Advances in Importance Sampling

Víctor Elvira⋆ and Luca Martino†
⋆ School of Mathematics, University of Edinburgh (United Kingdom)
† Universidad Rey Juan Carlos de Madrid (Spain)

Abstract

Importance sampling (IS) is a Monte Carlo technique for the approximation of intractable distributions and integrals with respect to them. The origin of IS dates from the early 1950s. In the last decades, the rise of the Bayesian paradigm and the increase of the available computational resources have propelled the interest in this theoretically sound methodology. In this paper, we first describe the basic IS algorithm and then revisit the recent advances in this methodology. We pay particular attention to two sophisticated lines. First, we focus on multiple IS (MIS), the case where more than one proposal is available. Second, we describe adaptive IS (AIS), the generic methodology for adapting one or more proposals.

Keywords: Monte Carlo methods, computational statistics, importance sampling.

1   Problem Statement

In many problems of science and engineering intractable integrals must be approximated. Let us denote an integral of interest

$$I(f) = E_{\widetilde{\pi}}[f(\mathbf{x})] = \int f(\mathbf{x})\bar{\pi}(\mathbf{x})d\mathbf{x}, \tag{1}$$

where $f : \mathbb{R}^{d_x} \to \mathbb{R}$, and $\widetilde{\pi}(\mathbf{x})$ is a distribution of the r.v. $\mathbf{X} \in \mathbb{R}^{d_x}$.[1] Note that although Eq. (1) involves a distribution, more generic integrals could be targeted with the techniques described below.

The integrals of this form appear often in the Bayesian framework, where a set of observations are available in $\mathbf{y} \in \mathbb{R}^{d_y}$, and the goal is in inferring some hidden parameters and/or latent variables $\mathbf{x} \in \mathbb{R}^{d_y}$ that are connected to the observations through a probabilistic model [61]. The information provided by the observations is compacted in the likelihood function $\ell(\mathbf{y}|\mathbf{x})$ and the prior knowledge on $\mathbf{x}$ is encoded in the prior distribution $p_0(\mathbf{x})$. Both sources of information are fused to create through the simple Bayes' rule the posterior probability density function (pdf), also called target distribution, given by

$$\widetilde{\pi}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\mathbf{x})p_0(\mathbf{x})}{Z(\mathbf{y})}, \tag{2}$$

---

[1]For the sake of easing the notation, from now on we use the same notation for denoting a random variable or one realization of a random variable.

Table 1: Summary of the notation.

d_x	dimension of the inference problem, $\mathbf{x} \in \mathbb{R}^{d}_x$
d_y	dimension of the observed data, $\mathbf{y} \in \mathbb{R}^{d}_y$
$\mathbf{x}$	r.v. of interest; parameter to be inferred
$\mathbf{y}$	observed data
$\ell(\mathbf{y}|\mathbf{x})$	likelihood function
$p_0(\mathbf{x})$	prior pdf
$\widetilde{\pi}(\mathbf{x})$	posterior pdf (target), $\widetilde{\pi}(\mathbf{x}) \equiv p(\mathbf{x}|\mathbf{y})$
$\pi(\mathbf{x})$	posterior density function (unnormalized target) , $\pi(\mathbf{x}) \equiv \ell(\mathbf{y}|\mathbf{x})g(\mathbf{x}) \propto \bar{\pi}(\mathbf{x})$
$q(\mathbf{x})$	proposal density
Z	normalizing constant or marginal likelihood, $Z \equiv Z(\mathbf{y})$
$I(f)$	integral to be approximated, $I(f) \equiv E_{\widetilde{\pi}}[f(\mathbf{x})]$.

where $Z(\mathbf{y}) = \int \ell(\mathbf{y}|\mathbf{x})p_0(\mathbf{x})d\mathbf{x}$ is the marginal likelihood (a.k.a., partition function, Bayesian evidence, model evidence, or normalizing constant) [7, 88]. In most models of interest $Z(\mathbf{y})$ is unknown and in many applications it must be approximated [7, 9, 88]. But even when its approximation is not needed, the unavailability of $Z(\mathbf{y})$ implies that the posterior can be evaluated only up to that (unknown) constant, i.e., we can only evaluate

$$\pi(\mathbf{x}) = \ell(\mathbf{y}|\mathbf{x})p_0(\mathbf{x}), \tag{3}$$

that we denote as unnormalized target distribution.[2] Table 1 summarizes the notation of this article.

The integral $I(f)$ cannot be computed in a closed form in many practical scenarios and hence must be approximated. The approximation methods can be divided into either deterministic or stochastic. While many deterministic numerical methods are available in the literature [1, 5, 11, 59, 87], it is in general accepted that they tend to become less efficient than stochastic approximations when the problem dimension d_x grows.

1.1 Standard Monte Carlo integration

The Monte Carlo approach consists in approximating the integral $I(f)$ in Eq. (1) with random samples [22, 51, 44, 58, 63, 78, 91]. In the standard Monte Carlo solution (often called instinctively vanilla/raw/classical/direct Monte Carlo), N samples $\mathbf{x}_n$ are independently simulated from $\widetilde{\pi}(\mathbf{x})$. The standard Monte Carlo estimator is built as

$$\overline{I}^N(f) = \frac{1}{N} \sum_{t=1}^{N} f(\mathbf{x}_n). \tag{4}$$

First, note that $\overline{I}^N(f)$ is unbiased since $E_{\widetilde{\pi}}[\overline{I}^N(f)] = I(f)$. Moreover, due to the weak law of large numbers, it can be shown that $\overline{I}_N$ is consistent and then converges in probability to the true value

---

[2]From now on, we drop $\mathbf{y}$ to ease the notation, e.g., $Z \equiv Z(\mathbf{y})$.

I, i.e., $\overline{I}^N(f) \xrightarrow{p} I(f)$, which is equivalent to stating that, for any positive number $\epsilon > 0$, we have $\lim_{N \to \infty} \Pr(|\overline{I}^N(f) - I(f)| > \epsilon) = 0$. The variance of $\overline{I}^N(f)$ is simply $\overline{\sigma}^2 = \frac{1}{N}(I(f^2) - I(f)^2)$. If the second moment is finite, $I(f^2) < \infty$, then the central limit theorem (CLT) applies and the estimator converges in distribution to a well-defined Gaussian when N grows to infinity i.e.,

$$\sqrt{N}\left(\overline{I}^N(f) - I(f)\right) \xrightarrow{d} \mathcal{N}(0, \overline{\sigma}^2). \tag{5}$$

There exist multiple families of Monte Carlo methods [91, 89, 100]. We address the interested reader to the articles in Markov chain Monte Carlo (including Metropolis-Hastings [90, 67] and Gibbs sampling [15]) and previous articles in importance sampling [104, 55].

2 Importance sampling

2.1 Origins

The first use of the importance sampling (IS) methodology dates from 1950 for rare event estimation in statistical physics, in particular for the approximation of the probability of nuclear particles penetrating shields [52]. IS was later used as a variance reduction technique when standard Monte Carlo integration was not possible and/or not efficient [48]. The renewed interest in IS has run in parallel with the hectic activity in the community of Bayesian analysis and its ever increasing computational demands. In most cases, the posterior in (2) is not available due to the intractability of the normalizing constant. See [101] for a previous review in IS.

2.2 Basics

Let us start defining the proposal pdf, $q(\mathbf{x})$, used to simulate the samples. It is widely accepted that the proposal is supposed to have heavier tails than the target, i.e., the target $\widetilde{\pi}(\mathbf{x})$ decays faster than $q(\mathbf{x})$ when $\mathbf{x}$ is *far* from the region where most of the probability mass is concentrated. However, this usual restriction is too vague and it will be clarified below. Here, we simply stick to the restriction that $q(\mathbf{x}) > 0$ for all $\mathbf{x}$ where $\widetilde{\pi}(\mathbf{x})f(\mathbf{x}) \neq 0$. IS is constituted of two simple steps:

1. **Sampling**: N samples are simulated as

$$\mathbf{x}_n \sim q(x), \qquad n = 1, ..., N. \tag{6}$$

2. **Weighting**: Each sample receives an associated importance weight given by

$$w_n = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}, \quad n = 1, \ldots, N. \tag{7}$$

The importance weights describe how representative the samples simulated from $q(\mathbf{x})$ are when one is interested in computing integrals w.r.t. $\widetilde{\pi}(\mathbf{x})$. The set of N weighted samples can be used to approximate the generic integral $I(f)$ of Eq. (1) by the two following IS estimators:

- Unnormalized (or nonnormalized) IS (UIS) estimator:

$$\widehat{I}^N(f) = \frac{1}{NZ} \sum_{n=1}^{N} w_n f(\mathbf{x}_n). \tag{8}$$

Note that the UIS estimator can be used only when Z is known.

- Self-normalized IS (SNIS) estimator:

$$\widetilde{I}^N(f) = \sum_{n=1}^{N} \overline{w}_n f(\mathbf{x}_n), \tag{9}$$

where

$$\overline{w}_n = \frac{w_n}{\sum_{j=1}^{N} w_j} \tag{10}$$

are the normalized weights.

The derivation of the SNIS estimator departs from the UIS estimator of Eq. (8), substituting Z by its unbiased estimator [91]

$$\widehat{Z} = \frac{1}{N} \sum_{n=1}^{N} w_n. \tag{11}$$

After a few manipulations, one recovers Eq. (10). The normalized weights also allow to approximate the target distribution by

$$\widetilde{\pi}^N(\mathbf{x}) = \sum_{n=1}^{N} \overline{w}_n \delta(\mathbf{x} - \mathbf{x}_n), \tag{12}$$

where δ represents the Dirac measure.

2.3   Theoretical analysis

The UIS estimator is unbiased since it can be easily proven that $\mathrm{E}_q[\widehat{I}^N(f)] = I(f)$. Its variance $\mathrm{E}_q[\widehat{I}^N(f)] = \frac{\sigma_q^2}{N}$ is given by

$$\sigma_q^2 = \int \frac{(f(\mathbf{x})\widetilde{\pi}(\mathbf{x}) - I(f)q(\mathbf{x}))^2}{q(\mathbf{x})} d\mathbf{x}, \tag{13}$$

if $q(\mathbf{x}) > 0$ for all $\mathbf{x}$ where $\widetilde{\pi}(\mathbf{x})f(\mathbf{x}) \neq 0$, as we have stated above [84]. We remark that it is not strictly necessary to have a proposal with heavier tails than the target distribution as long as $\sigma_q^2 < \infty$. One counter example is a case where $f(\mathbf{x})$ decays fast enough to compensate the heavier tails of the target distribution. Another counter example is a case where $f(\mathbf{x})$ takes non-zero and finite values only in a bounded set.

Note that $q(\mathbf{x})$ is chosen by the practitioner and a good choice is critical for the efficiency of IS. Let us first suppose that $\text{sign}(f(\mathbf{x}))$ is constant for all $\mathbf{x}$ and $I(f) \neq 0$. Let us also suppose that it is possible to simulate from

$$q^*(\mathbf{x}) = \frac{f(\mathbf{x})\widetilde{\pi}(\mathbf{x})}{\int f(\mathbf{z})\widetilde{\pi}(\mathbf{z})d\mathbf{z}}. \tag{14}$$

Then, the UIS estimator, for any $N \geq 1$ number of samples, yields a zero-variance unbiased estimator, since the numerator in (13) is zero, and hence $\sigma_q^2 = 0$. However, it is very unlikely to have access to the proposal of (14). The main reason is that its normalizing constant is exactly the intractable integral we are trying to approximate, $I(f)$. However, $q^*(\mathbf{x})$ gives the useful intuition that the proposal should have mass proportional to the targeted integrand in Eq. (1). More precisely, inspecting (13), we see that the efficiency is penalized with the mismatch of $f(\mathbf{x})\widetilde{\pi}(\mathbf{x})$ and $q(\mathbf{x})$, with this penalization amplified inversely proportional to the density $q(\mathbf{x})$. This explains the usual safe practice of over-spreading the proposal. The case where $\text{sign}(f(x))$ alternates can be easily modified by splitting the function as $f(x) = f_+(x) + f_-(x)$, where $f_+(x)$ is non-negative and $f_-(x)$ is non-positive. It is easy to show that with the use of two proposals and $N = 2$, a zero-variance estimator is possible (see [84, Section 9.13]). In summary, the UIS estimator, $\widehat{I}^N(f)$ is unbiased, while the $\widetilde{I}^N(f)$ is only asymptotically unbiased, i.e., with a bias that goes to 0 when N grows to infinity. Both UIS and SNIS are consistent estimators of I with a variance that depends on the discrepancy between $\pi(\mathbf{x})|f(\mathbf{x})|$ and $q(\mathbf{x})$, although the variance of the SNIS is more difficult to evaluate and its bias place also a central role when N is not large enough [84].

When several different moments f of the target must be estimated, a common strategy in IS is to decrease the mismatch between the proposal $q(\mathbf{x})$ and the target $\widetilde{\pi}(\mathbf{x})$ [20]. This is equivalent to minimizing the variance of the weights and consequently the variance of the estimator $\widehat{Z}$, and it is closely linked to the diagnostics of Section 2.4.

2.4   Diagnostics

It is a legitimate question to wonder about the efficiency of the set of simulated weighted samples in the task of approximating the target distribution and/or moments of it. Usual metrics of efficiency involve the computation of the variance of the IS estimators. However, the computation of those variances is intractable, and even more, their approximation is usually a harder problem than computing Eq. (1) (see [84, Chapter 9.3] for a discussion). A classic diagnostic metric in the IS literature [54] is

$$\widehat{\text{ESS}} = \frac{1}{\sum_{n=1}^{N} \bar{w}_n^2}. \tag{15}$$

Note that $1 \leq \widehat{\text{ESS}} \leq N$, taking the value $\widehat{\text{ESS}} = 1$, when one $w_j = 1$ and hence $w_i = 0$, for all $i \neq j$. Therefore, $\widehat{\text{ESS}} = N$ only when $w_j = 1/N$ for all $j = 1, ..., N$. Hence $\widehat{\text{ESS}}$ measures the discrepancy among normalized weights. This diagnostic is commonly called *effective sample size*, although it is an approximation of the more reasonable but intractable diagnostic given by [37]

$$\text{ESS}^* = N \frac{\text{Var}[\overline{I}^N]}{\text{MSE}[\widetilde{I}^N]}. \tag{16}$$

5

Then, ESS* can be interpreted as the number of standard Monte Carlo that are necessary to obtain the same performance (in terms of MSE) as with the SNIS estimator with N samples. The interested reader can find the derivation from ESS* to $\widehat{\text{ESS}}$ through a series of approximations and assumptions that rarely hold (see [37] for a thorough analysis). In practice, a low $\widehat{\text{ESS}}$ is a symptom of malfunctioning, but a high $\widehat{\text{ESS}}$ does not necessarily imply good behavior of the IS method.

New ESS-like methods have been proposed in the last years. In [72, 70], novel discrepancy measures with similar properties to $\widehat{\text{ESS}}$ are proposed and discussed, mitigating some of the deficiencies of the original diagnostic. For instance, an alternative to ESS* is using $1/\max(\bar{w}_n)$ instead, which preserves some of those properties (e.g., it takes values between 1 and N, being 1 if all the normalized weights are zero except one, and N if all weights are the same). Another metric in the same spirit has been recently proposed in [49]. Finally, the use of the importance trick within quadrature schemes has been recently proposed [38, 42]. Note that these *importance quadrature* schemes are not stochastic but strongly inspired in IS and its variants.

2.5   Other IS schemes

The research in IS methods has been very active in the last decade not only in the development of novel methodology but also for increasing the understanding and the theoretical behavior of IS-based methods. For instance, [2] unifies different perspectives about how many samples are necessary in IS for a given proposal and target densities, a problem that is usually related to some notion of distance (more precisely divergence) between the two densities. With a similar aim, in [14] it is shown that in a fairly general setting, IS requires a number of samples proportional to the exponential of the KL divergence between the target and the proposal densities. The notion of divergences between both densities is also explored in [92] through the Rényi generalized divergence, and in [79] in terms of the Pearson χ^2 divergence. Both divergences are connected with the variance of the $\widehat{Z}$ estimator in Eq. (11).

2.5.1   Transformation of the importance weights

As described in Section 2.4, a large variability in the importance weights is usually responsible for a large variance in the IS estimators. One alternative is adapting the proposals in order to diminish the mismatch with the target, as we describe in Section 4. However, this usually means throwing away past weighted samples (or stick to large variance estimators from the early iterations). Another alternative is the nonlinear transformation of the IS weights. The first work in this line is the *truncated importance sampling* [50] where the standard unnormalized weights w_n are truncated as $w'_n = \min(w_n, \tau)$, where τ is a maximum value allowed for the transformed/truncated weights. The consistency of the method and a central limit theorem of the modified estimator are proved. This transformation of the weights was also proposed in [53], and called *nonlinear importance sampling* within an adaptive IS scheme (N-PMC algorithm). The convergence of this method is analyzed in [53, 79, 80]. The underlying problem that those methods fight is the right heavy tail in the distribution of the importance weights when the proposal is not well fit. In [103], the authors go a step beyond by characterizing the distribution of the importance weights with

generalized Pareto distribution that fits the upper tail. Based on this fitting, a method is proposed for the stabilization of the importance weights. The authors provide proofs for consistency, finite variance, and asymptotic normality. See [76] for a review of the clipping methodologies.

2.5.2 Particle filtering (sequential Monte Carlo)

Particle filtering (also known as sequential Monte Carlo) is an IS-based methodology for performing approximate Bayesian inference on a hidden state that evolves over the time in state-space models, a class of probabilistic Markovian models. Due to the structure of the Bayesian network, it is possible to process sequentially and efficiently the observations related to the hidden state for building the sequence of filtering distributions (i.e., the posterior distribution of a given hidden state conditioned to all available observations). Particle filters (PFs) are based on importance sampling, incorporating in most cases a resampling step that helps to increase the diversity of the particle approximation [18, 62]. Since the publication of the seminal paper [45] where the bootstrap PF is developed (BPF), a plethora of PFs have been proposed in the literature [21, 86, 57, 17, 36]. Advanced MIS and AIS techniques are often implicit in those algorithms, but they are rarely explicit. In [39], a novel perspective of BPF and auxiliary PF (APF) based on MIS is introduced. In these state-space models, the ESS and its approximations are also used as diagnostics metrics for PF (see Section 2.4). Moreover, since the observations are dependent in these models, other metrics have been recently developed, mostly based on the predictive distribution of the observations [60, 8, 35, 41].

3   Multiple importance sampling (MIS)

The IS methodology can be easily extended when the samples are simulated from M proposals, $\{q_m(\mathbf{x})\}_{m=1}^M$, instead of only one. In a generic setting, one can consider that n_m samples are simulated from each proposal ($\sum_{j=1}^m n_j = 1$) and weighted appropriately. This extension is usually called multiple importance sampling (MIS), and it has strong connections with the case of standard IS with a single mixture proposal with components that are distributions, which is sometimes called mixture IS. Here we consider mixture IS as a subset of MIS methods when n_m are not deterministic number of samples but r.v.'s instead.

3.1   Generalized MIS

A unifying framework of MIS has been recently proposed in [40]. The framework encompasses most of existing IS methods with multiple proposals, proposes new schemes, and compares them in terms of variance. For the sake of clarity, the framework is described in the case where (a) no prior information about the adequateness of the proposals is available; and (b) $M = N$ proposals are available (i.e., exactly the same number of proposals than samples to be simulated). However, straightforward extensions are possible to more generic settings. According to this framework, a MIS is proper if it fulfills two conditions related to the sampling and weighting processes. A valid sampling scheme for the simulation of N samples, $\{\mathbf{x}_n\}_{n=1}^N$, can be agnostic to the dependence of those samples but must fulfill the following statistically property: a sample $\mathbf{x}$ randomly picked

from the whole set of N simulated samples must be distributed as the mixture of proposals $\psi(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} q_n(\mathbf{x})$. A valid weighting scheme must yield an unbiased and consistent UIS estimator, $\widehat{I}^N$. These properness conditions extend the standard properness in IS established by [63], and have been also used to assign proper importance weights to resampled particles [71]. The paper analyzes and ranks several resulting MIS schemes (different combination of valid sampling and weighting procedures) in terms of variance. Due to space restrictions, here we show only two MIS schemes commonly used in the literature. Let us simulate exactly one sample per proposal (sampling scheme $\mathcal{S}_3$ in [40]) as

$$\mathbf{x}_n \sim q_n(\mathbf{x}), \qquad n = 1, ..., N. \tag{17}$$

The next two weighting schemes are possible (among many others):

- **Option 1**: Standard MIS (s-MIS, also called `N1` scheme):

$$w_n = \frac{\pi(\mathbf{x}_n)}{q_n(\mathbf{x}_n)}, \quad n = 1, \ldots, N. \tag{18}$$

- **Option 2**: Deterministic mixture MIS (DM-MIS, also called `N3` scheme):

$$w_n = \frac{\pi(\mathbf{x}_n)}{\psi(\mathbf{x}_n)} = \frac{\pi(\mathbf{x}_n)}{\frac{1}{N} \sum_{j=1}^{N} q_j(\mathbf{x}_n)}, \quad n = 1, \ldots, N.$$

In both cases, it is possible to build the UIS and SNIS estimators. In [40], it is shown that

$$\mathrm{Var}[\widetilde{I}^N_{\texttt{N3}}] \leq \mathrm{Var}[\widetilde{I}^N_{\texttt{N1}}],$$

i.e., that using the second weighting option with the whole mixture in the denominator is always better than using just the proposal that simulated the sample (the equality in the variance relation happens only when all the proposals are the same). The result is relevant since `N1` is widely used in the literature but it should be avoided whenever possible. Note that both `N1` and `N3` require just one target evaluation per sample. However, `N3` requires N proposal evaluations per sample, while `N1` just one. For a small number of proposals, or when the target evaluation is very expensive (and hence the bottleneck), this extra complexity in `N3` may be not relevant, but it can become cumbersome otherwise. Several MIS strategies have been proposed in the literature to alleviate this problem. In [29], a partition of the proposals is done a priori, and then the `N3` scheme is applied within each cluster (i.e., small mixtures appear in the denominator of the weights). This method is called *partial deterministic mixture* and in some examples is able a similar variance reduction as in the `N3` method, while reducing drastically the number of proposal evaluations (see [29, Fig. 1]). The *overlapped partial deterministic mixture* method [32] extends the framework to the case where the proposals can belong to more than one cluster. However, the way the proposals are clustered remains an open problem and few attempts have been done for optimizing

8

the clustering (see [31] where the clusters are done after the sampling, using the information of the samples, and hence biasing the estimators).

When the selection of the proposals is also random, unlike in the sampling in (17), there exist options to evaluate only the proposals that have been used for sampling (scheme R2 in [40]) instead of using all of them in the numerator (scheme R3 in [40]). A recent paper explores the R2 scheme and some of its statistical properties [77].

3.1.1 MIS with different number of samples per proposal

Since the seminal works of [102, 48] in the computer graphics community, several works have addressed the case where the number of samples (also called counts) per proposal (also called techniques) can be different (see also [83] where the authors introduce control variates in MIS). In particular, the so-called *balance heuristic* estimator, proposed in [102] and very related to the scheme N3 in Section (3.1) has attracted attention due to its high performance. The UIS balance heuristic estimator is given by

$$\widehat{I}^N(f) = \sum_{j=1}^{M} \sum_{i=1}^{n_j} \frac{f(\mathbf{x}_{j,i})\bar{\pi}(\mathbf{x}_{j,i})}{\sum_{k=1}^{M} n_k q_k(\mathbf{x}_{j,i})}, \tag{19}$$

where again $\{q_m(\mathbf{x})\}_{m=1}^{M}$ is the set of available proposals, $\{n_m\}_{m=1}^{M}$ is the number of samples associated to each proposal, $N = \sum_{k=1}^{M} n_k$ is the total number of samples, and $\mathbf{x}_{j,i} \sim q_j(\mathbf{x})$, for $i = 1, ..., n_j$, and for $j = 1, ..., M$. Regarding the denominator in (19), it can be interpreted that the N samples are simulated from the mixture $\sum_{k=1}^{M} n_k q_k(\mathbf{x})$ via stratified sampling (a similar interpretation can be done in the aforementioned N3 scheme). In [96], this estimator is re-visited and novel bounds are obtained. In [93], the balance heuristic estimator of Eq. (19) is generalized, introducing more degrees of freedom that detach the sampling and the denominator of the importance weights, being able to obtain unbiased estimators that reduce the variance with respect to the standard balance heuristic. In [47], control variates are introduced in an IS scheme with a mixture proposal (similarly to [83]), and all parameters (including the mixture weights) are optimized to minimize the variance of the UIS estimator (which is jointly convex w.r.t. the mixture probabilities and the control variate regression coefficients). More works with a variable number of samples per proposal (either fixed or optimized) include [95, 94, 98].

3.2 Rare event estimation

Importance sampling is often considered as a variance reduction technique, not only in the case when sampling from $\widetilde{\pi}$ is not possible, but also when it is possible but not efficient. A classical example is the case of Eq. (1) when $f(\mathbf{x}) = \mathbb{I}_{\mathcal{S}}$, where $\mathbb{I}$ is the indicator function taking value 1 for all $\mathbf{x} \in \mathcal{S}$, and 0 otherwise. In rare event estimation, $\mathcal{S}$ is usually a set where the target $\widetilde{\pi}$ has few probability mass, and hence I is a small positive number. It is then not practical to simulate from the target, since most of the samples will not contribute to the estimator due to their evaluation in $\mathbb{I}_{\mathcal{S}}(\mathbf{x})$ being zero. IS allows for sampling from a different distribution that will increase the efficiency of the method when $q(\mathbf{x})$ is close to $\mathbb{I}_{\mathcal{S}}$. A recent MIS method called ALOE ("At least

one sample") is able to simulate from a mixture of proposals ensuring that all of them are in the integration region $\mathcal{S}$ in the case where $\widetilde{\pi}(\mathbf{x})$ is Gaussian and $\mathcal{S}$ is the union of half-spaces defined by a set of hyperplanes (linear constraints) [85]. As an example, the authors show successful results in a problem with 5772 constraints, in a 326-dimensional problem with a probability of $I \approx 10^{-22}$, with just $N = 10^4$ samples. ALOE has been recently applied for characterizing wireless communications systems through the estimation of the symbol error rate [27, 28].

3.3 Compressed and distributed IS

In the last years, several works have focused on alleviating the computational complexity, communication, or storage in intensive IS methods. This computational burden appears often when the inferential problem is challenging and requires a large amount of simulated samples. This can happen because the adaptive schemes may require many iterations, because of the high-dimensional nature of the tackled problem, and/or because a high precision (low variance) is required in the estimate. In [75], several compressing schemes are proposed and theoretically analyzed for assigned importance weights to groups of samples for distributed or decentralized Bayesian inference. The framework is extended in [68], where a stronger theoretical support is given, and new deterministic and random rules for compression are given. The approach in [56, 6] considers the case of a single node that keeps simulating samples and assigning them an importance weight. The bottleneck here is the storage of the samples so one needs to decide at each time if the sample is stored or discarded. A compression algorithm is introduced for building a dictionary based on greedy subspace projections and a kernel density estimator of the targeted distribution with a limited number of samples. It is shown that asymptotic bias of this method is a tunable constant depending on the kernel bandwidth parameter and a compression parameter. Finally, some works have studied the combination of IS estimators in the distributed setting. For instance, in [19, Section 4], independent estimators are linearly combined with the combination weights being the inverse of the variance of each estimator. A similar approach is followed in [82], using the $\widehat{\text{ESS}}$ instead of the variance of the estimator (which is unknown in most practical problems). A Bayesian combination of Monte Carlo estimators is considered in [65, 66]. Note that the MIS approach is, due to its own nature, an implicit linear combination of multiple estimators (each of them using samples from one or several proposals). This perspective is exploited for instance in [46, 97].

4 Adaptive importance sampling (AIS)

Since choosing a good proposal (or set of proposals) in advance is in general impossible, a common approach is the use of adaptive importance sampling (AIS) [10]. AIS algorithms are iterative methods for a gradual learning of one or multiple proposals that aim at approximating the target pdf. Algorithm 1 describes a generic AIS algorithm through three basic steps: the simulation of samples from a one or several proposals (sampling), the computation of the importance weight of each sample (weighting), and the update of the parameters that characterize the proposal(s) for repeating the previous steps in the next iteration (adaptation).

Most existing algorithms can be described in this framework that we describe with more detail. The generic AIS algorithm initializes N proposals $\{q_n(\mathbf{x}|\boldsymbol{\theta}_{n,1})\}_{n=1}^N$, parametrized each of them by a vector $\boldsymbol{\theta}_{n,1}$. Then, K samples are simulated from each proposals, $\mathbf{x}_{n,1}^{(k)}, \quad n = 1, \ldots, N, k = 1, \ldots, K$, and weighted properly. Here again many ways of sampling and weighting are possible, as it is described in Section 3.1. At the end of the weighting step, it is possible to approximate the integral of Eq. (1) with either UIS and SNIS, and the target distribution with a discrete random measure, by using the set of weighted samples $\{\mathbf{x}_{n,1}^{(k)}, w_{n,1}^{(k)}\}, \quad n = 1, \ldots, N, k = 1, \ldots, K$. Finally, the parameters of the n-th proposals are updated from $\boldsymbol{\theta}_{n,1}$ to $\boldsymbol{\theta}_{n,2}$. This three-step process is repeated until an iteration stoppage criterion is met (e.g., a maximum number of iterations, J, is reached). Note that at the end, the estimators can either use all weighted samples from iterations 1 to J, or only the samples from the last iteration.

Algorithm 1 Generic AIS algorithm

1: **Input:** Choose K, N, J, and $\{\boldsymbol{\theta}_{n,1}\}_{n=1}^N$
2: **for** $i = 1, \ldots, T$ **do**
3:     **Sampling:** Draw K samples from each of the N proposal pdfs, $\{q_{n,j}(\mathbf{x}|\boldsymbol{\theta}_{n,j})\}_{n=1}^N$, $\mathbf{x}_{n,j}^{(k)}, k = 1, \ldots, K, \ n = 1, \ldots, N$
4:     **Weighting:** Calculate the weights, $w_{n,j}^{(k)}$, for each of the generated KN samples.
5:     **Adaptation:** Update the proposal parameters $\{\boldsymbol{\theta}_{n,j}\}_{n=1}^N \longrightarrow \{\boldsymbol{\theta}_{n,j+1}\}_{n=1}^N$.
6: **end for**
7: **Output:** Return the KNJ pairs $\{\mathbf{x}_{n,j}^{(k)}, w_{n,j}^{(k)}\}$ for all $k = 1, \ldots, K, \ n = 1, \ldots, N, \ j = 1, \ldots, J$.

The literature is vast in AIS methods and a detailed description of all of them goes beyond the scope of this paper (see [10] for a thorough review). Most of the AIS algorithms can be classified within three categories, depending on how the proposals are adapted. Figure 1 shows graphically the three families of AIS algorithms, describing the dependencies for the adaptation of the proposal parameters and the simulation of the samples. Each subplot corresponds to each family, whose description and corresponding AIS algorithms of the literature are given below.

a) The proposal parameters are adapted using the last set of drawn samples (e.g., standard PMC [12], DM-PMC [33, 34], N-PMC [53], M-PMC [13], APIS [69]).

b) The proposal parameters are adapted using all drawn samples up to the latest iteration (e.g., AMIS [16], CAIS [23], Daisee [64], EAMIS [25], RS-CAIS [24]).

c) The proposal parameters are adapted using an independent process from the samples (LAIS [74, 73], GAPIS [30], GIS [99], IMIS [43], SL-PMC [26]).

In Table 2, we describe some relevant AIS algorithms according to different features: the number of proposals; the weighting scheme (*nonlinear* corresponds to the clipping strategies of Section 2.5.1, *standard* is equivalent to Option 1 in Section 3.1, *spatial mixture* corresponds to Option 2 with $\psi(\mathbf{x}) = \sum_{i=1}^N q_{i,j}(\mathbf{x}|\boldsymbol{\theta}_{i,j})$, *temporal mixture* corresponds to Option 2 with $\psi(\mathbf{x}) = \sum_{\tau=1}^j q_{n,\tau}(\mathbf{x}|\boldsymbol{\theta}_{n,\tau})$); and the parameters that are adapted (either location and scale, or only location). In Table 3, we describe the computational complexity of the same algorithms
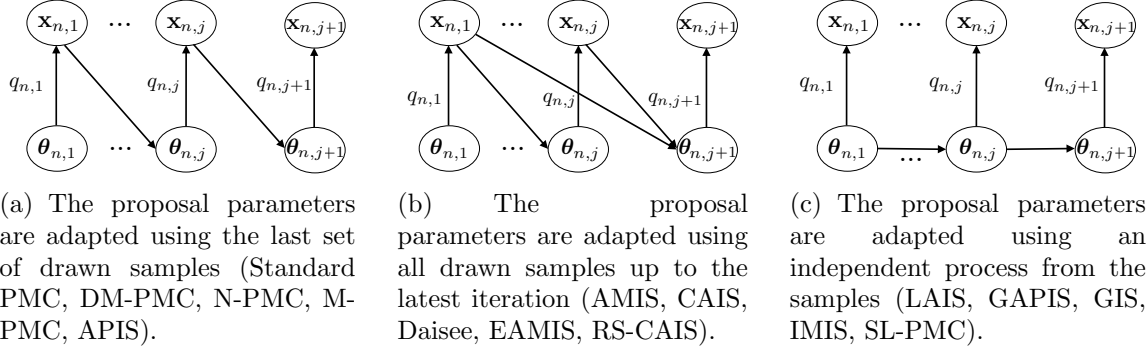
(a) The proposal parameters are adapted using the last set of drawn samples (Standard PMC, DM-PMC, N-PMC, M-PMC, APIS).

(b) The proposal parameters are adapted using all drawn samples up to the latest iteration (AMIS, CAIS, Daisee, EAMIS, RS-CAIS).

(c) The proposal parameters are adapted using an independent process from the samples (LAIS, GAPIS, GIS, IMIS, SL-PMC).

Figure 1: Graphical description of three possible dependencies between the adaptation of the proposal parameters $\boldsymbol{\theta}_{n,t}$ and the samples. Note that $q_{n,t} \equiv q_{n,t}(\mathbf{x}|\boldsymbol{\theta}_{n,t})$.

according to number of target evaluations, proposal evaluations, target evaluations per proposal, and proposal evaluations per proposal. In some AIS algorithms, the proposals converge with the number of iterations J, although proving this convergence (and the associated convergence rates) is in general a tough problem (see a recent result in [3]). For many other AIS algorithms (e.g., DM-PMC, LAIS, APIS), the proposals do not converge to any limiting distribution. Converge rates have been established only for simple classes of AIS algorithms which are based on optimized parametric proposals [3]. Note that AIS-based algorithms have been also used for optimization purposes [81, 4].

Table 2: Comparison of various AIS algorithms according to different features.

Algorithm	# proposals	Weighting	Adaptation strategy	Parameters adapted
Standard PMC	$N > 1$	standard	resampling	location
M-PMC	$N > 1$	spatial mixture	resampling	location
N-PMC	either	nonlinear	moment estimation	location/scale
LAIS	$N > 1$	generic mixture	MCMC	location
DM-PMC	$N > 1$	spatial mixture	resampling	location
AMIS	$N = 1$	temporal mixture	moment estimation	location/scale
GAPIS	$N > 1$	spatial mixture	gradient process	location/scale
APIS	$N > 1$	spatial mixture	moment estimation	location

Acknowledgements

Table 3: Comparison of various AIS algorithms according to the computational complexity.

Algorithm	# target eval	# proposal eval	# target eval/sample	# proposal eval/sample
Standard PMC	NJ	NJ	1	1
N-PMC	NJ	NJ	1	1
M-PMC	KJ	KNJ	1	N
LAIS	$K(N+1)J$	KN^2J	$1 + 1/N$	N
DM-PMC	KNJ	KN^2J	1	N
AMIS	KJ	KJ^2	1	J
GAPIS	KNJ	KN^2J	1	N
APIS	KNJ	KN^2J	1	N

References

[1] F. S. Acton. *Numerical Methods That Work*. The Mathematical Association of America, Washington, DC, 1990.

[2] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, A. Stuart, et al. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017.

[3] Ö. D. Akyildiz and J. Míguez. Convergence rates for optimised adaptive importance samplers. *arXiv preprint arXiv:1903.12044*, 2019.

[4] O. D. Akyildiz, I. P. Marino, and J. Míguez. Adaptive noisy importance sampling for stochastic optimization. In *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–5. IEEE, 2017.

[5] M. C. Ausín. Quadrature and numerical integration. *Wiley StatsRef: Statistics Reference Online (stat03875)*, pages 1–10, 2014.

[6] A. S. Bedi, A. Koppel, V. Elvira, and B. M. Sadler. Compressed streaming importance sampling for efficient representations of localization distributions. In *2019 Asilomar Conference on Signals, Systems and Computers*, pages 1–5. IEEE, 2019.

[7] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley & sons, 1994.

[8] A. Bhadra and E. L. Ionides. Adaptive particle allocation in iterated sequential Monte Carlo via approximating meta-models. *Statistics and Computing*, 26(1-2):393–407, 2016.

[9] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley & sons, 1973.

[10] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Míguez, and P. M. Djuric. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Process. Mag.*, 34 (4):60–79, 2017.

[11] R. L. Burden and J. D. Faires. *Numerical Analysis*. Brooks Cole, 2000.

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

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

[14] S. Chatterjee, P. Diaconis, et al. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

[15] J. A. Christen. Gibbs sampling. *Wiley StatsRef: Statistics Reference Online*, pages 1–9, 2014.

[16] 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.

[17] P. M. Djuric, T. Lu, and M. F. Bugallo. Multiple particle filtering. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 3, pages III–1181. IEEE, 2007.

[18] R. Douc, O. Cappé, and E. Moulines. Comparison of resampling schemes for particle filtering. In *Proc. 4th Int. Symp. on Image and Signal Processing and Analysis*, pages 64–69, September 2005.

[19] R. Douc, A. Guillin, J. M. Marin, and C. P. Robert. Minimum variance importance sampling via population Monte Carlo. *ESAIM: Probability and Statistics*, 11:427–447, 2007.

[20] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

[21] A. Doucet, N. De Freitas, K. Murphy, and S. Russell. Rao-blackwellised particle filtering for dynamic Bayesian networks. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 176–183. Morgan Kaufmann Publishers Inc., 2000.

[22] W. L. Dunn and J. K. Shultis. *Exploring Monte Carlo Methods*. Elsevier Science, Amsterdam (The Netherlands), 2011.

[23] Y. El-Laham, V. Elvira, and M. F. Bugallo. Robust covariance adaptation in adaptive importance sampling. *IEEE Signal Processing Letters*, 25(7):1049–1053, 2018.

[24] Y. El-Laham, V. Elvira, and M. F. Bugallo. Recursive shrinkage covariance learning in adaptive importance sampling. In *Proc. IEEE Int. Work. Comput. Adv. Multi-Sensor Adap. Process. (CAMSAP 2019)*, pages 1–5, 2019.

[25] Y. El-Laham, L. Martino, V. Elvira, and M. F. Bugallo. Efficient adaptive multiple importance sampling. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

[26] V. Elvira and É. Chouzenoux. Langevin-based strategy for efficient proposal adaptation in population Monte Carlo. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5077–5081. IEEE, 2019.

[27] V. Elvira and I. Santamaría. Efficient ser estimation for mimo detectors via importance sampling schemes. In *2019 Asilomar Conference on Signals, Systems and Computers*, pages 1–5. IEEE, 2019.

[28] V. Elvira and I. Santamaría. Multiple importance sampling for efficient symbol error rate estimation. *IEEE Signal Processing Letters*, 26(3):420–424, 2019.

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

[30] V. Elvira, L. Martino, L. Luengo, and J. Corander. A gradient adaptive population importance sampler. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2015)*, pages 4075–4079, Brisbane, Australia, 19-24 April 2015.

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

[32] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Multiple importance sampling with overlapping sets of proposals. *IEEE Workshop on Statistical Signal Processing (SSP)*, 2016.

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

[34] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Population Monte Carlo schemes with reduced path degeneracy. In *Proc. IEEE Int. Work. Comput. Adv. Multi-Sensor Adap. Process. (CAMSAP 2017)*, pages 1–5, 2017.

[35] V. Elvira, J. Míguez, and P. Djurić. Adapting the number of particles in sequential monte carlo methods through an online scheme for convergence assessment. *IEEE Transactions on Signal Processing*, 65(7):1781–1794, 2017.

[36] V. Elvira, L. Martino, M. F. Bugallo, and P. M. Djurić. In search for improved auxiliary particle filters. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1637–1641. IEEE, 2018.

[37] V. Elvira, L. Martino, and C. P. Robert. Rethinking the effective sample size. *arXiv preprint arXiv:1809.04129*, 2018.

[38] V. Elvira, P. Closas, and L. Martino. Gauss-Hermite quadrature for non-gaussian inference via an importance sampling interpretation. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

[39] V. Elvira, L. Martino, M. F. Bugallo, and P. M. Djuric. Elucidating the auxiliary particle filter via multiple importance sampling [lecture notes]. *IEEE Signal Processing Magazine*, 36(6):145–152, 2019.

[40] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155, 2019.

[41] V. Elvira, J. Míguez, and P. M. Djurić. New results on particle filters with adaptive number of particles. *arXiv preprint arXiv:1911.01383*, 2019.

[42] V. Elvira, L. Martino, and P. Closas. Importance Gaussian quadrature. *arXiv preprint arXiv:2001.03090*, 2020.

[43] M. Fasiolo, F. E. de Melo, and S. Maskell. Langevin incremental mixture importance sampling. *Stat. Comput.*, 28(3):549–561, 2018.

[44] J. E. Gentle. *Random Number Generation and Monte Carlo Methods*. Springer, 2004.

[45] 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.

[46] V. Havran and M. Sbert. Optimal combination of techniques in multiple importance sampling. In *Proceedings of the 13th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, pages 141–150, 2014.

[47] H. Y. He and A. B. Owen. Optimal mixture weights in multiple importance sampling. *arXiv preprint arXiv:1411.3954*, 2014.

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

[49] J. H. Huggins, D. M. Roy, et al. Sequential Monte Carlo as approximate sampling: bounds, adaptive resampling via $infty$-ess, and an application to particle gibbs. *Bernoulli*, 25(1): 584–622, 2019.

[50] E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

[51] P. Jaeckel. *Monte Carlo Methods in Finance*. Wiley, 2002.

[52] H. Kahn. Random sampling (Monte Carlo) techniques in neutron attenuation problems. *Nucleonics*, 6(5):27–passim, 1950.

[53] E. Koblents and J. Míguez. A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models. *Statistics and Computing*, 25(2):407–425, 2015.

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

[55] A. Kong. Importance sampling. *Wiley StatsRef: Statistics Reference Online*, 2014.

[56] A. Koppel, A. S. Bedi, V. Elvira, and B. M. Sadler. Approximate shannon sampling in importance sampling: Nearly consistent finite particle estimates. *arXiv preprint arXiv:1909.10279*, 2019.

[57] J. Kotecha and P. M. Djurić. Gaussian particle filtering. *IEEE Transactions Signal Processing*, 51(10):2592–2601, October 2003.

[58] D. Kroese, T. Taimre, and Z. Botev. *Handbook of Monte Carlo Methods*. Wiley Series in Probability and Statistics, John Wiley and Sons, New York, 2011.

[59] P. K. Kythe and M. R. Schaferkotter. *Handbook of Computational Methods for Integration*. Chapman and Hall/CRC, 2004.

[60] A. Lee and N. Whiteley. Variance estimation and allocation in the particle filter. *arXiv:1509.00394v1 [stat.CO]*, 2015.

[61] P. M. Lee. Bayesian inference. *Wiley StatsRef: Statistics Reference Online (stat00207.pub2)*, pages 1–9, 2014.

[62] T. Li, M. Bolic, and P. M. Djuric. Resampling methods for particle filtering: Classification, implementation, and strategies. *IEEE Signal Processing Magazine*, 32(3):70–86, 2015.

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

[64] X. Lu, T. Rainforth, Y. Zhou, J.-W. van de Meent, and Y. W. Teh. On exploration, exploitation and learning in adaptive importance sampling. *arXiv preprint arXiv:1810.13296*, 2018.

[65] D. Luengo, L. Martino, V. Elvira, and M. Bugallo. Bias correction for distributed bayesian estimators. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 253–256. IEEE, 2015.

[66] D. Luengo, L. Martino, V. Elvira, and M. Bugallo. Efficient linear fusion of partial estimators. *Digital Signal Processing*, 78:265–283, 2018.

[67] L. Martino and V. Elvira. Metropolis sampling. *Wiley StatsRef: Statistics Reference Online*, pages 1–18, 2017.

[68] L. Martino and V. Elvira. Compressed Monte Carlo for distributed Bayesian inference. *viXra:1811.0505*, 2018.

[69] 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, 2015.

[70] L. Martino, V. Elvira, and F. Louzada. Alternative effective sample size measures for importance sampling. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–5. IEEE, 2016.

[71] L. Martino, V. Elvira, and F. Louzada. Weighting a resampled particle in sequential Monte Carlo. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–5. IEEE, 2016.

[72] L. Martino, V. Elvira, and F. Louzada. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386–401, 2017.

[73] L. Martino, V. Elvira, and D. Luengo. Anti-tempered layered adaptive importance sampling. In *2017 22nd International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2017.

[74] L. Martino, V. Elvira, D. Luengo, and J. Corander. Layered adaptive importance sampling. *Statistics and Computing*, 27(3):599–623, 2017.

[75] L. Martino, V. Elvira, and G. Camps-Valls. Group importance sampling for particle filtering and mcmc. *Digital Signal Processing*, 82:133–151, 2018.

[76] L. Martino, V. Elvira, J. Míguez, A. Artés-Rodríguez, and P. Djurić. A comparison of clipping strategies for importance sampling. In *2018 IEEE Statistical Signal Processing Workshop (SSP)*, pages 558–562. IEEE, 2018.

[77] F. J. Medina-Aguayo and R. G. Everitt. Revisiting the balance heuristic for estimating normalising constants. *arXiv preprint arXiv:1908.06514*, 2019.

[78] E. Medova. Bayesian Analysis and Markov Chain Monte Carlo simulation. *Wiley StatsRef: Statistics Reference Online (stat03616)*, pages 1–12, 2015.

[79] J. Míguez. On the performance of nonlinear importance samplers and population Monte Carlo schemes. In *2017 22nd International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2017.

[80] J. Miguez, I. P. Mariño, and M. A. Vázquez. Analysis of a nonlinear importance sampling scheme for Bayesian parameter estimation in state-space models. *Signal Processing*, 142: 281–291, 2018.

[81] P. D. Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(3):411–436, 2006.

[82] T. L. T. Nguyen, F. Septier, G. W. Peters, and Y. Delignon. Improving smc sampler estimate by recycling all past simulated particles. In *Statistical Signal Processing (SSP), 2014 IEEE Workshop on*, pages 117–120. IEEE, 2014.

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

[84] A. B. Owen. *Monte Carlo theory, methods and examples.* 2013.

[85] A. B. Owen, Y. Maximov, M. Chertkov, et al. Importance sampling the union of rare events with an application to power systems analysis. *Electronic Journal of Statistics*, 13 (1):231–254, 2019.

[86] M. K. Pitt and N. Shephard. Auxiliary variable based particle filters. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, chapter 13, pages 273–293. Springer, 2001.

[87] B. F. Plybon. *An Introduction to Applied Numerical Analysis.* PWS-Kent, Boston, MA, 1992.

[88] C. P. Robert. *The Bayesian Choice.* Springer, 2007.

[89] C. P. Robert. Monte Carlo methods. *Wiley StatsRef: Statistics Reference Online*, pages 1–13, 2014.

[90] C. P. Robert. The metropolishastings algorithm. *Wiley StatsRef: Statistics Reference Online*, pages 1–13, 2015.

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

[92] E. K. Ryu and S. P. Boyd. Adaptive importance sampling via stochastic convex programming. *arXiv preprint arXiv:1412.4845*, 2014.

[93] M. Sbert and V. Elvira. Generalizing the balance heuristic estimator in multiple importance sampling. *arXiv preprint arXiv:1903.11908*, 2019.

[94] M. Sbert and V. Havran. Adaptive multiple importance sampling for general functions. *The Visual Computer*, 33(6-8):845–855, 2017.

[95] M. Sbert, V. Havran, and L. Szirmay-Kalos. Variance analysis of multi-sample and one-sample multiple importance sampling. In *Computer Graphics Forum*, volume 35, pages 451–460. Wiley Online Library, 2016.

[96] M. Sbert, V. Havran, and L. Szirmay-Kalos. Multiple importance sampling revisited: breaking the bounds. *EURASIP Journal on Advances in Signal Processing*, 2018(1):15, 2018.

[97] M. Sbert, V. Havran, L. Szirmay-Kalos, and V. Elvira. Multiple importance sampling characterization by weighted mean invariance. *The Visual Computer*, 34(6-8):843–852, 2018.

[98] M. Sbert, V. Havran, and L. Szirmay-Kalos. Optimal deterministic mixture sampling. In *Eurographics (Short Papers)*, pages 73–76, 2019.

[99] I. Schuster. Gradient importance sampling. Technical report, 2015. https://arxiv.org/abs/1507.05781.

[100] T. Taimre, D. P. Kroese, and Z. I. Botev. Monte Carlo methods. *Wiley StatsRef: Statistics Reference Online DOI*, 10:9781118445112, 2019.

[101] S. T. Tokdar and R. E. Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

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

[103] A. Vehtari, A. Gelman, and J. Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

[104] S. Wang. Importance sampling including the bootstrap. *Wiley StatsRef: Statistics Reference Online*, 2014.