Generalized Multiple Importance Sampling

Víctor Elvira[1*], **Luca Martino**[2], **David Luengo**[3], **and Mónica F. Bugallo**[4]

[1]Télécom Lille (France), [2]Universidad de Valencia (Spain), [3]Universidad Politécnica de Madrid (Spain), [4]Stony Brook University (USA)

Abstract. Importance sampling methods are broadly used to approximate posterior distributions or some of their moments. In its standard approach, samples are drawn from a single proposal distribution and weighted properly. However, since the performance 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, many works have addressed the selection and adaptation of the proposal distributions, interpreting the sampling and the weighting steps in different ways. In this paper, we establish a novel general framework for sampling and weighting procedures when more than one proposal is available. The most relevant MIS schemes in the literature are encompassed within the new framework, 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 (pdfs) [Robert and Casella, 2004; Liu, 2004]. The standard IS technique draws samples from a single proposal pdf and assigns them weights based on the ratio between the target and the proposal pdfs, both evaluated at the sample value. The choice of a suitable proposal pdf is crucial for obtaining a good approximation of the target pdf using the IS method. Indeed, although the validity of this approach is guaranteed under mild assumptions, the variance of the estimator depends notably on the discrepancy between the shape of the proposal and the target [Robert and Casella, 2004; Liu, 2004].

Therefore, several advanced strategies have been proposed in the literature to design more robust IS schemes [Liu, 2004, Chapter 2], [Owen, 2013, Chapter 9],

*Contact author for the manuscript (e-mail: victor.elvira@telecom-lille.fr)

[Liang, 2002]. A powerful approach is based on using a population of different proposal pdfs. 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, e.g. 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., 2015a]. In general, MIS strategies provide more robust algorithms, since they avoid entrusting the performance of the method to a single proposal. Moreover, many algorithms have been proposed in order to conveniently adapt the set of proposals in MIS [Cappé et al., 2004, 2008; Martino et al., 2015a].

When a set of proposal pdfs 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 non-adaptive) 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., 2015a; Martino et al., 2015a; Cornuet et al., 2012; Bugallo et al., 2015]. Namely, there are several possible combinations of sampling and weighting schemes, when a set of proposal pdfs is available, which lead to valid MIS approximations of the target pdf. 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 pdfs 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 pdfs. 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., 2015a; Cappé et al., 2004; Cornuet et al., 2012; Martino et al., 2015a; Bugallo et al., 2015]) 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 the existing ones) in terms of the variance of the estimators. A running example has been 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 pdfs are intentionally well chosen, to evidence the significant effects produced by the different interpretations of the sampling and weighting schemes. We illustrate the performance of all the methods by means of three other numerical examples (including a high-dimensional nonlinear challenging setup), highlighting the differences among the various MIS schemes in terms of performance and computational cost.

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.

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 made about the system, $\mathbf{y} \in \mathbb{R}^{d_y}$.[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 density function (pdf) 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 pdf, and $Z(\mathbf{y})$ is the normalization factor.[2] The objective is to approximate the pdf of interest (referred to as target pdf) by Monte Carlo-based sampling [Kong et al., 2003; Robert and Casella, 2004; Liu, 2004; Owen, 2013]. The resulting approximation of $\pi(\mathbf{x})$ will be denoted as $\hat{\pi}(\mathbf{x})$ and will be attained using importance sampling (IS) techniques.

2.1 Standard importance sampling

Importance sampling is a general Monte Carlo technique for the approximation of a pdf 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 pdf, $q(\mathbf{x})$, characterized by tails that are heavier than those of the target pdf, $\pi(\mathbf{x})$. A particular sample, $\mathbf{x}_n$, is assigned a weight, w_n, which measures the adequacy of that particular sample in the approximation of the posterior pdf. Namely, this importance weight is given by

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

which represents the ratio between the target pdf, π, and the proposal pdf, q, both evaluated at $\mathbf{x}_n$. The samples and the weights form the random measure

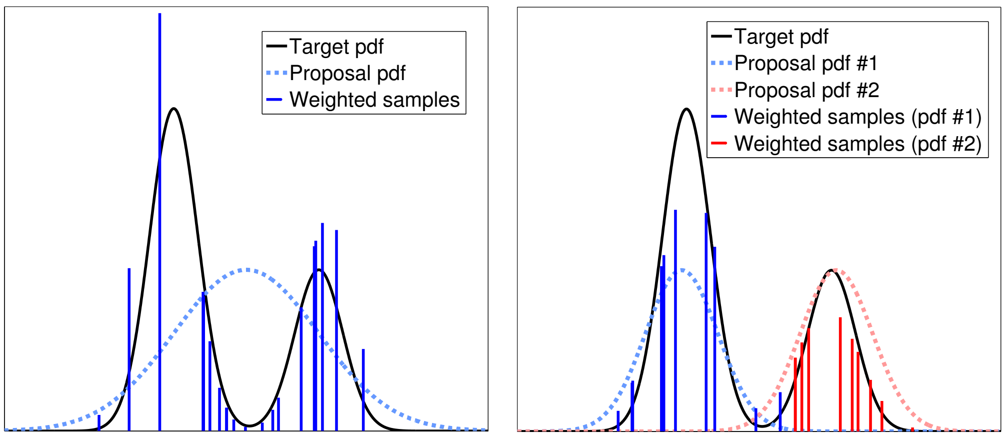
[1] Vectors are denoted by bold-faced letters, e.g., $\mathbf{x}$, while regular-faced letters are used for scalars, e.g., x.

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

$\chi = \{\mathbf{x}_n, w_n\}_{n=1}^N$ that approximates the measure of the target pdf as

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

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]. Fig. 1 (a) displays an example of a target pdf and a proposal pdf, as well as the samples and weights that form a random measure approximating the posterior.



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

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

3. SAMPLING IN MULTIPLE IMPORTANCE SAMPLING

MIS-based schemes consider a set of N proposal pdfs,

$$\{q_n(\mathbf{x})\}_{n=1}^N \equiv \{q_1(\mathbf{x}), \dots, q_N(\mathbf{x})\}.$$

and proceed by generating M samples, $\{\mathbf{x}_m\}_{m=1}^M$ (where $M \neq N$, in general) from them and properly weighting the samples drawn. As a visual example, Fig. 1 (b) displays a target pdf and two proposal pdfs, 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 $M = N$, i.e., the number of samples to be generated coincides with the number of proposal pdfs. All the considerations can be automatically extended to the case $M = kN$, with $k \geq 1$ and $k \in \mathbb{N}$.

Note that the use of the complete set of N proposal pdfs with no prior information about them can also represent a single equally weighted mixture proposal,

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

The previous interpretation motivates some of the sampling and weighting schemes discussed in this paper. Note that unequal weights could also be considered in the mixture, e.g. [He and Owen, 2014] shows how the weights can be optimized to minimize the variance for a certain integrand.

3.1 Sampling from a mixture of proposal pdfs

In order to provide a better explanation of the discussed sampling procedures, we employ a simile with the urn sampling problem. Let us consider an urn that contains N balls, where each ball is assigned an index $j \in \{1, \dots, N\}$, representing the j-th proposal of the complete set of available proposal pdfs, $\{q_j(\mathbf{x})\}_{j=1}^N$. Then, a generic sampling scheme for drawing N samples from $\psi(\mathbf{x})$ is given below. Starting with $n = 1$:

1. Draw a ball from the urn, i.e., choose an index $j_n \in \{1, \dots, N\}$ using some suitable approach. This corresponds to the selection of a proposal pdf, q_{j_n}.
2. Generate a sample $\mathbf{x}_n$ from the selected proposal pdf, i.e., $\mathbf{x}_n \sim q_{j_n}(\mathbf{x}_n)$.
3. Set $n = n + 1$ and go to step 1.

The graphical model corresponding to this sampling scheme is shown in Fig. 2.

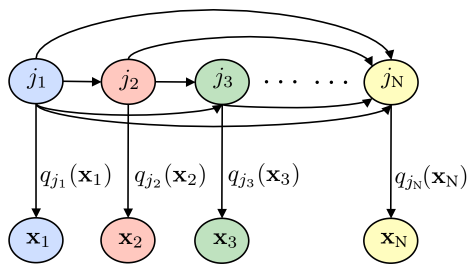


FIG 2. *Graphical model associated to the generic sampling scheme.*

Therefore, obtaining the set of samples $\{\mathbf{x}_n\}_{n=1}^N \equiv \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ from the mixture pdf ψ is in general a two step sequential procedure. First, the n-th index j_n is drawn according to some conditional pdf, $P(j_n|j_{1:n-1})$, where $j_{1:n-1} \equiv \{j_1, \dots, j_{n-1}\}$ is the sequence of the previously generated indexes. [3] Then, the n-th sample is drawn from the selected proposal pdf 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 then

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

where $p(\mathbf{x}_n|j_n) = q_{j_n}(\mathbf{x}_n)$ for all $n \in \{1, \dots, N\}$ is the conditional pdf of the n-th sample given the n-th selected index, i.e., the selected proposal pdf, $q_{j_n}(\mathbf{x}_n)$. The

[3] We use a simplified argument-wise notation, where $p(\mathbf{x}_n)$ denotes the pdf of the continuous random variable (r.v.) $\mathbf{X}_n$, while $P(j_n)$ denotes the probability mass function (pmf) of the discrete r.v. J_n. Also, $p(\mathbf{x}_n, j_n)$ denotes the joint pdf and $p(\mathbf{x}_n|j_n)$ is the conditional pdf 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 pdf as a function, e.g., $p(\mathbf{z}|j_n)$ denotes the pdf $p(\mathbf{x}_n|j_n)$ evaluated at $\mathbf{x}_n = \mathbf{z}$.

full joint distribution of all samples and indexes is given by

$$p(\mathbf{x}_{1:N}, j_{1:N}) = P(j_1)\left[\prod_{i=2}^{N} P(j_i|j_{1:i-1})\right]\left[\prod_{i=1}^{N} p(\mathbf{x}_i|j_i)\right]. \tag{3.3}$$

3.2 Selection of the proposal pdfs

In the sequel, we describe three mechanisms for obtaining the sequence of indexes, $j_{1:N}$: two random schemes (with and without replacement) and a deterministic procedure. The resulting sampling methods will be labeled as $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$, respectively. All the mechanisms share the property that

$$\frac{1}{N}\sum_{n=1}^{N} P(J_n = k) = \frac{1}{N}, \quad \forall k \in \{1, \dots, N\}, \tag{3.4}$$

i.e., all the indexes have the same probability of being selected.

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

$$P(j_n|j_{1:n-1}) = P(j_n) = \frac{1}{N}. \tag{3.5}$$

With this type of index sampling, there may be more than one sample drawn from some proposals and proposal pdfs that are not used at all.

$\mathcal{S}_2$: *Random index selection without replacement*: In this case, when an index is selected from the set of available values, that particular index is removed from the urn. This means that indexes are uniformly and sequentially drawn from different sets, i.e., $j_1 \in \mathcal{I}_1 = \{1, \dots, N\}$ and $j_n \in \mathcal{I}_n = \{1, \dots, N\} \setminus \{j_{1:n-1}\}$ for $n = 2, \dots, N$. Hence, the conditional probability mass function (pmf) of the n-th index given the previous ones is now

$$P(J_n = k|j_{1:n-1}) = \begin{cases} \dfrac{1}{N-n+1} & \text{if } k \in \mathcal{I}_n, \\ 0 & \text{if } k \notin \mathcal{I}_n, \end{cases} \tag{3.6}$$

where $|\mathcal{I}_n| = N - n + 1$. Note that the marginal pmf of the j-th index is still given by (3.5).[4] However, exactly one sample is drawn from each of the proposal pdfs 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 pmf of the n-th index given the $n-1$ previous ones becomes

$$P(j_n|j_{1:n-1}) = P(j_n) = \mathbb{1}_{j_n=n}, \tag{3.7}$$

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

where $\mathbb{1}$ denotes the indicator function. Again, each of the N proposal pdfs 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., APIS [Martino et al., 2015a]), and it is also implicitly used in some particle filters (PFs), such as the bootstrap PF [Gordon et al., 1993].

3.3 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]).

3.4 Running example

Let us consider $N = 3$ Gaussian proposal pdfs $q_1(x) = \mathcal{N}(x;\mu_1,\sigma_1^2)$, $q_2(x) = \mathcal{N}(x;\mu_2,\sigma_2^2)$ 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 $\{j_1, j_2, j_3\} = \{1, 2, 3\}$.

3.5 Distributions of interest of the n-th 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.

Firstly, note that the distribution of the n-th 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 the random index selection with replacement ($\mathcal{S}_1$), this distribution corresponds to $p(\mathbf{x}_n|j_{1:n-1}) = \psi(\mathbf{x}_n)$. For the random index selection without replacement ($\mathcal{S}_2$), 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 the deterministic index selection scheme ($\mathcal{S}_3$), $p(\mathbf{x}_n|j_{1:n-1}) = q_n(\mathbf{x}_n)$.

Once the n-th index j_n has been selected, the n-th 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 n-th sample, $\mathbf{x}_n$, is then given by

$$p(\mathbf{x}_n) = \sum_{k=1}^{N} q_k(\mathbf{x}_n) P(J_n = k), \tag{3.8}$$

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 (with ($\mathcal{S}_1$) or without ($\mathcal{S}_2$) replacement), $P(J_n = k) = \frac{1}{N}, \forall n, k$, 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 $p(\mathbf{x}_n) = q_n(\mathbf{x}_n)$, i.e., the distribution of the r.v. $\mathbf{X}_n$ is the n-th proposal pdf, and not the whole mixture as in the other two sampling schemes with random index selection.

3.6 Distributions of interest beyond $\mathbf{x}_n$

The traditional IS approach focuses just on the distribution of the r.v. $\mathbf{X}_n$, whereas we are also interested in the distribution of the samples regardless of their index n. The reason is that, in MIS schemes, the N samples can be used jointly regardless of their order of appearance. Hence, we introduce a generic r.v.,

$$\mathbf{X} = \mathbf{X}_n \quad \text{with} \quad n \sim \mathcal{U}\{1, 2, \ldots, N\}, \tag{3.9}$$

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

$$f(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} p_{\mathbf{x}_n}(\mathbf{x}) = \psi(\mathbf{x}), \tag{3.10}$$

where $p_{\mathbf{x}_n}(\mathbf{x})$ denotes the marginal pdf of $\mathbf{X}_n$, given by Eq. (3.8), evaluated at $\mathbf{x}$, and $\psi(\mathbf{x})$ is the mixture pdf.[5] Moreover, one can also obtain the conditional pdf of $\mathbf{X}$ given the sequence of indexes as

$$f(\mathbf{x}|j_{1:N}) = \frac{1}{N} \sum_{k=1}^{N} p_{\mathbf{x}_k}(\mathbf{x}|j_{1:N}) = \frac{1}{N} \sum_{k=1}^{N} q_{j_k}(\mathbf{x}). \tag{3.11}$$

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 ($\mathcal{S}_1$), i.e., conditioned to the selection of the indexes, some proposal pdfs 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 pdf of the r.v. $\mathbf{X}$ defined in Eq. (3.9) 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 5 (in the Appendix) summarizes all the distributions of interest. Note that, the pdf 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 $\mathbf{X}_1, \ldots, \mathbf{X}_N$, i.e., $p(\mathbf{x}_{1:N}) = \prod_{n=1}^{N} \psi(\mathbf{x}_n)$ and $p(\mathbf{x}_{1:N}) = \prod_{n=1}^{N} q_n(\mathbf{x}_n)$ for the sampling with replacement ($\mathcal{S}_1$) and deterministic selection ($\mathcal{S}_3$), respectively. For the sampling without replacement and random selection this joint distribution is

$$p(\mathbf{x}_{1:N}) = \psi(\mathbf{x}_1) \prod_{n=2}^{N} \frac{1}{|\mathcal{I}_n|} \sum_{\ell \in \mathcal{I}_n} q_\ell(\mathbf{x}_n), \tag{3.12}$$

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

[5] For the sake of clarity, in Eq. (3.10) we have used the notation $p_{\mathbf{x}_n}(\mathbf{x})$, instead of $p(\mathbf{x})$ as in Eq. (3.8) and the rest of the paper, to denote the marginal pdf of $\mathbf{X}_n$ evaluated at $\mathbf{x}$.

4. WEIGHTING IN MULTIPLE IMPORTANCE SAMPLING

Let us consider the integral $I = \int g(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}$, where $\pi(\mathbf{x})$ is the target distribution and g is any square integrable function w.r.t. $\pi(\mathbf{x})$. The generic IS estimator of I is given by

$$\hat{I} = \frac{1}{NZ}\sum_{n=1}^{N} w_n g(\mathbf{x}_n) \tag{4.1}$$

where w_n is the importance weight of the n-th sample, $\mathbf{x}_n$, and $Z = \int \pi(\mathbf{x})d\mathbf{x}$ is the normalizing constant. In standard IS, when Z is known, $\hat{I}$ is an unbiased estimator of I. Otherwise, if the target distribution is only known up to the normalizing constant, Z, then one can estimate it as

$$\hat{Z} = \frac{1}{N}\sum_{n=1}^{N} w_n, \tag{4.2}$$

which is an unbiased estimator and asymptotically consistent with the number of samples under some mild assumptions regarding the tails of the proposal and target distributions [Robert and Casella, 2004]. Therefore, $\hat{I}$ is also asymptotically consistent, even when Z is unknown and is replaced with $\hat{Z}$ instead [Robert and Casella, 2004].

The weight assigned to the n-th sample is proportional to the ratio between the target pdf evaluated at the sample value, $\pi(\mathbf{x}_n)$, and the proposal pdf evaluated at the sample value, i.e.,

$$w_n = \frac{\pi(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}, \tag{4.3}$$

where the generic function $\varphi_{\mathcal{P}_n}$ represents the proposal pdf from which it is interpreted that the n-th sample is drawn. This function may be parameterized by a subset or the entire sequence of indexes $j_{1:N}$, i.e., $\mathcal{P}_n \subseteq \{j_1, ..., j_N\}$ (further details are given below).

It is on this interpretation of what the proposal pdf used for the generation of the sample is (the evaluation of the denominator in the weight calculation) that different weighting strategies can be devised.

4.1 Mathematical justification

Our approach is based on analyzing which weighting functions yield *proper* MIS estimators. We propose a generalized properness condition in the MIS scenario over the whole estimator. This perspective has some connections with the definition of *properly* weighted sampled proposed in [Liu, 2004, Section 2.5] (see Section 4.3) for further details).

In particular, we consider that the set of weighting functions $\{w_n\}_{n=1}^{N}$ is proper if

$$\frac{E_{p(\mathbf{x}_{1:N}, j_{1:N})}\left[\frac{1}{N}\sum_{n=1}^{N} w_n g(\mathbf{x}_n)\right]}{E_{p(\mathbf{x}_{1:N}, j_{1:N})}\left[\frac{1}{N}\sum_{n=1}^{N} w_n\right]} = E_{\pi}[g(\mathbf{x})]. \tag{4.4}$$

This is equivalent to impose the restriction

$$\frac{E_{p(\mathbf{x}_{1:N},j_{1:N})}\left[Z\hat{I}\right]}{E_{p(\mathbf{x}_{1:N},j_{1:N})}\left[\hat{Z}\right]} = I, \tag{4.5}$$

which is fulfilled if $E[\hat{I}] = I$ and $E[\hat{Z}] = Z$. Note that the MIS properness is fulfilled by any set of weighting functions $\{w_n\}_{n=1}^N$ that yield an unbiased generic estimator $\hat{I}$, i.e., $E[\hat{I}] = I$. At this point, 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 $\pi(\mathbf{x}_n)$ is the target density and $\varphi_n(\mathbf{x}_n)$ is a generic function parametrized by a set of parameters $\mathcal{P}_n$, and both terms are evaluated at $\mathbf{x}_n$.[6] The expectation of the generic estimator $\hat{I}$ of Eq. (4.1) can be computed as

$$E[\hat{I}] \quad = \quad \frac{1}{ZN}\sum_{n=1}^{N}\sum_{j_{1:N}}\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}P(j_{1:N})p(\mathbf{x}_n|j_n)d\mathbf{x}_n, \tag{4.6}$$

where we use the joint distribution of indexes and samples from Eq. (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 $E[\hat{I}] \equiv I$ in Eq. (4.6).*

In the following, we show that various distributions related to the generation of the samples (discussed in Section 3) can be used as the denominator of the weight $\varphi_{\mathcal{P}_n}$, yielding valid estimators.

4.2 Weighting functions

Here we present several possible functions $\varphi_{\mathcal{P}_n}$, that yield an unbiased estimator of I according to Eq. (4.6). The different choices for $\varphi_{\mathcal{P}_n}$ come naturally from the sampling densities discussed in the previous section. More precisely, they correspond to the appropriate evaluation at $\mathbf{x}_n$ of the five different functions in Table 5 related to the distributions of the generated samples. From now on, $p(\cdot)$ and $f(\cdot)$, which correspond to the pdfs 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 pdf 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, φ is the proposal q_{j_n}.

$\mathcal{W}_3$: $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = p(\mathbf{x}_n)$
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).

[6] Note that, in an even more generalized framework, the n-th weight 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 different generic functions $\varphi_{\mathcal{P}_n}$. The distributions depend on the specific sampling scheme used for drawing the samples as shown in Table 6.

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

$\mathcal{W}_4$: $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi_{j_{1:N}}(\mathbf{x}_n) = f(\mathbf{x}_n|j_{1:N}) = \frac{1}{N}\sum_{k=1}^{N} q_{j_k}(\mathbf{x}_n)$
This interpretation makes use of the distribution of the r.v. $\mathbf{X}$ conditioned on the whole set of indexes (defined in Section 3.6).

$\mathcal{W}_5$: $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi(\mathbf{x}_n) = f(\mathbf{x}_n) = \frac{1}{N}\sum_{k=1}^{N} q_k(\mathbf{x}_n)$
This option considers that all the $\mathbf{x}_n$ are realizations of the r.v. $\mathbf{X}$ defined in Section 3.6 (see Appendix A for a thorough discussion of this interpretation).

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 B. Table 1 summarizes the discussed functions $\varphi_{\mathcal{P}_n}$ that can be used to evaluate the denominator for the weight calculation, $w_n = \frac{\pi(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}$. Other proper weighting functions are described in Section 7.2.

4.3 Connection with Liu-properness of single IS

We consider the definition of properness by Liu [Liu, 2004, 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,

$$\frac{E_q[g(\mathbf{x})w(\mathbf{x})]}{E_q[\pi(\mathbf{x})]} = E_\pi[g(\mathbf{x})], \tag{4.7}$$

i.e., w can be in any form as long as the condition of Eq. (4.7) 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, i.e., any Liu-proper weighting scheme is also proper a according to our definition, but not vice versa.

4.4 Running example

Here we follow the running example of Section 3.4. For instance, let us consider the sampling method $\mathcal{S}_1$ and let the realization of the indexes be the sequence $\{j_1, j_2, j_3\} = \{3, 3, 1\}$. 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 $w_3 = \frac{\pi(\mathbf{x}_3)}{\frac{1}{3}(q_1(\mathbf{x}_3)+2q_3(\mathbf{x}_3))}$.

5. MULTIPLE IMPORTANCE SAMPLING SCHEMES

In this section, we describe the different possible combinations of the sampling strategies considered in Section 3 and the weighting functions devised in Section 4. We note that, even though we have discussed three sampling procedures and five alternatives for weight calculation, 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 $\mathcal{S}_3$). Note that for each specific sampling (i.e., with or without replacement), different weighting options can yield the same function in the denominator (e.g. for the deterministic sampling without replacement, $\mathcal{S}_3$, the denominators for weighting options 1, 2 and 3 are identical, always yielding $w_n = \frac{\pi(\mathbf{x}_n)}{q_n(\mathbf{x}_n)}$). Table 6 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` indicating that the method uses sampling with *replacement* or with an `N` to denote that the method corresponds to a sampling scheme 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 n-th 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 n-th sample, only the mixand 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 $n = 1, ..., N$, one forms a mixture composed by all the corresponding proposal pdfs. The weight calculation of the n-th sample considers this *a posteriori* mixture evaluated at the n-th sample in the denominator, i.e., 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 n-th sample, the denominator applies the value of the n-th sample to the whole mixture ψ composed of the set of initial proposal pdfs (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 pdf. 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 n-th weight, the specific mixand used for the generation of the sample is used. This is the approach frequently used in particle filtering [Gordon et al., 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 mixand, but the order matters (it must be random) since the calculation of the n-th weight uses for the evaluation of the denominator the mixture pdf formed by the proposal pdfs that were still available at the generation of the n-th sample.

`[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 $\mathcal{S}_3$):*

TABLE 2
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
R2	$\mathcal{S}_1$	$\frac{\pi(\mathbf{x}_n)}{\frac{1}{N}\sum_{k=1}^{N} q_{j_k}(\mathbf{x}_n)}$	Novel
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
N3	$\mathcal{S}_3$	$\frac{\pi(\mathbf{x}_n)}{\psi(\mathbf{x}_n)}$	[Martino et al., 2015a; Cornuet et al., 2012]

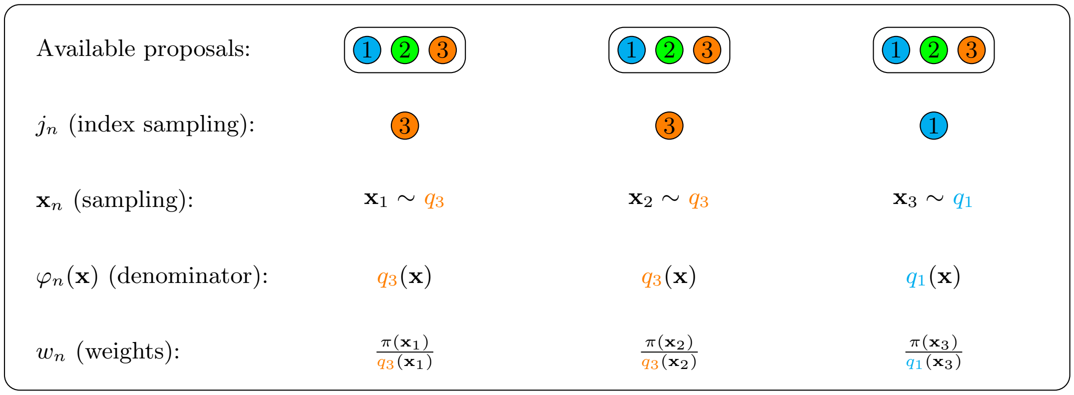
In the calculation of the n-th weight, one uses for the denominator the whole mixture. This is the approach, for instance, of [Martino et al., 2015a; Cornuet et al., 2012]. As showed in Section 6, this scheme has several benefits over the others.

Table 2 summarizes the six resulting MIS schemes and their references in literature, indicating the sampling procedure and weighting function that are applied to obtain the n-th 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 a simpler sampler than $\mathcal{S}_2$.

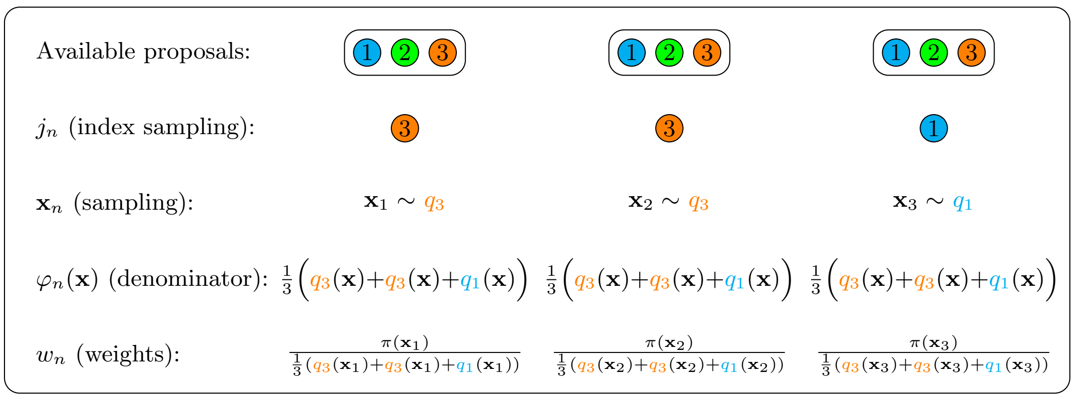
Within the proposed framework, we have considered three sampling procedures and five general weighting functions. All the different algorithms in the literature (that we are aware of) correspond to one of the MIS schemes described above. Section 7.3 provides more details about the MIS schemes used by the different algorithms available in literature. Several new valid schemes have also appeared naturally. Namely, schemes R1, R2, and N2 are novel, and their advantages and drawbacks are analyzed in the following sections. Furthermore, following the sampling and weighting remarks provided above, new proper MIS schemes can easily be proposed within this framework.

5.3 Running example

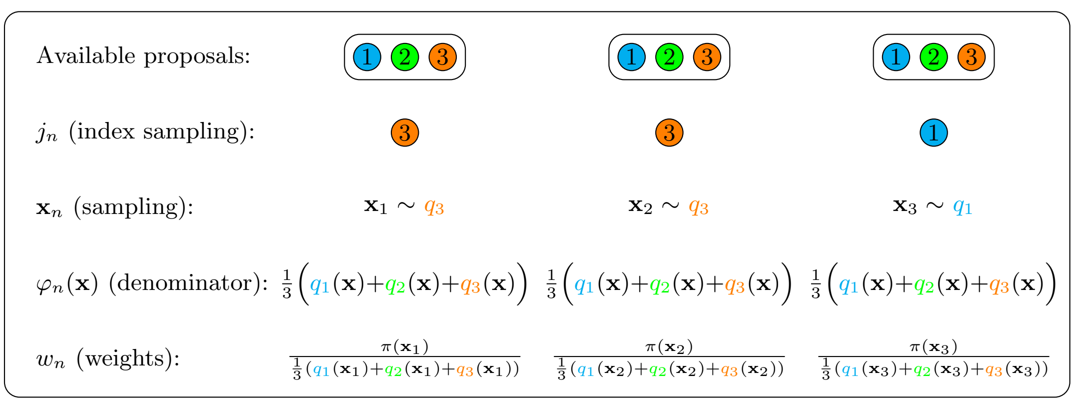
Let us consider the example from Section 3.4 where the realizations of the sequence of indexes for the samplings schemes $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$ are respectively $\{j_1, j_2, j_3\} = \{3, 3, 1\}$, $\{j_1, j_2, j_3\} = \{3, 1, 2\}$, and $\{j_1, j_2, j_3\} = \{1, 2, 3\}$. Figure 3 shows the three first schemes of Table 2 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 pdfs. For the n-th sample, we show the set of available proposals, the index j_n of the proposal pdf that was actually selected to draw the sample, the function φ_n, and the importance weight. Similarly, Fig. 4 depicts the three schemes of Table 2 related to the sampling without replacement, $\mathcal{S}_2$ and $\mathcal{S}_3$, where exactly one sample is drawn from each of the proposal pdfs of the available set.



(a) Scheme R1



(b) Scheme R2



(c) Scheme R3

FIG 3. *Example of a realization of the indexes selection ($N = 3$) with the procedure $\mathcal{S}_1$ (with replacement), and the corresponding possible denominators for weighting: (a) [R1]: $\varphi_n(\mathbf{x}_n) = q_{j_n}(\mathbf{x}_n)$; (b) [R2]: $\varphi_n(\mathbf{x}_n) = \frac{1}{N}\sum_{k=1}^{N} q_{j_k}(\mathbf{x}_n)$; (c) [R3]: $\varphi_n(\mathbf{x}_n) = \psi(\mathbf{x}_n) = \frac{1}{N}\sum_{k=1}^{N} q_k(\mathbf{x}_n)$.*

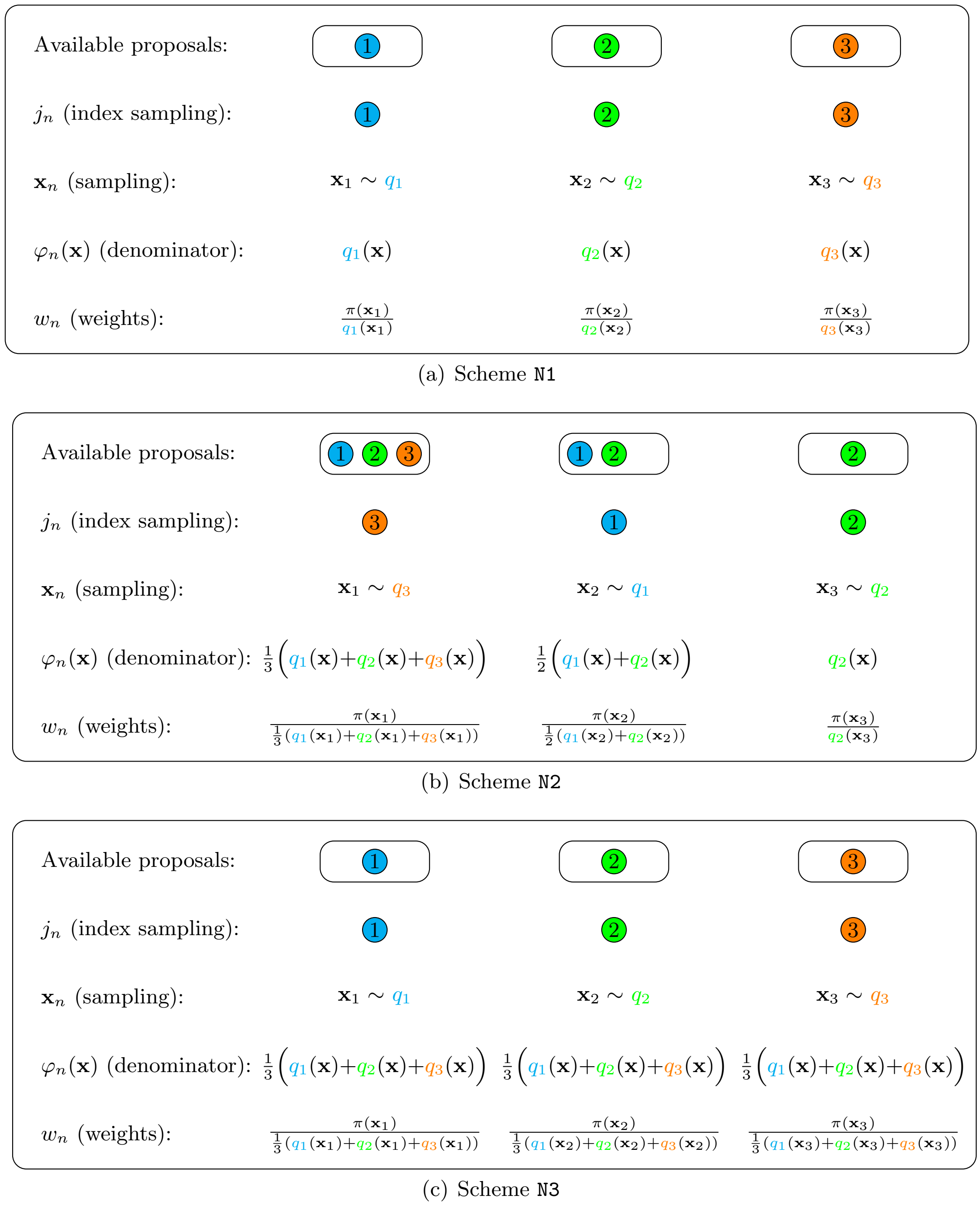


FIG 4. *Example of a realization of the indexes selection ($N = 3$) with the procedure $\mathcal{S}_1$ and $\mathcal{S}_2$ (without replacement), and the corresponding possible denominators for weighting: (a) [N1]: $\varphi_n(\mathbf{x}_n) = q_n(\mathbf{x}_n)$; (b) [N2]: $\varphi_n(\mathbf{x}_n) = \frac{1}{|\mathcal{I}_n|}\sum_{\forall k \in \mathcal{I}_n} q_k(\mathbf{x}_n)$; (c) [N3]: $\varphi_n(\mathbf{x}_n) = \frac{1}{N}\sum_{k=1}^{N} q_k(\mathbf{x}_n)$.*

6. VARIANCE ANALYSIS OF THE SCHEMES

In this section we provide an exhaustive variance analysis of the MIS schemes presented in the previous section. A central objective in importance sampling entails the computation of a particular moment of r.v. with pdf $\tilde{\pi}(\mathbf{x}) = \frac{\pi(\mathbf{x})}{Z}$. For sake of completeness of this section, let us revisit the general forms of the estimators. We recall that the goal is approximating

$$I = \frac{1}{Z}\int g(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}, \tag{6.1}$$

where g can be any square integrable function of $\mathbf{x}$ [Liu, 2004].

In standard importance sampling, the moment in Eq. (6.1) can be estimated by drawing N independent samples $\mathbf{x}_n$ from a single proposal density $q(\mathbf{x})$ and building the estimator

$$\tilde{I} = \frac{1}{N\hat{Z}}\sum_{n=1}^{N} w_n g(\mathbf{x}_n), \tag{6.2}$$

where $w_n = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}$ for $n = 1, ..., N$, and $\hat{Z} = \frac{1}{N}\sum_{j=1}^{N} w_j$. Under mild assumptions about the tails of the distributions, $\hat{I}$ provides a consistent estimator of I [Robert and Casella, 2004]. If the normalizing constant Z of the target $\pi(\mathbf{x})$ is known, the estimator

$$\hat{I} = \frac{1}{NZ}\sum_{n=1}^{N} w_n g(\mathbf{x}_n), \tag{6.3}$$

is also unbiased [Robert and Casella, 2004; Liu, 2004]. Furthermore, it is well known that the variance of both estimators is directly related to the discrepancy between $\tilde{\pi}(\mathbf{x})|g(\mathbf{x})|$ and $q(\mathbf{x})$ (for a specific choice of g) [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 $\tilde{\pi}(\mathbf{x})$.

In MIS, a set of N proposal pdfs $\{q_n(\mathbf{x})\}_{n=1}^{N}$ is used to draw the N samples. While the MIS estimators preserve the same structure as in Eqs. (6.2) and (6.3), the way the samples are drawn (see the sampling procedures in Section 3) and the function used for the weight calculation (see Section 4) can make a substantial difference in the performance. In fact, although the six different MIS schemes that appear in Section 5 yield an unbiased estimator (see Appendix B), the performance of that estimator can be dramatically different. In the following, we focus on the variance of the unbiased estimator $\hat{I}$ of Eq. (6.3) in all the studied schemes. The details of the derivations are in Appendix C.2. In particular, the estimators of the three methods with replacement present the following variances

$$\text{Var}(\hat{I}_{\text{R1}}) = \frac{1}{Z^2N^2}\sum_{k=1}^{N}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_k(\mathbf{x})}d\mathbf{x} - \frac{I^2}{N}, \tag{6.4}$$

$$\begin{aligned}\text{Var}(\hat{I}_{\text{R2}}) &= \frac{1}{Z^2N}\frac{1}{N^N}\sum_{j_{1:N}}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{f(\mathbf{x}|j_{1:N})}d\mathbf{x} \\ &-\frac{1}{Z^2N^2}\frac{1}{N^N}\sum_{j_{1:N}}\sum_{n=1}^{N}\left(\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{f(\mathbf{x}_n|j_{1:N})}q_{j_n}(\mathbf{x}_n)d\mathbf{x}_n\right)^2,\end{aligned} \tag{6.5}$$

and

$$\text{Var}(\hat{I}_{\texttt{R3}}) = \frac{1}{Z^2 N} \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{\psi(\mathbf{x})} d\mathbf{x} - \frac{I^2}{N}. \tag{6.6}$$

On the other hand, the variances associated to the estimators of the three methods without replacement are

$$\text{Var}(\hat{I}_{\texttt{N1}}) = \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{q_n(\mathbf{x}_n)} d\mathbf{x}_n - \frac{I^2}{N}, \tag{6.7}$$

$$\begin{aligned}\text{Var}(\hat{I}_{\texttt{N2}}) = &\left[\frac{1}{Z^2 N^2} \sum_{n=1}^{N} \sum_{j_{1:n-1}} \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{p(\mathbf{x}_n | j_{1:n-1})} P(j_{1:n-1}) d\mathbf{x}_n \right] \\ &- \left[\frac{1}{Z^2 N^2} \sum_{n=1}^{N} \sum_{j_{1:n}} \left(\int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{p(\mathbf{x}_n | j_{1:n-1})} q_{j_n} d\mathbf{x}_n \right)^2 \right] P(j_{1:n}),\end{aligned} \tag{6.8}$$

and

$$\text{Var}(\hat{I}_{\texttt{N3}}) = \frac{1}{Z^2 N} \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{\psi(\mathbf{x})} d\mathbf{x} - \frac{1}{Z^2 N^2} \sum_{n=1}^{N} \left(\int \frac{\pi(\mathbf{x}) g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x}) d\mathbf{x} \right)^2. \tag{6.9}$$

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,*

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

Proof: See Appendix C.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,*

$$Var(\hat{I}_{\texttt{R1}}) = Var(\hat{I}_{\texttt{N1}}) \geq Var(\hat{I}_{\texttt{R2}}) = Var(\hat{I}_{\texttt{N2}}) \geq Var(\hat{I}_{\texttt{N3}}) \tag{6.10}$$

Proof: See Appendix C.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.

The variance analysis of $\tilde{I}$ in Eq. (6.2) implies a ratio of dependent r.v.'s, and therefore, it cannot be performed without resorting to an approximation, e.g., by means of a Taylor expansion as it is performed in [Kong, 1992; Kong et al., 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, although we prove several relations in terms of the variance for $\hat{I}$ and $\hat{Z}$, the same conclusions for the normalized estimator $\hat{I}_{\text{MIS}}$ cannot strictly be proved in the general case. 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.

6.1 Running example: Estimation of the normalizing constant in MIS

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 pdfs 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 pdf $\pi(x) = \frac{1}{2}\mathcal{N}(x; \nu_1, c_1^2) + \frac{1}{2}\mathcal{N}(x; \nu_2, c_2^2)$ and set $\nu_1 = \mu_1$, $\nu_2 = \mu_2$, and $c_1^2 = c_2^2 = \sigma^2$. Then, both proposal pdfs can be seen as a whole mixture that exactly replicates the target, i.e., $\pi(x) = \frac{1}{2}q_1(x) + \frac{1}{2}q_2(x)$. 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 Eq. (4.2) and the estimator $\hat{I}$ of (4.1) 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

$$
\begin{aligned}
\text{Var}(\hat{Z}_{\texttt{R1}}) = \text{Var}(\hat{Z}_{\texttt{N1}}) &= \frac{3 + \exp\left(\frac{4\mu^2}{\sigma^2}\right)}{8} - \frac{1}{2} = \frac{\exp(36) - 1}{8} \approx 5.4 \cdot 10^{14}, \\
\text{Var}(\hat{Z}_{\texttt{R2}}) = \text{Var}(\hat{Z}_{\texttt{N2}}) &= \frac{3 + \exp\left(\frac{4\mu^2}{\sigma^2}\right)}{16} - \frac{1}{4} = \frac{\exp(36) - 1}{16} \approx 2.7 \cdot 10^{14},
\end{aligned}
$$

and

$$
\text{Var}(\hat{Z}_{\texttt{R3}}) = \text{Var}(\hat{Z}_{\texttt{N3}}) = 0.
$$

The variances of the estimators of the target mean (true value $I = \int x\pi(x)dx =$

0) are given by

$$\begin{aligned}\mathrm{Var}(\hat{I}_{\mathtt{R1}}) = \mathrm{Var}(\hat{I}_{\mathtt{N1}}) &= \frac{3(\sigma^2+\mu^2)}{8} + \frac{\sigma^2+9\mu^2}{8}\exp\left(\frac{4\mu^2}{\sigma^2}\right) \\ &= \frac{30}{8} + \frac{82}{8}\exp(36) \approx 4.42\cdot 10^{16}, \\ \mathrm{Var}(\hat{I}_{\mathtt{R2}}) = \mathrm{Var}(\hat{I}_{\mathtt{N2}}) &= \frac{3(\sigma^2+\mu^2)}{16} + \frac{\sigma^2+9\mu^2}{16}\exp\left(\frac{4\mu^2}{\sigma^2}\right) + \frac{\sigma^2}{4} \\ &= \frac{30}{16} + \frac{82}{16}\exp(36) + \frac{1}{4} \approx 2.21\cdot 10^{16}, \\ \mathrm{Var}(\hat{I}_{\mathtt{R3}}) &= \frac{\sigma^2+\mu^2}{2} = 5,\end{aligned}$$

and

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

The derivations can be found in Appendix C.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

In previous section, we have compared the MIS schemes in terms of performance, whereas here we discuss their computational complexity. Table 3 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 use N weighted samples where the general weight is given by Eq. (4.3).[7] Therefore, all of them 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 yield 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 many data in the likelihood function. However, in some other scenarios, e.g. 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 of MCMC methods, different 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

[7] We recall that, in general, one can draw $M = kN$ samples, with $k \geq 1$ and $k \in \mathbb{N}$.

the adaptivity is the bottleneck of the parallelization. The different six schemes discussed above 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 scheme `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.

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 propose a framework that limits the number of proposal evaluations, let us first define a partition of the set of the indexes of all proposals, $\{1, \ldots, N\}$, into P disjoint subsets of L elements (indexes), $\mathcal{J}_p$ with $p = 1, \ldots, P$, s.t.

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

where $\mathcal{J}_k \cap \mathcal{J}_q = \emptyset$ for all $k, q = 1, \ldots, P$ and $k \neq q$.[8] Therefore, each subset, $\mathcal{J}_p = \{j_{p,1}, j_{p,2}, \ldots, j_{p,L}\}$, contains L indexes, $j_{p,\ell} \in \{1, \ldots, N\}$ for $\ell = 1, \ldots, L$ and $p = 1, \ldots, P$.

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., 2015a]. 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 $\mathcal{S}_3$, i.e., 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 p-th 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., 2015a], 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 Eq. (6.3).

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 Eq. (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

[8] Note that, for 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

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

TABLE 3
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.

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 pdfs using the information of the past samples. In that way, they decrease the mismatch between the proposal and the target, and thus improve the performance of the MIS scheme [Cappé et al., 2004; Bugallo et al., 2015]. The sampling and weighting options, described in this work within a static framework for sake of simplicity, can be straightforwardly applied in the adaptive context.

More specifically, let us consider a set of proposal pdfs $\{q_{j,t}(\mathbf{x})\}$, with $j = 1, \ldots, J$ and $t = 1, \ldots, T$, 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 pdfs. A general adaptation procedure takes into account, at the t-th iteration, statistical information about the target pdf gathered in all of the previous iterations, $1, \ldots, t-1$, using one of the many algorithms that have been proposed in the literature [Cappé et al., 2008, 2004; Cornuet et al., 2012; Martino et al., 2015a; Elvira et al., 2017].

Hence, the sampling and the weighting procedures described in previous sections (and therefore the six MIS schemes considered in Section 5) can be directly applied to the whole set of N proposal pdfs. Moreover, in the adaptive context, when many proposals are considered (the number of proposals grow over the time), the *a priori* partition approach of Section 7.2 can be useful to limit the computational cost of the different MIS schemes.

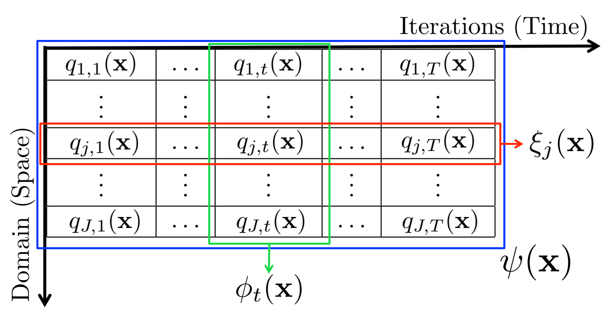


FIG 5. *Graphical representation of the $N = JT$ proposal pdfs used in the generalized adaptive MIS scheme, spread through the state space $\mathbb{R}^{d_x}$ ($j = 1, \ldots, J$) and adapted over time ($t = 1, \ldots, T$).*

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

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

$j = 1, \ldots, J$ 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. Fig. 5 provides a graphical representation of this scenario, by showing both the spatial and temporal evolution of the $J = NT$ proposal pdfs. 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 6.2. This method, with different choices of the partitioning of the set of proposals, implicitly appears in several methodologies that have been proposed independently in the literature of adaptive MIS algorithms. 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 pdfs. This case corresponds to the blue rectangle in Fig. 5. However, note that the computational complexity can become prohibitive if the product JT is increased. 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 pdf in the IS weight denominator, i.e. using the temporal evolution of the j-th 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 Fig. 5). Another possibility is considering the mixture of all the $q_{j,t}$'s at the t-th iteration, i.e.,

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

with $t = 1, \ldots, T$, so that we have $P = T$ mixtures, each one formed by $L = J$ components (green rectangle in Fig. 5). The function $\varphi_{j,t}$ in Eq. (7.4) is used in the standard PMC scheme [Cappé et al., 2004]; Eq. (7.6), in the particular case of $J = 1$, has been considered in the *adaptive multiple importance sampling* (AMIS) algorithm [Cornuet et al., 2012]. The choice in Eq. (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., 2015b], 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,b; Cappé et al., 2008], a similar strategy is employed, but using sampling $\mathcal{S}_1$ in the mixture $\phi_t(\mathbf{x})$, i.e., with the MIS scheme R3.

Table 7 summarizes all the possible cases discussed above. The last row corresponds to a generic grouping strategy of the proposal pdfs $q_{j,t}$. 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 p-th 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 $\mathcal{J}_p \cap \mathcal{J}_q = \emptyset$ for any pair $p, q = 1, \ldots, P$, and $p \neq q$. In this scenario, we have

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

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 t-th 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.

Through this subsection, we have shown that some of the most relevant adaptive MIS algorithms can be cast within the proposed generalized MIS framework. Besides this unifying perspective, new adaptive algorithms can be naturally proposed by modifying the sampling or the weighting schemes of the existing algorithms in the literature.

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 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., 2015a] and [Elvira et al., 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 Fig. 5). However, the computational complexity for large number of proposals is prohibitive, and further theoretical analysis about the bias of the estimators is need (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., 2015b], 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.

8. NUMERICAL EXAMPLES

In the previous sections, we have provided several theoretical results for comparing different MIS schemes according to different quality measures, e.g., 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 choice of sampling and the weighting procedure dramatically affects the performance of the MIS estimator.

8.1 Running exmple: Estimation of the target mean

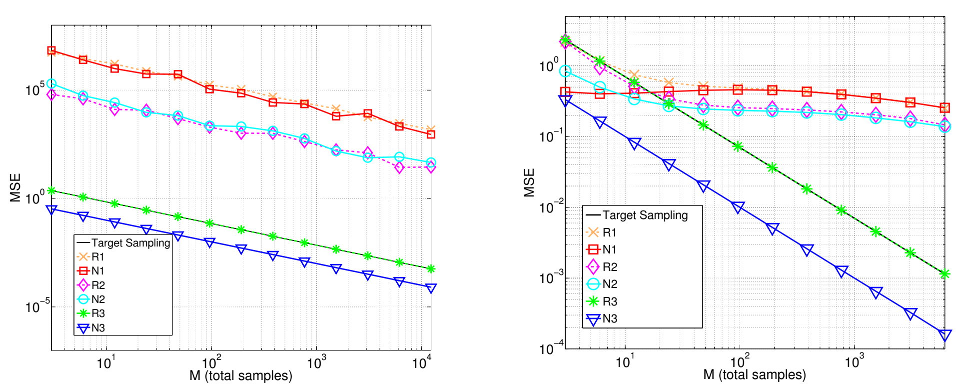
Let us consider again the target pdf of the running example

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

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 $\sigma^2 = 1$, i.e., the proposal pdfs can be seen as a whole mixture that exactly replicates the target, i.e., $\pi(x) = \psi(x) = \frac{1}{3}q_1(x) + \frac{1}{3}q_2(x) + \frac{1}{3}q_3(x)$.

The goal is estimating the mean of the target pdf with the six MIS schemes. Fig. 6(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, i.e. sampling directly from the target pdf (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 $w = 1$, i.e., $\hat{I}_{\mathtt{R3}}$ 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, i.e., 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 $\hat{I}_{\mathtt{N2}}$ behave also bad in terms of variance.

Figure 6(b) shows the variance of the estimator $\tilde{I}$ of Eq. (6.2) for all methods. First, note that the MSE of R3 and N3 is the same as in Fig. 6 (b), since the estimators $\hat{I}$ and $\tilde{I}$ are equivalent in this scenario (since they perfectly estimate the normalizing constant, i.e., $\hat{Z} = Z$). 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, i.e. the same number of target evaluations. Note that each MIS scheme requires a different number of proposal evaluations per sample (see Table 3). 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.



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

FIG 6. *(Ex. of Section. 8.1) Performance of the estimators of the target mean for the different MIS schemes.*

8.2 Multidimensional mixture of generalized Gaussian distributions

Let us consider a mixture of multivariate generalized Gaussian distributions (GGD) as a target pdf. In particular

$$\pi(\mathbf{x}) = \frac{1}{3}\sum_{k=1}^{3} \mathcal{GG}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k), \quad \mathbf{x} \in \mathbb{R}^{d_x}, \tag{8.1}$$

where $\boldsymbol{\mu}_k = [\mu_{k,1}, ..., \mu_{k,d_x}]^\top$, $\boldsymbol{\alpha}_k = [\alpha_{k,1}, ..., \alpha_{k,d_x}]^\top$, and $\boldsymbol{\beta}_k = [\beta_{k,1}, ..., \beta_{k,d_x}]^\top$ are respectively the mean, scale, and shape parameters of each component of the mixture. Each component of the mixture factorizes in all dimensions, i.e., the multivariate GGD pdf is the product of N unidimensional GGD pdfs. 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 $\kappa_{k,d} = \frac{\beta_{k,d}}{2\alpha_{k,d}\Gamma\left(\frac{1}{\beta_{k,d}}\right)}$, $\Gamma(\cdot)$ is the gamma function, and x_d is the d-th 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 $d = 1, ..., d_x$. The expected value of the target $\pi(\mathbf{x})$ is then $E_\pi[X_d] = 1$ for $d = 1, ..., d_x$. In order to study the performance of the different MIS schemes, we vary the dimension of the state space in Eq. (8.1) 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 $N = 500$ non-standardized 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 $n = 1, ..., N$ and $d = 1, ..., d_x$. For each method, we draw $M = kN$ samples, with $k = 32$, and we average all the results over 200 runs.

Fig. 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 in this example, the proposal pdfs are fixed in random locations of the space, which can be considered covered at low dimensions (since we are using $N = 500$ pdfs), but this coverage becomes worse as the dimension increases. This can probably explain the similar performance of all the methods in higher dimensions. If we performed adaptive MIS algorithms in order to adapt the proposal pdfs, we would expect that the MIS scheme `N3` outperformed the other methods substantially as in previous examples.

8.3 Applying the MIS schemes in adaptive IS (AIS)

We apply the different MIS schemes within an adaptive IS (AIS) context. In particular, we focus on the *layered adaptive importance sampling* (LAIS) algorithm, recently proposed in [Martino et al., 2015b]. 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 pdfs 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 pdfs. For instance, the `N3` scheme is

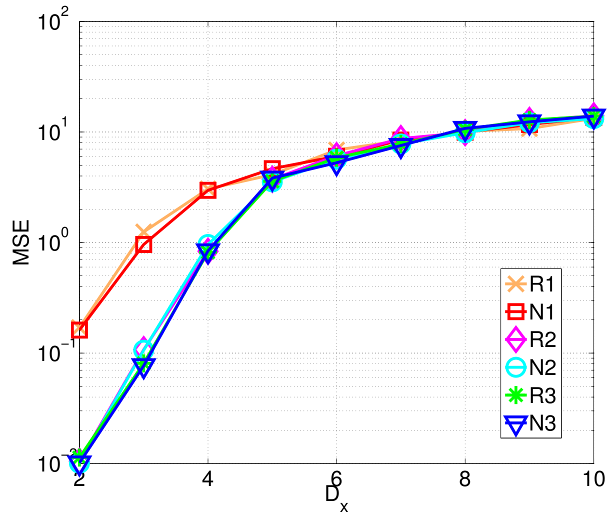


FIG 7. ***(Ex. of Section. 8.2)*** *MSE of the MIS estimator $\tilde{I}$ (normalized weights) for the different MIS schemes when we increase the dimension d_x of the state space.*

implemented by sampling exactly one sample from each of the J proposals at the t-th iteration, and applying at the denominator of the IS weight the whole mixture of J proposals as in Eq. (7.7) (see the green square of Fig. 5).

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

$$\pi(\mathbf{x}) = \frac{1}{5}\sum_{i=1}^{5}\mathcal{N}(\mathbf{x};\boldsymbol{\nu}_i,\boldsymbol{\Sigma}_i), \quad \mathbf{x}\in\mathbb{R}^2, \tag{8.2}$$

where $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\mathbf{C})$ denotes a Gaussian pdf with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$, $\nu_1 = [-10,-10]^\top$, $\boldsymbol{\nu}_2 = [0,16]^\top$, $\boldsymbol{\nu}_3 = [13,8]^\top$, $\boldsymbol{\nu}_4 = [-9,7]^\top$, $\boldsymbol{\nu}_5 = [14,-14]^\top$, $\boldsymbol{\Sigma}_1 = [2,\ 0.6;0.6,\ 1]$, $\boldsymbol{\Sigma}_2 = [2,\ -0.4;-0.4,\ 2]$, $\boldsymbol{\Sigma}_3 = [2,\ 0.8;0.8,\ 2]$, $\boldsymbol{\Sigma}_4 = [3,\ 0;0,\ 0.5]$, 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 4 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 $E[\mathbf{X}] = [1.6, 1.4]^\top$ and $Z = 1$, respectively). 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.

8.3.2 Multidimensional banana-shaped distribution. We consider the banana shape target example used in [Haario et al., 1999, 2001] which "can be be calibrated to become extremely challenging" [Cornuet et al., 2012]. The target is

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

TABLE 4
(Ex. of Section 8.3.1) *MSE of the LAIS and PMC algorithms with the different MIS schemes at the lower layer. $J = 100$ proposals and $T = 200$ iterations.*

based on a d_x-dimensional multivariate Gaussian $\mathbf{x} \sim \mathcal{N}(\mathbf{x}; \mathbf{0}_{d_x}, \mathbf{\Sigma})$ with $\Sigma = \text{diag}(\sigma^2, 1, ..., 1)$, 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 $d_x \geq 2$).

A preliminary attempt to apply static MIS schemes in this challenging example has shown that they cannot directly work without an adaptive procedure on top of them. We implement the MIS schemes within the LAIS algorithm as described in 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 Fig. 8, 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. Note that, if N were smaller, N3 would clearly outperform R3. When the dimensions increases, the performance of all schemes decays, but the same hierarchy in performance holds for all schemes. Note that N2 presents a similar performance than N3 in high dimensions.

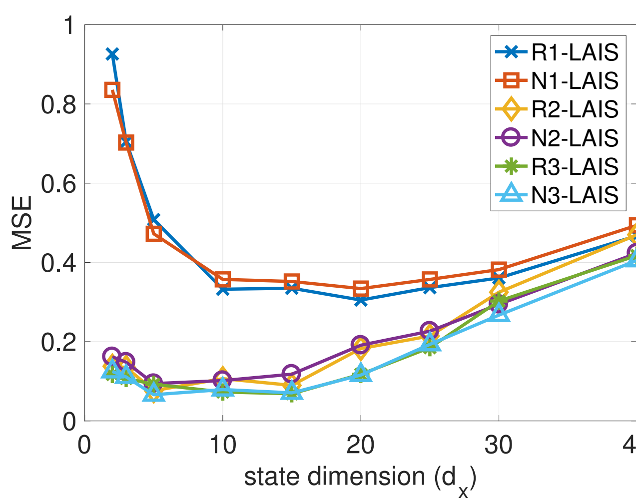


FIG 8. ***(Ex. of Section 8.3.2)*** *LAIS algorithm with different MIS schemes in a multidimensional banana-shaped target. $J = 100$ proposals and $T = 200$ iterations.*

8.4 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, 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 latter for the case of $N = 2$ proposals). The same hierarchy is represented graphically in Fig. 7. Furthermore, Fig. 6 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 Fig. 6 (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. A possible interpretation is the following: `N3` and `R3` apply 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. 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 C). The scheme `N3` goes a step further w.r.t. `R3`, drawing deterministically one sample from each mixand 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 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 $N > 2$, e.g. in Figs. 6(a) and 6(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 Fig. 6, when a small number of samples M is employed, the schemes `N1`, `N2` and `N3`, i.e., those with index selection without replacement ($\mathcal{S}_2$ and $\mathcal{S}_3$), behave better. This occurs because, in this case, the variance associated to the index selection is reduced by guaranteeing that all proposal pdfs 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.

REFERENCES

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer, 2004.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer, 2004.

A. Owen. *Monte Carlo theory, methods and examples.* http://statweb.stanford.edu/∼owen/mc/, 2013.

F. Liang. Dynamically weighted importance sampling in Monte Carlo computation. *Journal of the American Statistical Association*, 97(459):807–821, 2002.

E. Veach and L. Guibas. Optimally combining sampling techniques for Monte Carlo rendering. *In SIGGRAPH 1995 Proceedings*, pages 419–428, 1995.

T. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

A. Owen and Y. Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

Z. Tan. On a likelihood approach for monte carlo integration. *Journal of the American Statistical Association*, 99(468):1027–1036, 2004.

H. Y. He and A. B. Owen. Optimal mixture weights in multiple importance sampling. *arXiv preprint (arXiv:1411.3954)*, Nov. 2014.

V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Efficient multiple importance sampling estimators. *Signal Processing Letters, IEEE*, 22(10):1757–1761, 2015a.

O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert. Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929, 2004.

O. Cappé, R. Douc, A. Guillin, J. M. Marin, and C. P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18:447–459, 2008.

L. Martino, V. Elvira, D. Luengo, and J. Corander. An adaptive population importance sampler: Learning from the uncertainty. *IEEE Transactions on Signal Processing*, 63(16):4422–4437, 2015a.

J. M. Cornuet, J. M. Marin, A. Mira, and C. P. Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, December 2012.

M. F. Bugallo, L. Martino, and J. Corander. Adaptive importance sampling in signal processing. *(To appear) Digital Signal Processing*, 2015.

A Kong, P McCullagh, X-L Meng, D Nicolae, and Z Tan. A theory of statistical models for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):585–604, 2003.

N. Gordon, D. Salmond, and A. F. M. Smith. Novel approach to nonlinear and non-Gaussian Bayesian state estimation. *IEE Proceedings-F Radar and Signal Processing*, 140:107–113, 1993.

R. Douc and O. Cappé. Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pages 64–69. IEEE, 2005.

Herman Kahn and Andy W Marshall. Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278, 1953.

J. Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 24:1317–1399, 1989.

A. Kong. A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348, 1992.

A. Kong, J. S. Liu, and W. H. Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 9:278–288, 1994.

V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Heretical multiple importance sampling. *IEEE Signal Processing Letters*, 23(10):1474–1478, 2016.

V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Improving Population Monte Carlo: Alternative weighting and resampling schemes. *Signal Processing*, 131(12):77–91, 2017.

L. Martino, V. Elvira, D. Luengo, and J. Corander. Layered adaptive importance sampling. *Statistics and Computing*, pages 1–25, 2015b.

G.R. Douc, J.M. Marin, and C. Robert. Convergence of adaptive mixtures of importance sampling schemes. *Annals of Statistics*, 35:420–448, 2007a.

G.R. Douc, J.M. Marin, and C. Robert. Minimum variance importance sampling via population Monte Carlo. *ESAIM: Probability and Statistics*, 11:427–447, 2007b.

V. Elvira, L. Martino, L. Luengo, and J. Corander. A gradient adaptive population importance sampler. In *IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4075–4079, 2015b.

H. Haario, E. Saksman, and J. Tamminen. Adaptive proposal distribution for random walk metropolis algorithm. *Computational Statistics*, 14(3):375–396, 1999.

H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2): 223–242, April 2001.

H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods.* Society for Industrial Mathematics, 1992.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge Univ. Press, 1952.

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables.* Dover Pub., number 55., 1972.

P. W. Gwanyama. The HM-GM-AM-QM inequalities. *The College Mathematics Journal*, 35 (1):47–50, Jan. 2004.

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

Let us recall again the sampling procedure $\mathcal{S}_3$ (i.e., with deterministic selection of the index): $\mathbf{X}_n \sim q_n(\mathbf{x})$ for $n = 1, \ldots, N$. Note that the realizations (samples) $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are used in the importance sampling estimators regardless of their order. Then, we can interpret 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-Blackewillization (see [Owen, 2013, Section 9.12] for more details). More formally, if we define the r.v. $\mathbf{Z}$ equal to a r.v. uniformly chosen from the set $\{\mathbf{X}_n\}_{n=1}^{N}$, then $\mathbf{Z} \sim \psi(\mathbf{x})$. 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, 2004], 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 $j_1 = 1, j_2 = 2, \ldots, j_N = N$ (uniform, in the sense of low-discrepancy sequence [Niederreiter, 1992]) and then drawing $\mathbf{x}_n \sim q_{j_n}(\mathbf{x}) = q_n(\mathbf{x})$ for $n = 1, \ldots, N$. 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: PROOFS OF UNBIASEDNESS OF THE MIS ESTIMATORS

In this appendix we prove the unbiasedness of the estimator $\hat{I}$ of Eq. (4.1) 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

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

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 Eq. (B.1) over all indexes that do not affect the n-th weight ($j_{n:N}$):

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

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

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

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 Eq. (B.1), which cancels the denominator:

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

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 Eq. (B.1):

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

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

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

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:

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

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 Eq. (B.4), and cancelling the denominator:

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

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

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

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 3.6 for more details). Substituting $\varphi(\mathbf{x}) = \psi(\mathbf{x})$ in Eq. (B.5)

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

APPENDIX C: VARIANCE ANALYSIS OF THE MIS ESTIMATORS

Let us consider the unbiased estimator,

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

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

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

In the general case of Eq. (C.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, i.e.,

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

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

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

C.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 6), 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 Eq. (C.1) are independent, so we can use Eq. (C.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 C.3,

$$\begin{aligned}\text{Var}(\hat{I}_{\texttt{R1}}) &= \frac{1}{Z^2N^2}\sum_{n=1}^{N}\sum_{j_n=1}^{N}\left[\int \frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{p^2(\mathbf{x}_n|j_n)}p(\mathbf{x}_n|j_n)P(j_n)d\mathbf{x}_n\right] - \frac{I^2}{N} \\ &= \frac{1}{Z^2N^2}\sum_{n=1}^{N}\left[\int \sum_{j_n=1}^{N}\frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{q_{j_n}(\mathbf{x}_n)}P(j_n)d\mathbf{x}_n\right] - \frac{I^2}{N} \\ &= \frac{1}{Z^2N^2}\sum_{n=1}^{N}\left[\int \frac{1}{N}\sum_{k=1}^{N}\frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{q_k(\mathbf{x}_n)}d\mathbf{x}_n\right] - \frac{I^2}{N} \\ &= \frac{1}{Z^2N^2}\sum_{k=1}^{N}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_k(\mathbf{x})}d\mathbf{x} - \frac{I^2}{N},\end{aligned} \tag{C.5}$$

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

2. `[R2]` **Sampling 1 / Weighting 4:** The expression for the conditional independence of Eq. (C.4) is now 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 $j_{1:N}$:

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

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 Eq. (C.3), marginalize over the indexes, and substitute $\varphi_n(\mathbf{x}_n) = p(\mathbf{x}_n) = \psi(\mathbf{x}_n)$,

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

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 n-th 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 Eq. (C.3) over the indexes, and substituting $\varphi_n(\mathbf{x}_n) = p(\mathbf{x}_n) = q_n(\mathbf{x}_n)$:

$$\begin{aligned}
\mathrm{Var}(\hat{I}_{\mathtt{N1}}) &= \frac{1}{Z^2N^2}\sum_{n=1}^{N}\int \frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{\varphi_n^2(\mathbf{x}_n)}p(\mathbf{x}_n)d\mathbf{x}_n - \frac{1}{Z^2N^2}\sum_{n=1}^{N}\left(\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi_n(\mathbf{x}_n)}p(\mathbf{x}_n)d\mathbf{x}_n\right)^2 \\
&= \frac{1}{Z^2N^2}\sum_{n=1}^{N}\int \frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{q_n^2(\mathbf{x}_n)}q_n(\mathbf{x}_n)d\mathbf{x}_n - \frac{1}{Z^2N^2}\sum_{n=1}^{N}\left(\int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{q_n(\mathbf{x}_n)}q_n(\mathbf{x}_n)d\mathbf{x}_n\right)^2 \\
&= \frac{1}{Z^2N^2}\sum_{n=1}^{N}\int \frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{q_n(\mathbf{x}_n)}d\mathbf{x}_n - \frac{I^2}{N}.
\end{aligned} \tag{C.8}$$

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

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

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 Eq. (C.3) over all the set of indexes $j_{1:N}$, and substituting $\varphi_n(\mathbf{x}_n) = f(\mathbf{x}_n) = \psi(\mathbf{x}_n)$:

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

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

C.2 Proof of Theorem 6.1

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

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

Proof: See that Eqs. (C.5) and (C.8) are equivalent. □

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

Proof: Subtracting Eqs. (C.7) and (C.8), we get

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

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

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

Now, let us note that the left-hand side of Eq. (C.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 Eq. (C.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 Eq. (C.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 et al., 1952; Abramowitz and Stegun, 1972; Gwanyama, 2004]. □

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

Proof: Subtracting (C.7) and (C.10), we get

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

Therefore, the proposition is proved if

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

If we substitute I with the expression of Eq. (6.1), multiplying both numerator and denominator by $\psi(\mathbf{x})$ in the integral of the right-hand side,

$$\begin{aligned}
\frac{1}{Z^2}\sum_{n=1}^{N}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})d\mathbf{x}\right)^2 &\geq N\left(\frac{1}{Z}\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})}\psi(\mathbf{x})d\mathbf{x}\right)^2 \\
\sum_{n=1}^{N}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})d\mathbf{x}\right)^2 &\geq N\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})}\left(\frac{1}{N}\sum_{n=1}^{N} q_n(\mathbf{x})\right)d\mathbf{x}\right)^2 \\
\sum_{n=1}^{N}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})d\mathbf{x}\right)^2 &\geq \frac{1}{N}\left(\sum_{n=1}^{N}\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})d\mathbf{x}\right)^2 \\
N\sum_{n=1}^{N} a_n^2 &\geq \left(\sum_{n=1}^{N} a_n\right)^2 \qquad \text{(C.12)}
\end{aligned}$$

with $a_n = \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x})d\mathbf{x}$. The inequality of Eq. (C.12) holds, since it is the definition of the Cauchy-Schwarz inequality [Hardy et al., 1952],

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

with $b_n = 1$ for $n = 1, ..., N$. □

Proof of Theorem 6.1. The proof is obtained by applying Propositions C.1, C.2, and C.3. □

C.3 Proof of Theorem 6.2

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

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

From Eq. (C.7),

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

From Eq. (C.10),

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

From Eq. (C.6),

$$\begin{aligned}\mathrm{Var}(\hat{I}_{\texttt{R2}}) &= \frac{1}{8Z^2}\left(\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_1(\mathbf{x})}d\mathbf{x} + \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_2(\mathbf{x})}d\mathbf{x}\right) - \frac{I^2}{4} + \frac{1}{4Z^2}\int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}}d\mathbf{x} \\ &\quad - \frac{1}{8Z^2}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}}q_1(\mathbf{x})d\mathbf{x}\right)^2 - \frac{1}{8Z^2}\left(\int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}}q_2(\mathbf{x})d\mathbf{x}\right)^2.\end{aligned} \tag{C.17}$$

From Eq. (C.9),

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

PROPOSITION C.4. *For $N = 2$, $\mathrm{Var}(\hat{I}_{\texttt{R2}}) = \mathrm{Var}(\hat{I}_{\texttt{N2}})$*

Proof: See that Eqs. (C.17) and (C.18) are equivalent. □

PROPOSITION C.5. *For $N = 2$, $\mathrm{Var}(\hat{I}_{\texttt{N1}}) \geq \mathrm{Var}(\hat{I}_{\texttt{R2}}) \geq \mathrm{Var}(\hat{I}_{\texttt{N3}})$*

Proof: Analyzing Eqs. (C.14) and (C.16), we see that Eq. (C.17) can be rewritten as

$$\mathrm{Var}(\hat{I}_{\texttt{R2}}) = \frac{1}{2}\mathrm{Var}(\hat{I}_{\texttt{N1}}) + \frac{1}{2}\mathrm{Var}(\hat{I}_{\texttt{N3}}). \tag{C.19}$$

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

Proof of Theorem 6.2. The proof is obtained by applying Propositions C.4 and C.5. □

REMARK C.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 C.2. *Note that the scheme R3 does not appear in Theorem 6.2. Eq. (C.17) can be rewritten as*

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

The question is then whether the last four terms are larger than $\frac{1}{2}\mathrm{Var}(\hat{I}_{R3})$. We hypothesize that no inequality can be established in a general case, i.e., 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 $q_2(\mathbf{x})$.

C.4 Example with closed-form variances

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

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

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, i.e. $\psi(\mathbf{x}) = \pi(\mathbf{x})$. 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 Eq. (4.1) for the six MIS schemes. From Eq. (C.5),

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

Let us first compute

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

Since the proposals are Gaussian,

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

Then,

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

Similarly,

$$
S_2 = \frac{1}{4}\left[\int g^2(\mathbf{x})q_2(\mathbf{x})d\mathbf{x} + 2\int g^2(\mathbf{x})q_1(\mathbf{x})d\mathbf{x} + \int g^2(\mathbf{x})\frac{q_1(\mathbf{x})}{q_2(\mathbf{x})}q_1(\mathbf{x})d\mathbf{x}\right],
$$

where

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

Finally, from Eq. (C.21),

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

Note that $\mathrm{Var}(\hat{I}_{\mathtt{R1}}) = \mathrm{Var}(\hat{I}_{\mathtt{N1}})$. From Eq. (C.7),

$$
\begin{aligned}
\mathrm{Var}(\hat{I}_{\mathtt{R3}}) &= \frac{1}{2}\int \frac{g^2(\mathbf{x})\pi(\mathbf{x})}{\pi(x)}\pi(\mathbf{x})d\mathbf{x} - \frac{I}{2} \\
&= \frac{1}{2}\int g^2(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} - \frac{1}{2}\int g(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} \\
&= \frac{1}{2}\int g(\mathbf{x})(g(\mathbf{x})-1)\pi(x)d\mathbf{x}.
\end{aligned}
\tag{C.21}
$$

From Eq. (C.10),

$$
\mathrm{Var}(\hat{I}_{\mathtt{N3}}) = \frac{1}{2}\int g^2(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} - \frac{1}{4}\left[\left(\int g(\mathbf{x})q_1(\mathbf{x})d\mathbf{x}\right)^2 + \left(\int g(\mathbf{x})q_2(\mathbf{x})d\mathbf{x}\right)^2\right].
$$

From Eq. (C.19), $\mathrm{Var}(\hat{I}_{\mathtt{R2}}) = \frac{\mathrm{Var}(\hat{I}_{\mathtt{N1}})+\mathrm{Var}(\hat{I}_{\mathtt{N3}})}{2}$. Therefore,

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

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

TABLE 5
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. (3.5) and (3.7)
$J_n \vert J_{1:n-1} \sim P(j_n \vert j_{1:n-1})$	$\frac{1}{N}$	$\frac{1}{\vert\mathcal{I}_n\vert}\mathbb{1}_{j_n \in \mathcal{I}_n}$	$\mathbb{1}_{j_n=n}$	Eqs. (3.5)-(3.6)-(3.7)
$\mathbf{X}_n \vert J_{1:n-1} \sim p(\mathbf{x}_n \vert j_{1:n-1})$	$\psi(\mathbf{x}_n)$	$\frac{1}{\vert\mathcal{I}_n\vert}\sum_{\forall k \in \mathcal{I}_n} q_k(\mathbf{x}_n)$	$q_n(\mathbf{x}_n)$	Sect. 3.5
$\mathbf{X}_n \vert J_n \sim p(\mathbf{x}_n \vert j_n)$	$q_{j_n}(\mathbf{x}_n)$	$q_{j_n}(\mathbf{x}_n)$	$q_{j_n}(\mathbf{x}_n) = q_n(\mathbf{x}_n)$	Sect. 3.1
$\mathbf{X}_n \sim p(\mathbf{x}_n)$	$\psi(\mathbf{x}_n)$	$\psi(\mathbf{x}_n)$	$q_n(\mathbf{x}_n)$	Eq. (3.8)
$\mathbf{X} \vert J_{1:N} \sim f(\mathbf{x} \vert j_{1:N})$	$\frac{1}{N}\sum_{n=1}^{N} q_{j_n}(\mathbf{x})$	$\psi(\mathbf{x})$	$\psi(\mathbf{x})$	Eq. (3.11)
$\mathbf{X} \sim f(\mathbf{x})$	$\psi(\mathbf{x})$	$\psi(\mathbf{x})$	$\psi(\mathbf{x})$	Eq. (3.10)
$\mathbf{X}_{1:N} \sim p(\mathbf{x}_{1:N})$	$\prod_{n=1}^{N} \psi(\mathbf{x}_n)$	$\psi(\mathbf{x}_1)\prod_{n=2}^{N} \frac{1}{\vert\mathcal{I}_n\vert}\sum_{\ell \in I_n} q_\ell(\mathbf{x}_n)$	$\prod_{n=1}^{N} q_n(\mathbf{x}_n)$	Sect. 3.6; Eq. (3.12)

TABLE 6
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 3.6) and weighting functions (Section 4.2).

$\varphi_{\mathcal{P}_n}$	$\mathcal{W}_1$ $p(\mathbf{x}_n \vert j_{1:n-1})$	$\mathcal{W}_2$ $p(\mathbf{x}_n \vert j_n)$	$\mathcal{W}_3$ $p(\mathbf{x}_n)$	$\mathcal{W}_4$ $f(\mathbf{x} \vert j_{1:N})$	$\mathcal{W}_5$ $f(\mathbf{x})$
$\mathcal{S}_1$: **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)**	$\frac{1}{\vert\mathcal{I}_n\vert}\sum_{\forall k \in \mathcal{I}_n} q_k(\mathbf{x}_n)$ [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]

Table 7

Summary of possible MIS strategies in an adaptive framework.

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