal{W}_5$): $\varphi_{\mathcal{P}_n}(\mathbf{x}) = \varphi(\mathbf{x}) = f(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} q_n(\mathbf{x}) = \psi(\mathbf{x})$. Now the expectation of $\hat{I}$ becomes

$$\begin{aligned}\mathrm{E}[\hat{I}] &= \frac{1}{Z}\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\varphi(\mathbf{x})}\\ &\quad\cdot\sum_{j_{1:N}}\left[\frac{1}{N}\sum_{n=1}^{N} q_{j_n}(\mathbf{x})\right]P(j_{1:N})\,d\mathbf{x}\\ &= \frac{1}{Z}\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\varphi(\mathbf{x})}\psi(\mathbf{x})\,d\mathbf{x},\end{aligned} \tag{C.5}$$

where, in the last step, we have used the identity

$$\sum_{j_{1:N}}\left[\frac{1}{N}\sum_{n=1}^{N} q_{j_n}(\mathbf{x})\right]P(j_{1:N}) = f(\mathbf{x}) = \psi(\mathbf{x})$$

for any valid sampling procedure within this framework (see Remark 3.1 and Section 3.5 for more details). Substituting $\varphi(\mathbf{x}) = \psi(\mathbf{x})$ in equation (C.5)

$$\begin{aligned}\mathrm{E}[\hat{I}] &= \frac{1}{Z}\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})}\psi(\mathbf{x})\,d\mathbf{x}\\ &= \frac{1}{Z}\int \pi(\mathbf{x})g(\mathbf{x})\,d\mathbf{x} = I.\end{aligned} \tag{C.6}$$

APPENDIX D: VARIANCE ANALYSIS OF THE MIS ESTIMATORS

Let us consider the unbiased estimator,

$$\hat{I} = \frac{1}{ZN}\sum_{n=1}^{N} w_n(\mathbf{x}_n)g(\mathbf{x}_n), \tag{D.1}$$

that approximates I. The variance of $\hat{I}$ can be expressed in the general form as

$$\begin{aligned}\mathrm{Var}(\hat{I}) &= E_{p(\mathbf{x}_{1:N}, j_{1:N})}\big[(\hat{I} - E_{p(\mathbf{x}_{1:N}, j_{1:N})}[\hat{I}])^2\big]\\ &= E_{p(\mathbf{x}_{1:N}, j_{1:N})}[\hat{I}^2] - E^2_{p(\mathbf{x}_{1:N}, j_{1:N})}[\hat{I}].\end{aligned} \tag{D.2}$$

In the general case of equation (D.2), the N terms of the sum of the estimator in $\hat{I}$ are dependent. However, in the specific cases where they are independent, the variance of a sum of r.v.'s can be simplified as the sum of the variances, that is,

$$\begin{aligned}\mathrm{Var}(\hat{I}) &= \frac{1}{Z^2N^2}\left[\sum_{n=1}^{N} E_{p(\mathbf{x}_n, j_n)}\big[w_n^2(\mathbf{x}_n)g^2(\mathbf{x}_n)\big]\right.\\ &\quad\left.-\sum_{n=1}^{N} E^2_{p(\mathbf{x}_n, j_n)}\big[w_n(\mathbf{x}_n)g(\mathbf{x}_n)\big]\right]\\ &= \frac{1}{Z^2N^2}\left[\sum_{n=1}^{N}\sum_{j_n=1}^{N}\int \frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{\varphi^2_{\mathcal{P}_n}(\mathbf{x}_n)}\right.\\ &\quad\cdot p(\mathbf{x}_n|j_n)P(j_n)\,d\mathbf{x}_n\\ &\quad-\sum_{n=1}^{N}\left(\sum_{j_n=1}^{N}\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}\right.\\ &\quad\left.\left.\cdot p(\mathbf{x}_n|j_n)P(j_n)\,d\mathbf{x}_n\right)^2\right].\end{aligned} \tag{D.3}$$

In some MIS schemes, the N terms are dependent (due to a sampling without replacement or because the nth weight depends on several indexes j_k, with at least one $k \neq n$). However, conditioned to the whole set of indexes $j_{1:N}$, the terms of the sum in equation (D.1) are always conditionally independent, so we can apply

$$\begin{aligned}&\mathrm{Var}(\hat{I})\\ &= \frac{1}{Z^2N^2}\sum_{j_{1:N}}\left[\sum_{n=1}^{N} E_{p(\mathbf{x}_n|j_n)}\big[w_n^2(\mathbf{x}_n)g^2(\mathbf{x}_n)\big]\right.\\ &\quad\left.-\sum_{n=1}^{N} E^2_{p(\mathbf{x}_n|j_n)}\big[w_n(\mathbf{x}_n)g(\mathbf{x}_n)\big]\right]P(j_{1:N})\\ &= \frac{1}{Z^2N^2}\sum_{j_{1:N}}\left[\sum_{n=1}^{N}\int \frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{\varphi^2_{\mathcal{P}_n}(\mathbf{x}_n)}p(\mathbf{x}_n|j_n)\,d\mathbf{x}_n\right.\\ &\quad\left.-\sum_{n=1}^{N}\left(\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}p(\mathbf{x}_n|j_n)\,d\mathbf{x}_n\right)^2\right]\\ &\quad\cdot P(j_{1:N}).\end{aligned} \tag{D.4}$$

D.1 Variance of the Estimators of the MIS Schemes

In the following, we analyze the variance of the six MIS schemes discussed through this paper under the assumptions described in Theorem 6.1 (see Section 6 for more details). Since some schemes arise under more than one sampling/weighting combination (see Table 3), here we always use the combination that facilitates the analysis.

1. [R1] *Sampling 1/Weighting 2*: In this scheme, all the terms of the sum in equation (D.1) are independent, so we can use equation (D.3) for computing the variance of $\hat{I}$. Substituting $\varphi_{j_n}(\mathbf{x}_n) = p(\mathbf{x}_n|j_n) = q_{j_n}(\mathbf{x}_n)$ in D.3,

$$\begin{aligned}
&\mathrm{Var}(\hat{I}_{\mathtt{R1}}) \\
&\quad= \frac{1}{Z^2N^2}\sum_{n=1}^{N}\sum_{j_n=1}^{N}\Bigg[\int \frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{p^2(\mathbf{x}_n|j_n)} \\
&\qquad\cdot p(\mathbf{x}_n|j_n)P(j_n)\,d\mathbf{x}_n\Bigg] - \frac{I^2}{N} \\
&\quad= \frac{1}{Z^2N^2}\sum_{n=1}^{N}\Bigg[\int \sum_{j_n=1}^{N}\frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{q_{j_n}(\mathbf{x}_n)} \\
&\qquad\cdot P(j_n)\,d\mathbf{x}_n\Bigg] - \frac{I^2}{N} \\
&\quad= \frac{1}{Z^2N^2}\sum_{n=1}^{N}\Bigg[\int \frac{1}{N}\sum_{k=1}^{N}\frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{q_k(\mathbf{x}_n)}\,d\mathbf{x}_n\Bigg] \\
&\qquad- \frac{I^2}{N} \\
&\quad= \frac{1}{Z^2N^2}\sum_{k=1}^{N}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_k(\mathbf{x})}\,d\mathbf{x} - \frac{I^2}{N},
\end{aligned} \tag{D.5}$$

were we have used that $P(j_n) = \frac{1}{N}, \forall j_n \in \{1,\dots,N\}$.

2. [R2] *Sampling 1/Weighting 4*: The expression for the conditional independence of equation (D.4) is used substituting $\varphi_{j_{1:N}}(\mathbf{x}_n) = f(\mathbf{x}_n|j_{1:N}) = \frac{1}{N}\sum_{k=1}^{N} q_{j_k}(\mathbf{x}_n)$ and averaging it over the N^N equiprobable sequences of indexes $j_{1:N}$:

$$\begin{aligned}
&\mathrm{Var}(\hat{I}_{\mathtt{R2}}) \\
&\quad= \frac{1}{Z^2N^2}\Bigg[\sum_{j_{1:N}}\Bigg[\sum_{n=1}^{N}\int \frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{\varphi^2_{j_{1:N}}(\mathbf{x}_n)}p(\mathbf{x}_n|j_n)\,d\mathbf{x}_n \\
&\qquad- \sum_{n=1}^{N}\Bigg(\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi_{j_{1:N}}(\mathbf{x}_n)}p(\mathbf{x}_n|j_n)\,d\mathbf{x}_n\Bigg)^2\Bigg] \\
&\qquad\cdot P(j_{1:N})\Bigg] \\
&\quad= \frac{1}{Z^2N^2}\frac{1}{N^N}\Bigg[\sum_{j_{1:N}}\sum_{n=1}^{N}\int \frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{f^2(\mathbf{x}_n|j_{1:N})}q_{j_n}(\mathbf{x}_n)\,d\mathbf{x}_n \\
&\qquad- \sum_{j_{1:N}}\sum_{n=1}^{N}\Bigg(\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{f(\mathbf{x}_n|j_{1:N})}q_{j_n}(\mathbf{x}_n)\,d\mathbf{x}_n\Bigg)^2\Bigg] \\
&\quad= \frac{1}{Z^2N}\frac{1}{N^N}\Bigg[\sum_{j_{1:N}}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{f^2(\mathbf{x}|j_{1:N})}\Bigg(\frac{1}{N}\sum_{n=1}^{N}q_{j_n}(\mathbf{x})\Bigg)d\mathbf{x} \\
&\qquad- \frac{1}{N}\sum_{j_{1:N}}\sum_{n=1}^{N}\Bigg(\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{f(\mathbf{x}_n|j_{1:N})}q_{j_n}(\mathbf{x}_n)\,d\mathbf{x}_n\Bigg)^2\Bigg] \\
&\quad= \frac{1}{Z^2N}\frac{1}{N^N}\Bigg[\sum_{j_{1:N}}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{f(\mathbf{x}|j_{1:N})}\,d\mathbf{x} \\
&\qquad- \frac{1}{N}\sum_{j_{1:N}}\sum_{n=1}^{N}\Bigg(\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{f(\mathbf{x}_n|j_{1:N})}q_{j_n}(\mathbf{x}_n)\,d\mathbf{x}_n\Bigg)^2\Bigg],
\end{aligned} \tag{D.6}$$

where we have used the identity $f(\mathbf{x}|j_{1:N}) = \frac{1}{N}\sum_{n=1}^{N} q_{j_n}(\mathbf{x}_n)$. This expression for the variance resembles that of scheme [N3], averaged over the N^N possible mixtures (combinations) that can arise with sampling $\mathcal{S}_1$.

3. [R3] *Sampling 1/Weighting 3*: All the elements are independent in the sum, and the weights do not depend on any index of the set $j_{1:N}$. Therefore, we can start with equation (D.3), marginalize over the indexes and substitute $\varphi_n(\mathbf{x}_n) = p(\mathbf{x}_n) = \psi(\mathbf{x}_n)$,

$$\begin{aligned}
&\mathrm{Var}(\hat{I}_{\mathtt{R3}}) \\
&\quad= \frac{1}{Z^2N^2}\sum_{n=1}^{N}\int \frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{\varphi_n^2(\mathbf{x}_n)}p(\mathbf{x}_n)\,d\mathbf{x}_n \\
&\qquad- \frac{1}{Z^2N^2}\sum_{n=1}^{N}\Bigg(\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi_n(\mathbf{x}_n)}p(\mathbf{x}_n)\,d\mathbf{x}_n\Bigg)^2 \\
&\quad= \frac{1}{Z^2N^2}\sum_{n=1}^{N}\int \frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{\psi^2(\mathbf{x}_n)}\psi(\mathbf{x}_n)\,d\mathbf{x}_n \\
&\qquad- \frac{1}{Z^2N^2}\sum_{n=1}^{N}\Bigg(\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\psi(\mathbf{x}_n)}\psi(\mathbf{x}_n)\,d\mathbf{x}_n\Bigg)^2 \\
&\quad= \frac{1}{Z^2N}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{\psi(\mathbf{x})}\,d\mathbf{x} - \frac{I^2}{N}.
\end{aligned} \tag{D.7}$$

4. [N1] *Sampling 3/Weighting 3*: The methods that use sampling without replacement introduce correlation at the selection of the proposals. However, under the perspective of the deterministic sampling ($\mathcal{S}_3$), the nth sample $\mathbf{x}_n$ is a realization of the r.v. $X_n \sim q_n$ and is independent of the other samples. Marginalizing first equation (D.3) over the indexes, and substituting $\varphi_n(\mathbf{x}_n) = p(\mathbf{x}_n) = q_n(\mathbf{x}_n)$:

$$\begin{aligned}
&\mathrm{Var}(\hat{I}_{\mathtt{N1}}) \\
&\quad= \frac{1}{Z^2N^2}\sum_{n=1}^{N}\int \frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{\varphi_n^2(\mathbf{x}_n)}p(\mathbf{x}_n)\,d\mathbf{x}_n
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \left(\int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_n(\mathbf{x}_n)} p(\mathbf{x}_n)\, d\mathbf{x}_n \right)^2 \\
&= \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{q_n^2(\mathbf{x}_n)} q_n(\mathbf{x}_n)\, d\mathbf{x}_n \\
&\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \left(\int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{q_n(\mathbf{x}_n)} q_n(\mathbf{x}_n)\, d\mathbf{x}_n \right)^2 \\
&= \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{q_n(\mathbf{x}_n)}\, d\mathbf{x}_n - \frac{I^2}{N}.
\end{aligned} \tag{D.8}
$$

5. [N2] *Sampling 2/Weighting 1*: In this scheme, we use again the expression for conditional independence of equation (D.4). Substituting $\varphi_{j_{1:n-1}} = p(\mathbf{x}_n | j_{1:n-1})$,

$$
\begin{aligned}
&\mathrm{Var}(\hat{I}_{\mathrm{N2}}) \\
&= \frac{1}{Z^2 N^2} \sum_{j_{1:N}} \left[\sum_{n=1}^{N} \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{\varphi^2_{j_{1:n-1}}(\mathbf{x}_n)} p(\mathbf{x}_n | j_n)\, d\mathbf{x}_n \right. \\
&\quad \left. - \sum_{n=1}^{N} \left(\int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_{j_{1:n-1}}(\mathbf{x}_n)} p(\mathbf{x}_n | j_n)\, d\mathbf{x}_n \right)^2 \right] P(j_{1:N}) \\
&= \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \sum_{j_{1:n}} \left[\int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{p^2(\mathbf{x}_n | j_{1:n-1})} p(\mathbf{x}_n | j_n)\, d\mathbf{x}_n \right. \\
&\quad \left. - \left(\int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{p(\mathbf{x}_n | j_{1:n-1})} p(\mathbf{x}_n | j_n)\, d\mathbf{x}_n \right)^2 \right] P(j_{1:n}) \\
&= \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \sum_{j_{1:n-1}} \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{p(\mathbf{x}_n | j_{1:n-1})} P(j_{1:n-1})\, d\mathbf{x}_n \\
&\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \sum_{j_{1:n}} \left(\int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{p(\mathbf{x}_n | j_{1:n-1})} q_{j_n}\, d\mathbf{x}_n \right)^2 \\
&\quad \cdot P(j_{1:n}).
\end{aligned} \tag{D.9}
$$

Since the the integrals only depend on the set of indexes $j_{1:n}$, each term of the sum has been first marginalized over $j_{n+1:N}$. The first term in the sum can then be further marginalized over j_n to obtain the final expression. Note that the variance is the average of the variance of all the $N!$ possible sequences of indexes in the sampling without replacement.

6. [N3] *Sampling 3/Weighting 5*: We have followed the same arguments of scheme N1. Marginalizing equation (D.3) over all the set of indexes $j_{1:N}$, and substituting $\varphi_n(\mathbf{x}_n) = f(\mathbf{x}_n) = \psi(\mathbf{x}_n)$:

$$
\begin{aligned}
&\mathrm{Var}(\hat{I}_{\mathrm{N3}}) \\
&= \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{\psi^2(\mathbf{x}_n)} q_n(\mathbf{x}_n)\, d\mathbf{x}_n \\
&\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \left(\int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\psi(\mathbf{x}_n)} q_n(\mathbf{x}_n)\, d\mathbf{x}_n \right)^2 \\
&= \frac{1}{Z^2 N} \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{\psi^2(\mathbf{x})} \left(\frac{1}{N} \sum_{n=1}^{N} q_n(\mathbf{x}) \right) d\mathbf{x} \\
&\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \left(\int \frac{\pi(\mathbf{x}) g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})\, d\mathbf{x} \right)^2 \\
&= \frac{1}{Z^2 N} \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{\psi(\mathbf{x})}\, d\mathbf{x} \\
&\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \left(\int \frac{\pi(\mathbf{x}) g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})\, d\mathbf{x} \right)^2,
\end{aligned} \tag{D.10}
$$

where we have used the identity $\psi(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} q_n(\mathbf{x})\, d\mathbf{x}$.

D.2 Proof of Theorem 6.1

The proof of Theorem 6.1 is split in the next three propositions.

PROPOSITION D.1. $\mathrm{Var}(\hat{I}_{\mathrm{R1}}) = \mathrm{Var}(\hat{I}_{\mathrm{N1}})$

PROOF. See that equations (D.5) and (D.8) are equivalent. □

PROPOSITION D.2. $\mathrm{Var}(\hat{I}_{\mathrm{N1}}) \geq \mathrm{Var}(\hat{I}_{\mathrm{R3}})$.

PROOF. Subtracting equations (D.7) and (D.8), we get

$$
\begin{aligned}
&\mathrm{Var}(\hat{I}_{\mathrm{R3}}) - \mathrm{Var}(\hat{I}_{\mathrm{N1}}) \\
&= \frac{1}{Z^2 N^2} \int \left(\frac{N}{\frac{1}{N} \sum_{j=1}^{N} q_j(\mathbf{x})} \right. \\
&\quad \left. - \sum_{i=1}^{N} \frac{1}{q_i(\mathbf{x})} \right) g^2(\mathbf{x}) \pi^2(\mathbf{x})\, d\mathbf{x}.
\end{aligned}
$$

Since $g^2(\mathbf{x})\pi^2(\mathbf{x}) \geq 0\ \forall \mathbf{x} \in \mathbb{R}^{d_x}$, it is sufficient to show that

$$
\frac{1}{\frac{1}{N} \sum_{j=1}^{N} q_j(\mathbf{x})} \leq \frac{1}{N} \sum_{i=1}^{N} \frac{1}{q_i(\mathbf{x})}. \tag{D.11}
$$

Now let us note that the left-hand side of equation (D.11) is the inverse of the arithmetic mean of

$q_1(\mathbf{x}), \ldots, q_N(\mathbf{x})$,

$$A_N = \frac{1}{N}\sum_{j=1}^{N} q_j(\mathbf{x}),$$

whereas the right-hand side of equation (D.11) is the inverse of the harmonic mean of $q_1(\mathbf{x}), \ldots, q_N(\mathbf{x})$,

$$\frac{1}{H_N} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{q_i(\mathbf{x})}.$$

Therefore, the inequality in equation (D.11) is equivalent to stating that $\frac{1}{A_N} \leq \frac{1}{H_N}$, or equivalently $A_N \geq H_N$, which is the well-known arithmetic mean-harmonic mean inequality for positive real numbers (Hardy, Littlewood and Pólya, 1952, Abramowitz and Stegun, 1992, Gwanyama, 2004). □

PROPOSITION D.3. $\mathrm{Var}(\hat{I}_{\mathrm{R3}}) \geq \mathrm{Var}(\hat{I}_{\mathrm{N3}})$.

PROOF. Subtracting (D.7) and (D.10), we get

$$\mathrm{Var}(\hat{I}_{\mathrm{N3}}) - \mathrm{Var}(\hat{I}_{\mathrm{R3}}) = -\frac{I^2}{N} + \frac{1}{Z^2N^2}\sum_{n=1}^{N}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})\,d\mathbf{x}\right)^2.$$

Therefore, the proposition is proved if

$$\frac{1}{Z^2}\sum_{n=1}^{N}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})\,d\mathbf{x}\right)^2 \geq NI^2.$$

If we substitute $I = \int g(\mathbf{x})\tilde{\pi}(\mathbf{x})\,d\mathbf{x}$, multiplying both numerator and denominator by $\psi(\mathbf{x})$ in the integral of the right-hand side,

$$\frac{1}{Z^2}\sum_{n=1}^{N}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})\,d\mathbf{x}\right)^2 \geq N\left(\frac{1}{Z}\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})}\psi(\mathbf{x})\,d\mathbf{x}\right)^2,$$

$$\sum_{n=1}^{N}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})\,d\mathbf{x}\right)^2 \geq N\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})}\left(\frac{1}{N}\sum_{n=1}^{N} q_n(\mathbf{x})\right)d\mathbf{x}\right)^2, \tag{D.12}$$

$$\sum_{n=1}^{N}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})\,d\mathbf{x}\right)^2 \geq \frac{1}{N}\left(\sum_{n=1}^{N}\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})\,d\mathbf{x}\right)^2,$$

$$N\sum_{n=1}^{N} a_n^2 \geq \left(\sum_{n=1}^{N} a_n\right)^2$$

with $a_n = \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})\,d\mathbf{x}$. The inequality of equation (D.12) holds, since it is the definition of the Cauchy–Schwarz inequality (Hardy, Littlewood and Pólya, 1952),

$$\left(\sum_{n=1}^{N} a_n^2\right)\left(\sum_{n=1}^{N} b_n^2\right) \geq \left(\sum_{n=1}^{N} a_n b_n\right)^2, \tag{D.13}$$

with $b_n = 1$ for $n = 1, \ldots, N$. □

PROOF OF THEOREM 6.1. The proof is obtained by applying Propositions D.1, D.2 and D.3. □

D.3 Proof of Theorem 6.2

Let us first particularize the variance expression for $N = 2$. From equation (D.8),

$$\mathrm{Var}(\hat{I}_{\mathrm{N1}}) = \mathrm{Var}(\hat{I}_{\mathrm{R1}}) = \frac{1}{4Z^2}\left(\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_1(\mathbf{x})}\,d\mathbf{x} + \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_2(\mathbf{x})}\,d\mathbf{x}\right) - \frac{I^2}{2}. \tag{D.14}$$

From equation (D.7),

$$\mathrm{Var}(\hat{I}_{\mathrm{R3}}) = \frac{1}{2Z^2}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}}\,d\mathbf{x} - \frac{I^2}{2}. \tag{D.15}$$

From equation (D.10),

$$\mathrm{Var}(\hat{I}_{\mathrm{N3}}) = \frac{1}{2Z^2}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}}\,d\mathbf{x} - \frac{1}{4Z^2}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_1(\mathbf{x})\,d\mathbf{x}\right)^2 - \frac{1}{4Z^2}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_2(\mathbf{x})\,d\mathbf{x}\right)^2. \tag{D.16}$$

From equation (D.6),

$$\mathrm{Var}(\hat{I}_{\mathrm{R2}}) = \frac{1}{8Z^2}\left(\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_1(\mathbf{x})}\,d\mathbf{x} + \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_2(\mathbf{x})}\,d\mathbf{x}\right) - \frac{I^2}{4} + \frac{1}{4Z^2}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}}\,d\mathbf{x} - \frac{1}{8Z^2}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_1(\mathbf{x})\,d\mathbf{x}\right)^2 - \frac{1}{8Z^2}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_2(\mathbf{x})\,d\mathbf{x}\right)^2. \tag{D.17}$$

From equation (D.9),

$$\begin{aligned}
&\operatorname{Var}(\hat{I}_{\mathtt{N2}}) \\
&\quad = \frac{1}{4Z^2}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}}\,d\mathbf{x} \\
&\qquad + \frac{1}{8Z^2}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_1(\mathbf{x})}\,d\mathbf{x} \\
&\qquad + \frac{1}{8Z^2}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_2(\mathbf{x})}\,d\mathbf{x} \\
&\qquad - \frac{1}{8Z^2}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_1(\mathbf{x})\,d\mathbf{x}\right)^2 \\
&\qquad - \frac{1}{8Z^2}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_2(\mathbf{x})\,dx\right)^2 - \frac{I^2}{4}. \qquad \square
\end{aligned} \tag{D.18}$$

PROPOSITION D.4. *For* $N = 2$, $\operatorname{Var}(\hat{I}_{\mathtt{R2}}) = \operatorname{Var}(\hat{I}_{\mathtt{N2}})$.

PROOF. See that equations (D.17) and (D.18) are equivalent. $\square$

PROPOSITION D.5. *For* $N = 2$, $\operatorname{Var}(\hat{I}_{\mathtt{N1}}) \geq \operatorname{Var}(\hat{I}_{\mathtt{R2}}) \geq \operatorname{Var}(\hat{I}_{\mathtt{N3}})$.

PROOF. Analyzing equations (D.14) and (D.16), we see that equation (D.17) can be rewritten as

$$\operatorname{Var}(\hat{I}_{\mathtt{R2}}) = \frac{1}{2}\operatorname{Var}(\hat{I}_{\mathtt{N1}}) + \frac{1}{2}\operatorname{Var}(\hat{I}_{\mathtt{N3}}). \tag{D.19}$$

Since in Theorem 6.1 it is proved that $\operatorname{Var}(\hat{I}_{\mathtt{N1}}) \geq \operatorname{Var}(\hat{I}_{\mathtt{N3}})$ for any N, the proposition holds at least for $N = 2$. $\square$

PROOF OF THEOREM 6.2. The proof is obtained by applying Propositions D.4 and D.5. $\square$

REMARK D.1. We hypothesize that Theorem 6.2 might also hold for $N > 2$. The MIS schemes `R2` and `N2` seem to average estimators with variance reduction (related to `N3`) with estimators with worse variance (related to `N1`).

REMARK D.2. Note that the scheme `R3` does not appear in Theorem 6.2. Equation (D.17) can be rewritten as

$$\begin{aligned}
&\operatorname{Var}(\hat{I}_{\mathtt{R2}}) \\
&\quad = \frac{1}{2}\operatorname{Var}(\hat{I}_{\mathtt{R3}}) + \frac{1}{8Z^2}\left(\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_1(\mathbf{x})}\,d\mathbf{x}\right. \\
&\qquad \left. + \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_2(\mathbf{x})}\,d\mathbf{x}\right) \\
&\qquad - \frac{1}{8Z^2}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_1(\mathbf{x})\,d\mathbf{x}\right)^2 \\
&\qquad - \frac{1}{8Z^2}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_2(\mathbf{x})\,d\mathbf{x}\right)^2.
\end{aligned}$$

The question is then whether the last four terms are larger than $\frac{1}{2}\operatorname{Var}(\hat{I}_{\mathtt{R3}})$. We hypothesize that no inequality can be established in a general case, that is, whether the scheme `R3` would outperform `R2` or not for a given $\pi(\mathbf{x})$ and $g(\mathbf{x})$, might depend on the proposals $q_1(\mathbf{x})$ and $q_2(\mathbf{x})$.

D.4 Example with Closed-Form Variances

Let us derive the expressions of the example of Section 6.1 by considering the targeted distribution

$$\pi(\mathbf{x}) = \frac{1}{2}\left[\mathcal{N}(\mathbf{x}|-\mu,\sigma^2) + \mathcal{N}(\mathbf{x}|\mu,\sigma^2)\right]. \tag{D.20}$$

We consider $N = 2$ proposal densities, $q_1(\mathbf{x}) = \mathcal{N}(\mathbf{x}|-\mu,\sigma^2)$ and $q_2(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu,\sigma^2)$. Note that the mixture of proposals is exactly the targeted distribution, that is, $\psi(\mathbf{x}) = \pi(\mathbf{x})$. We address the case where we want to estimate a specific moment g of π with the $M = 2$ samples. In the following, we provide explicit variances of the unnormalized estimator of equation (2.4) for the six MIS schemes. From equation (D.5),

$$\begin{aligned}
\operatorname{Var}(\hat{I}_{\mathtt{N1}}) &= \frac{1}{4}\left[\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_1(\mathbf{x})}\,d\mathbf{x} + \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_2(\mathbf{x})}\,d\mathbf{x}\right] \\
&\quad - \frac{I}{2} \\
&= \frac{1}{4}[S_1 + S_2] - \frac{I}{2}.
\end{aligned} \tag{D.21}$$

Let us first compute

$$\begin{aligned}
S_1 &= \int \frac{g^2(\mathbf{x})\frac{1}{2}(q_1(\mathbf{x}) + q_2(\mathbf{x}))}{q_1(\mathbf{x})}\pi(\mathbf{x})\,d\mathbf{x} \\
&= \frac{1}{2}\left[\int g^2(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x} + \int \frac{q_2(\mathbf{x})}{q_1(\mathbf{x})} g^2(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x}\right] \\
&= \frac{1}{4}\left[\int g^2(\mathbf{x})q_1(\mathbf{x})\,d\mathbf{x} + \int g^2(\mathbf{x})q_2(\mathbf{x})\,d\mathbf{x}\right. \\
&\quad \left. + \int g^2(\mathbf{x})\frac{q_1(\mathbf{x}) + q_2(\mathbf{x})}{q_1(\mathbf{x})} q_2(\mathbf{x})\,d\mathbf{x}\right] \\
&= \frac{1}{4}\left[\int g^2(\mathbf{x})q_1(\mathbf{x})\,d\mathbf{x} + 2\int g^2(\mathbf{x})q_2(\mathbf{x})\,d\mathbf{x}\right. \\
&\quad \left. + \int g^2(\mathbf{x})\frac{q_2(\mathbf{x})}{q_1(\mathbf{x})} q_2(\mathbf{x})\,d\mathbf{x}\right].
\end{aligned}$$

Since the proposals are Gaussian,

$$\begin{aligned}
&\frac{q_2(\mathbf{x})}{q_1(\mathbf{x})}q_2(\mathbf{x})\\
&\quad=\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\mathbf{x}-\mu)^2}{2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\mathbf{x}+\mu)^2}{2\sigma^2}\right)}\frac{1}{\sqrt{2\pi\sigma^2}}\\
&\qquad\cdot\exp\left(-\frac{(\mathbf{x}-\mu)^2}{2\sigma^2}\right)\\
&\quad=\exp\left(\frac{4\mu\mathbf{x}}{2\sigma^2}\right)\frac{1}{\sqrt{2\pi\sigma^2}}\exp-\frac{(\mathbf{x}-\mu)^2}{2\sigma^2}\\
&\quad=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\mathbf{x}^2+\mu^2-2\mu\mathbf{x}-4\mu\mathbf{x}}{2\sigma^2}\right)\\
&\quad=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\mathbf{x}-3\mu)}{2\sigma^2}\right)\exp\left(-\frac{8\mu^2}{2\sigma^2}\right).
\end{aligned}$$

Then

$$\begin{aligned}
S_1&=\frac{1}{4}\Bigg[\int g^2(\mathbf{x})q_1(\mathbf{x})\,d\mathbf{x}+2\int g^2(\mathbf{x})q_2(\mathbf{x})\,d\mathbf{x}\\
&\qquad+\exp\left(-\frac{8\mu^2}{2\sigma^2}\right)\int g^2(\mathbf{x})\frac{1}{\sqrt{2\pi\sigma^2}}\\
&\qquad\cdot\exp\left(-\frac{(\mathbf{x}-3\mu)}{2\sigma^2}\right)d\mathbf{x}\Bigg]\\
&=\frac{1}{4}\Bigg[\int g^2(\mathbf{x})q_1(\mathbf{x})\,d\mathbf{x}\\
&\qquad+2\int g^2(\mathbf{x})q_2(\mathbf{x})\,d\mathbf{x}\\
&\qquad+\exp\left(-\frac{8\mu^2}{2\sigma^2}\right)\int g^2(\mathbf{x})\mathcal{N}\big(3\mu,\sigma^2\big)\,d\mathbf{x}\Bigg].
\end{aligned}$$

Similarly,

$$\begin{aligned}
S_2=\frac{1}{4}\Bigg[&\int g^2(\mathbf{x})q_2(\mathbf{x})\,d\mathbf{x}\\
&+2\int g^2(\mathbf{x})q_1(\mathbf{x})\,d\mathbf{x}\\
&+\int g^2(\mathbf{x})\frac{q_1(\mathbf{x})}{q_2(\mathbf{x})}q_1(\mathbf{x})\,d\mathbf{x}\Bigg],
\end{aligned}$$

where

$$\begin{aligned}
&\frac{q_1(\mathbf{x})}{q_2(\mathbf{x})}q_1(\mathbf{x})\\
&\quad=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\mathbf{x}^2+\mu^2+2\mu\mathbf{x}+4\mu\mathbf{x}}{2\sigma^2}\right)\\
&\quad=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\mathbf{x}+3\mu)}{2\sigma^2}\right)\exp\left(\frac{8\mu^2}{2\sigma^2}\right).
\end{aligned}$$

Finally, from equation (D.21),

$$\begin{aligned}
&\operatorname{Var}(\hat{I}_{\mathrm{N1}})\\
&\quad=\frac{1}{16}\Bigg[3\int g^2(\mathbf{x})q_1(\mathbf{x})\,d\mathbf{x}+3\int g^2(\mathbf{x})q_2(\mathbf{x})\,d\mathbf{x}\\
&\qquad+\Bigg(\int g^2(\mathbf{x})\mathcal{N}\big(\mathbf{x}|3\mu,\sigma^2\big)\,d\mathbf{x}\\
&\qquad+\int g^2(\mathbf{x})\mathcal{N}\big(\mathbf{x}|-3\mu,\sigma^2\big)\,d\mathbf{x}\Bigg)\exp\left(\frac{4\mu^2}{\sigma^2}\right)\Bigg]\\
&\qquad-\frac{I}{2}.
\end{aligned}$$

Note that $\operatorname{Var}(\hat{I}_{\mathrm{R1}})=\operatorname{Var}(\hat{I}_{\mathrm{N1}})$. From equation (D.7),

$$\begin{aligned}
\operatorname{Var}(\hat{I}_{\mathrm{R3}})&=\frac{1}{2}\int\frac{g^2(\mathbf{x})\pi(\mathbf{x})}{\pi(x)}\pi(\mathbf{x})\,d\mathbf{x}-\frac{I}{2}\\
&=\frac{1}{2}\int g^2(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x}-\frac{1}{2}\int g(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x}\qquad\text{(D.22)}\\
&=\frac{1}{2}\int g(\mathbf{x})\big(g(\mathbf{x})-1\big)\pi(x)\,d\mathbf{x}.
\end{aligned}$$

From equation (D.10),

$$\begin{aligned}
&\operatorname{Var}(\hat{I}_{\mathrm{N3}})\\
&\quad=\frac{1}{2}\int g^2(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x}-\frac{1}{4}\Bigg[\left(\int g(\mathbf{x})q_1(\mathbf{x})\,d\mathbf{x}\right)^2\\
&\qquad+\left(\int g(\mathbf{x})q_2(\mathbf{x})\,d\mathbf{x}\right)^2\Bigg].
\end{aligned}$$

From equation (D.19), $\operatorname{Var}(\hat{I}_{\mathrm{R2}})=\frac{\operatorname{Var}(\hat{I}_{\mathrm{N1}})+\operatorname{Var}(\hat{I}_{\mathrm{N3}})}{2}$. Therefore,

$$\begin{aligned}
&\operatorname{Var}(\hat{I}_{\mathrm{R2}})\\
&\quad=\frac{1}{32}\Bigg[3\int g^2(\mathbf{x})q_1(\mathbf{x})\,d\mathbf{x}+3\int g^2(\mathbf{x})q_2(\mathbf{x})\,d\mathbf{x}\\
&\qquad+\Bigg(\int g^2(\mathbf{x})\mathcal{N}\big(\mathbf{x}|3\mu,\sigma^2\big)\,d\mathbf{x}\\
&\qquad+\int g^2(\mathbf{x})\mathcal{N}\big(\mathbf{x}|-3\mu,\sigma^2\big)\,d\mathbf{x}\Bigg)\\
&\qquad\cdot\exp\left(\frac{4\mu^2}{\sigma^2}\right)\Bigg]-\frac{I}{4}\\
&\qquad+\frac{1}{4}\int g^2(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x}-\frac{1}{8}\Bigg[\left(\int g(\mathbf{x})q_1(\mathbf{x})\,d\mathbf{x}\right)^2\\
&\qquad+\left(\int g(\mathbf{x})q_2(\mathbf{x})\,d\mathbf{x}\right)^2\Bigg].
\end{aligned}$$

Moreover, from Proposition D.4, $\hat{I}_{\mathrm{N2}}=\hat{I}_{\mathrm{R2}}$.

APPENDIX E: MULTIDIMENSIONAL MIXTURE OF GENERALIZED GAUSSIAN DISTRIBUTIONS

Let us consider a mixture of multivariate generalized Gaussian distributions (GGD) as a target p.d.f. In particular,

$$\text{(E.1)} \quad \pi(\mathbf{x}) = \frac{1}{3}\sum_{k=1}^{3} \mathcal{GG}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k), \quad \mathbf{x} \in \mathbb{R}^{d_x},$$

where $\boldsymbol{\mu}_k = [\mu_{k,1}, \ldots, \mu_{k,d_x}]^\top$, $\boldsymbol{\alpha}_k = [\alpha_{k,1}, \ldots, \alpha_{k,d_x}]^\top$ and $\boldsymbol{\beta}_k = [\beta_{k,1}, \ldots, \beta_{k,d_x}]^\top$ are respectively the mean, scale and shape parameters of each component of the mixture. Each component of the mixture factorizes in all dimensions, that is, the multivariate GGD p.d.f. is the product of N unidimensional GGD p.d.f.'s. Namely,

$$\mathcal{GG}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) = \prod_{d=1}^{d_x} \kappa_{k,d} \exp\left(-\left(\frac{|x_d - \mu_{k,d}|}{\alpha_{k,d}}\right)^{\beta_{k,d}}\right),$$

where $\kappa_{k,d} = \frac{\beta_{k,d}}{2\alpha_{k,d}\Gamma(\frac{1}{\beta_{k,d}})}$, $\Gamma(\cdot)$ is the gamma function, and x_d is the dth dimension of $\mathbf{x}$. This family of distributions includes both Gaussian and Laplace distributions with $\beta = 2$ and $\beta = 1$, respectively. In this example, $\mu_{1,d} = -3$, $\mu_{2,d} = 1$, $\mu_{3,d} = 5$, $\beta_{1,d} = 1.1$, $\beta_{2,d} = 1.8$, $\beta_{3,d} = 5$, $\alpha_{1,d} = \alpha_{2,d} = \alpha_{3,d} = 1$ for all $d = 1, \ldots, d_x$. The expected value of the target $\pi(\mathbf{x})$ is then $E_\pi[X_d] = 1$ for $d = 1, \ldots, d_x$. In order to study the performance of the different MIS schemes, we vary the dimension of the state space in equation (E.1) testing different values of d_x (with $2 \leq d_x \leq 10$). We consider the problem of approximating via Monte Carlo the expected value of the target density, and we compare the performance of all MIS schemes. In this example, we use $N = 500$ nonstandardized t-student densities as proposal functions, where each location parameter has been selected uniformly within the $[-6, 6]^{d_x}$ square, and the scale parameters and the degree of freedom parameters have been selected as $\sigma_{n,d} = 5$ and $\nu_{n,d} = 5$, respectively, for $n = 1, \ldots, N$ and $d = 1, \ldots, d_x$. For each method, we draw $M = kN$ samples, with $k = 32$, and we average all the results over 200 runs.

Figure 7 shows the MSE in the estimation of the mean of the target (averaged over all dimensions) when we increase the dimension d_x. Note that the hierarchy established in Section 6 also holds in this example regardless the dimension. In this case, methods `R1` and `N1` behave poorly even at lower dimensions, while the other MIS schemes have a similar behavior. When we increase the dimension, all the methods degrade, and, at certain point ($d_x \geq 6$), the performance of all of them is similar. Note that the proposal p.d.f.'s are fixed in random locations of the space, which is well covered at low dimensions (since we are using $N = 500$ p.d.f.'s), but this coverage becomes worse as the dimension increases. This can probably explain the similar performance of all the methods in higher dimensions.

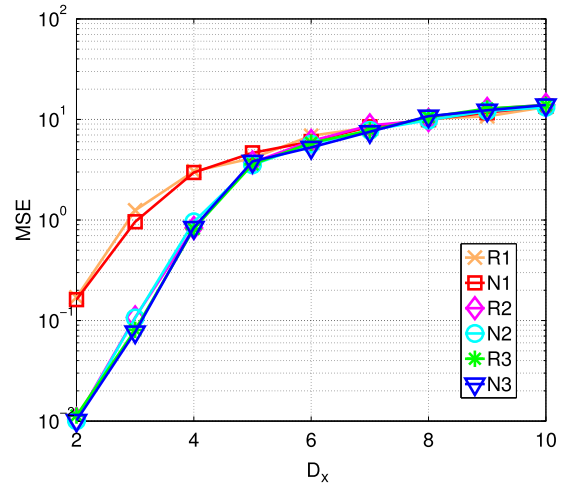


FIG. 7. (Ex. of Section E) *MSE of the self-normalized estimator $\tilde{I}$ for all MIS schemes when we increase the dimension d_x of the state space.*

ACKNOWLEDGMENTS

We thank the Editor, Associate Editor and referees for their constructive comments that helped to improve the paper. V. Elvira acknowledges support from the *Agence Nationale de la Recherche* of France under PISCES project (ANR-17-CE40-0031-01). D. Luengo thanks the support of *Ministerio de Economía, Industria y Competitividad* through projects MIMOD-PLC (TEC2015-64835-C3-3-R) and KERMES (TEC2016-81900-REDT). M. F. Bugallo thanks the support of the National Science Foundation (NSF) under Award CCF-1617986.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A., eds. (1992). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York. MR1225604

BUGALLO, M. F., ELVIRA, V., MARTINO, L., LUENGO, D., MÍGUEZ, J. and DJURIC, P. M. (2017). Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Process. Mag.* **34** 60–79.

CAPPÉ, O., GUILLIN, A., MARIN, J. M. and ROBERT, C. P. (2004). Population Monte Carlo. *J. Comput. Graph. Statist.* **13** 907–929. MR2109057

CAPPÉ, O., DOUC, R., GUILLIN, A., MARIN, J.-M. and ROBERT, C. P. (2008). Adaptive importance sampling in general mixture classes. *Stat. Comput.* **18** 447–459. MR2461888

CORNUET, J.-M., MARIN, J.-M., MIRA, A. and ROBERT, C. P. (2012). Adaptive multiple importance sampling. *Scand. J. Stat.* **39** 798–812. MR3000850

DOUC, R. and CAPPÉ, O. (2005). Comparison of resampling schemes for particle filtering. In *ISPA* 2005. *Proceedings of the* 4*th International Symposium on Image and Signal Processing and Analysis* 64–69. IEEE, New York.

DOUC, R., GUILLIN, A., MARIN, J.-M. and ROBERT, C. P. (2007a). Convergence of adaptive mixtures of importance sampling schemes. *Ann. Statist.* **35** 420–448. MR2332281

DOUC, R., GUILLIN, A., MARIN, J.-M. and ROBERT, C. P. (2007b). Minimum variance importance sampling via population Monte Carlo. *ESAIM Probab. Stat.* **11** 427–447. MR2339302

ELVIRA, V., MARTINO, L., LUENGO, D. and BUGALLO, M. F. (2015). Efficient multiple importance sampling estimators. *IEEE Signal Process. Lett.* **22** 1757–1761.

ELVIRA, V., MARTINO, L., LUENGO, D. and CORANDER, J. (2015b). A gradient adaptive population importance sampler. In *IEEE International Conf. on Acoustics, Speech and Signal Processing* (*ICASSP*) 4075–4079.

ELVIRA, V., MARTINO, L., LUENGO, D. and BUGALLO, M. F. (2016). Heretical multiple importance sampling. *IEEE Signal Process. Lett.* **23** 1474–1478.

ELVIRA, V., MARTINO, L., LUENGO, D. and BUGALLO, M. F. (2017). Improving Population Monte Carlo: Alternative weighting and resampling schemes. *Signal Process.* **131** 77–91.

GEWEKE, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57** 1317–1339. MR1035115

GORDON, N., SALMOND, D. and SMITH, A. F. M. (1993). Novel approach to nonlinear and non-Gaussian Bayesian state estimation. *IEE Proc. F, Commun. Radar Signal Process.* **140** 107–113.

GWANYAMA, P. W. (2004). The HM-GM-AM-QM inequalities. *College Math. J.* **35** 47–50.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Comput. Statist.* **14** 375–396.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

HARDY, G. H., LITTLEWOOD, J. E. and PÓLYA, G. (1952). *Inequalities*, 2nd ed. Cambridge Univ. Press, Cambridge. MR0046395

HE, H. Y. and OWEN, A. B. (2014). Optimal mixture weights in multiple importance sampling. Preprint. Available at arXiv:1411.3954.

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

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

KONG, A. (1992). A note on importance sampling using standardized weights. Technical Report 348, Dept. Statistics, Univ. Chicago, Chicago, IL.

KONG, A., LIU, J. S. and WONG, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **9** 278–288.

KONG, A., MCCULLAGH, P., MENG, X.-L., NICOLAE, D. and TAN, Z. (2003). A theory of statistical models for Monte Carlo integration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 585–618. MR1998624

LIANG, F. (2002). Dynamically weighted importance sampling in Monte Carlo computation. *J. Amer. Statist. Assoc.* **97** 807–821. MR1941411

LIU, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR2401592

MARTINO, L., ELVIRA, V., LUENGO, D. and CORANDER, J. (2015a). An adaptive population importance sampler: Learning from uncertainty. *IEEE Trans. Signal Process.* **63** 4422–4437. MR3368395

MARTINO, L., ELVIRA, V., LUENGO, D. and CORANDER, J. (2017). Layered adaptive importance sampling. *Stat. Comput.* **27** 599–623. MR3613588

NIEDERREITER, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF Regional Conference Series in Applied Mathematics* **63**. SIAM, Philadelphia, PA. MR1172997

OWEN, A. (2013). Monte Carlo Theory, Methods and Examples. Available at http://statweb.stanford.edu/~owen/mc/.

OWEN, A. and ZHOU, Y. (2000). Safe and effective importance sampling. *J. Amer. Statist. Assoc.* **95** 135–143. MR1803146

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York. MR2080278

TAN, Z. (2004). On a likelihood approach for Monte Carlo integration. *J. Amer. Statist. Assoc.* **99** 1027–1036. MR2109492

VEACH, E. and GUIBAS, L. (1995). Optimally combining sampling techniques for Monte Carlo rendering. In *SIGGRAPH* 1995 *Proceedings* 419–428.