Adaptive Population Importance Samplers: A General Perspective

Luca Martino , Víctor Elvira , David Luengo , Francisco Louzada

Institute of Mathematical Sciences and Computing, Universidade de São Paulo, Brazil.
Dep. of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain.
Dep. of Circuits and Systems Engineering, Universidad Politécnica de Madrid, Spain.

Abstract—Importance sampling (IS) is a well-known Monte Carlo method, widely used to approximate a distribution of interest using a random measure composed of a set of weighted samples generated from another proposal density. Since the performance of the algorithm depends on the mismatch between the target and the proposal densities, a set of proposals is often iteratively adapted in order to reduce the variance of the resulting estimator. In this paper, we review several well-known adaptive population importance samplers, providing a unified common framework and classifying them according to the nature of their estimation and adaptive procedures. Furthermore, we interpret the underlying motivation for the different adaptation schemes, opening the door for novel and more efficient algorithms. Finally, we compare the performance of different algorithms available in the literature through a toy example.

I. INTRODUCTION

Monte Carlo methods are widely used in signal processing, communications, machine learning and many other scientific disciplines [1], [2], [3], [4]. Importance sampling (IS) [5], [6] is a well-known Monte Carlo (MC) methodology that can be used to compute integrals involving a complicated multidimensional target probability density function (pdf), $\pi(\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^n$. The method proceeds by drawing samples from a simple proposal pdf, $q(\mathbf{x})$, and assigning weights to them according to the ratio between the target and the proposal, i.e., $w(\mathbf{x}) = \frac{\pi(\mathbf{x})}{q(\mathbf{x})}$. However, although the validity of this approach is guaranteed under mild assumptions, the variance of the estimator depends critically on the discrepancy between the proposal and the target densities.

In order to solve this issue, several works are devoted to the design of adaptive IS (AIS) schemes [7], where the proposal density is updated by learning from previously generated samples [6]. Some well-known methods available in the literature are Population Monte Carlo (PMC) [8], [9], [10], Adaptive Multiple Importance Sampling (AMIS) [11], Adaptive Population Importance Sampling (APIS) [12], [13], or Layered Adaptive Importance Sampling (LAIS) [14]. All of them are iterated importance samplers, and most of them

employ the multiple importance sampling (MIS) approach [15] for building the IS estimators. In all of these schemes, the proposal or the population of proposals is updated at each iteration, using the statistical information that has been collected from the target in the previous iterations through the samples generated. However, the aforementioned methods differ substantially, both in the specific MIS scheme used to perform the estimation (different weighting functions can be employed within a MIS framework, as shown in [15]) and in the adaptation procedure used to update the proposals.

In this work, we describe a general framework for adaptive population importance samplers. Firstly, we introduce a generic AIS scheme that encompasses several state of the art methods currently available in the literature. For the sake of simplicity, we focus on the adaptation of the mean vectors of the proposal pdfs, although some variants of these techniques also consider the adaptation of the covariance matrices. Then, each step of this general AIS scheme is analyzed and discussed in detail. Furthermore, several examples are provided and the numerical performance of the different alternatives is compared through a toy example. The theoretical reinterpretation of these adaptive IS procedures and the practical considerations provided throughout the paper can be very useful for the better understanding of the existing methods, as well as for designing novel and more efficient techniques.

II. STATIC IMPORTANCE SAMPLING

Let us consider the variable of interest, $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n$, and the target density

$$\bar{\pi}(\mathbf{x}) = \frac{1}{Z}\pi(\mathbf{x}), \tag{1}$$

that usually corresponds to a posterior pdf. Our goal is computing efficiently some moment of $\mathbf{x}$, i.e.,

$$I = E_\pi[f(\mathbf{X})] = \int_{\mathcal{X}} f(\mathbf{x})\bar{\pi}(\mathbf{x})d\mathbf{x} = \frac{1}{Z}\int_{\mathcal{X}} f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}, \tag{2}$$

where $Z = \int_{\mathcal{X}} \pi(\mathbf{x})d\mathbf{x}$ and $f(\mathbf{x})$ can be any square-integrable function. In general, we are not able to draw samples from $\bar{\pi}(\mathbf{x})$. The importance sampling technique is based on the following equality

$$I = E_q[w(\mathbf{X})f(\mathbf{X})] = \frac{1}{Z}\int_{\mathcal{X}} f(\mathbf{x})w(\mathbf{x})q(\mathbf{x})d\mathbf{x}, \tag{3}$$

where q is a proposal pdf (simpler to draw from and with heavier tails than $\bar{\pi}$), and

$$w(\mathbf{x}) = \frac{\pi(\mathbf{x})}{q(\mathbf{x})}. \tag{4}$$

Hence, drawing M samples $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}$ from $q(\mathbf{x})$, the standard IS estimator when Z is known is given by

$$\widehat{I} = \frac{1}{M} \frac{1}{Z} \sum_{m=1}^{M} w^{(m)} f(\mathbf{x}^{(m)}), \tag{5}$$

where $w^{(m)} = w(\mathbf{x}^{(m)}) = \frac{\pi(\mathbf{x}^{(m)})}{q(\mathbf{x}^{(m)})}$. Alternatively, when Z is unknown the standard *self-normalized* IS estimator is

$$\widetilde{I} = \sum_{m=1}^{M} \bar{w}^{(m)} f(\mathbf{x}^{(m)}), \tag{6}$$

where $\bar{w}^{(m)} = \frac{w^{(m)}}{\sum_{\ell=1}^{M} w^{(\ell)}}$. Given a fixed function $f(\mathbf{x})$, the optimal proposal q that minimizes the variance of the IS estimator is $q(\mathbf{x}) \propto |f(\mathbf{x})|\bar{\pi}(\mathbf{x})$ (at least, when $\widehat{I}$ is applied). However, in practical applications we are often interested in computing expectations w.r.t. several f's. In this context, a more appropriate strategy is minimizing the variance of the importance weights $w^{(m)}$ for $m = 1, \ldots, M$. In this case, the minimum variance is attained when $q(\mathbf{x}) = \bar{\pi}(\mathbf{x})$ [16]. Thus, we are interested in choosing the proposal pdf q as close as possible to $\bar{\pi}$. In the rest of this work, we consider the joint use of N different proposal pdfs, $q_1, \ldots, q_N$ [15]. In this case, drawing $\{\mathbf{x}_n^{(m)}\}_{m=1}^{M} \sim q_n(\mathbf{x})$ for all $n = 1, \ldots, N$, several choices of weights

$$w^{(m)} = \frac{\pi(\mathbf{x}_n^{(m)})}{\Phi_n(\mathbf{x}_n^{(m)})}, \tag{7}$$

are valid. More specifically, several choices of the denominator $\Phi_n(\mathbf{x})$ that lead to an unbiased estimator $\widehat{I}$ in Eq. (5) are possible, as proved in [15].

III. GENERIC SCHEME FOR ADAPTIVE POPULATION IMPORTANCE SAMPLERS

Table I shows a generic population importance sampler that encompasses most of the state of the art approaches currently available in the literature. A generic adaptive importance sampler is a sequential algorithm where a set of N proposals are adapted over time. At the t-th iteration, the set of available proposals is $\{q_{1,t}, \ldots, q_{N,t}\}$. These proposals can change over time, not only updating their parameters but even completely changing their shape (i.e., the family of the distribution), and we only assume that all the $q_{n,t}$'s have heavier tails than the target $\bar{\pi}$ [6]. For the sake of simplicity, we only consider location-scale densities such that each $q_{n,t}$ is completely characterized by a mean vector, $\boldsymbol{\mu}_{n,t}$, and a covariance matrix, $\mathbf{C}_n$. In this work, we focus on the adaptation of the mean vectors, i.e., the $\boldsymbol{\mu}_{n,t}$'s.

At the t-th iteration, M independent samples are drawn from each proposal in Step 1 of Table I, i.e.,

$$\mathbf{x}_{n,t}^{(m)} \sim q_{n,t}(\mathbf{x}|\boldsymbol{\mu}_{n,t}, \mathbf{C}_n), \tag{8}$$

Table I
GENERIC ADAPTIVE POPULATION IMPORTANCE SAMPLER.

Initialization: Choose T, N, M, $q_{n,0}$, $\boldsymbol{\mu}_{n,0}$ and $\mathbf{C}_n$ for $n = 1, \ldots, N$.
For $t = 1, \ldots, T$:   1) **Sampling:** Draw M samples $\mathbf{x}_{n,t}^{(1)}, \ldots, \mathbf{x}_{n,t}^{(M)}$ from each of the N proposal pdfs in the population $\{q_1, \ldots, q_N\}$, i.e., $\qquad \mathbf{x}_{n,t}^{(m)} \sim q_{n,t}(\mathbf{x}|\boldsymbol{\mu}_{n,t}, \mathbf{C}_n),$ for $m = 1, \ldots, M$.   2) **Weighting:** Weight the samples, $\{\mathbf{x}_{n,t}^{(m)}\}_{m=1}^{M}$, with $\qquad w_{n,t}^{(m)} = \dfrac{\pi(\mathbf{x}_{n,t}^{(m)})}{\Phi_{n,t}(\mathbf{x}_{n,t}^{(m)})}, \qquad (10)$ according to a proper function $\Phi_{n,t}$ that leads to unbiased IS estimators $\widehat{I}$ [5, Section 2.5.4][14], [15].   3) **Adaptation of the means:** Apply some suitable procedure to update the mean vectors, $\qquad \{\boldsymbol{\mu}_{n,t}\}_{n=1}^{N} \longrightarrow \{\boldsymbol{\mu}_{n,t+1}\}_{n=1}^{N},$ without jeopardizing the consistency of the IS estimators.   **Outputs:** Return $\{\mathbf{x}_{n,t}^{(m)}, w_{n,t}^{(m)}\}$ for all $m = 1, \ldots, M$ and $t = 1, \ldots, T$.

Table II
EXAMPLES OF $\Phi_{n,t}$ AND $\Omega_{n,t}$ IN LITERATURE.

Weight den.	PMC [8]	AMIS ($N = 1$) [11]	APIS [13]
$\Phi_{n,t}(\mathbf{x})$	$q_{n,t}(\mathbf{x})$	$\frac{1}{t}\sum_{\tau=1}^{t} q_\tau(\mathbf{x})$	$\frac{1}{N}\sum_{n=1}^{N} q_{n,t}(\mathbf{x})$
$\Omega_{n,t}(\mathbf{x})$	$q_{n,t}(\mathbf{x})$	$\frac{1}{t}\sum_{\tau=1}^{t} q_\tau(\mathbf{x})$	$q_{n,t}(\mathbf{x})$

for $m = 1, \ldots, M$ and $n = 1, \ldots, N$. In Step 2, a weight of the form

$$w_{n,t}^{(m)} = \frac{\pi(\mathbf{x}_{n,t}^{(m)})}{\Phi_{n,t}(\mathbf{x}_{n,t}^{(m)})}, \tag{9}$$

is associated to these MN samples. In the denominator, we have a function $\Phi_{n,t}$, properly chosen in such a way that the estimator of Eq. (5) is unbiased [5, Section 2.5.4], [14], [15]. Table II shows some examples of the $\Phi_{n,t}$'s used in literature, denoting $q_{n,t}(\mathbf{x}) \equiv q_{n,t}(\mathbf{x}|\boldsymbol{\mu}_{n,t}, \mathbf{C}_n)$ to ease the notation. Note that $q_{n,t}(\mathbf{x}) \equiv q_t(\mathbf{x}|\boldsymbol{\mu}_t, \mathbf{C})$ for AMIS, since $N = 1$ in this case. Finally, Step 3 performs the adaptation of the mean vectors $\boldsymbol{\mu}_{n,t}$. Several adaptive schemes have been proposed in the literature, and here we divide them into two main families.

1) The adaptation procedure employs previous samples, weighted using either the same or other denominator w.r.t. the weights used for the estimation. This approach is summarized in Table III, and includes the possible application of resampling steps over a subset of these weighted samples [8], [9], [10].

2) More sophisticated procedures are employed. For instance, information about the gradient of $\bar{\pi}$ can be used [17], [18], or MCMC techniques applied to adapt the mean vectors [14], [19], [20], [21], [22]. These approaches tend to provide better performance, but they are more computationally demanding in general (see Section V). Table IV summarizes them.
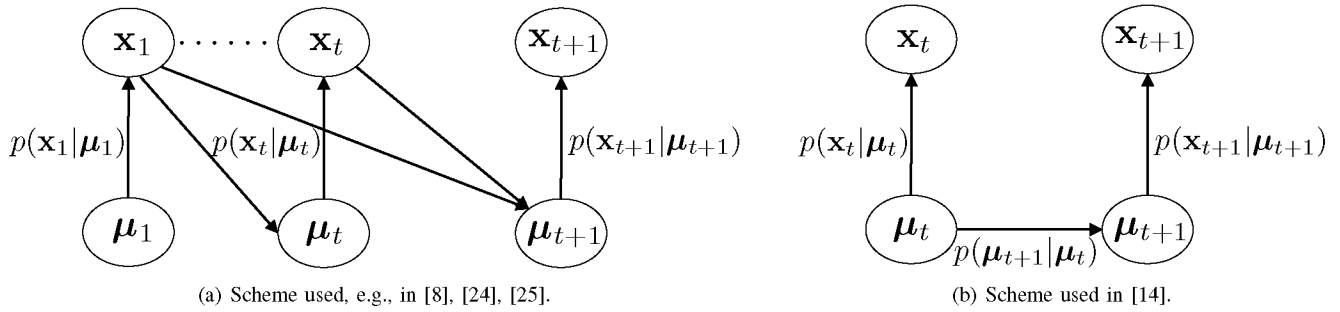
(a) Scheme used, e.g., in [8], [24], [25].

(b) Scheme used in [14].

Figure 1. Graphical representation of two possible generation procedures of the pair $(\boldsymbol{\mu}_{t+1}, \mathbf{x}_{t+1})$ given $(\boldsymbol{\mu}_t, \mathbf{x}_{1:t})$. Note that, for the sake of simplicity, we have set $N = 1$. The PMC method corresponds to a simplification of (a), where $p(\boldsymbol{\mu}_{1:N}|\mathbf{x}_{1:N,1:t}) = p(\boldsymbol{\mu}_{1:N}|\mathbf{x}_{1:N,t})$.

Table III
ADAPTATION BASED ON WEIGHTED SAMPLES.

Let us consider the set of IS weights

$$\rho_{n,t}^{(m)} = \frac{\pi(\mathbf{x}_{n,t}^{(m)})}{\Omega_{n,t}(\mathbf{x}_{n,t}^{(m)})},$$

where $\Omega_{n,t}$ are chosen in such a way that they do not jeopardize the consistency of the IS estimators, and they can be equal to $w_{n,t}^{(m)}$ in Eq. (10) or not (see Table II). Two different procedures are used in literature:

P1 Apply some resampling strategy to $\{\boldsymbol{\mu}_{n,t}\}_{n=1}^N$, with probabilities according to the weights $\rho_{n,t}^{(m)}$, to obtain $\{\boldsymbol{\mu}_{n,t+1}\}_{n=1}^N$ [8], [10], [23]. Nonlinear transformations of $\rho_{n,t}^{(m)}$ can be also applied [9].

P2 Building estimators of some moments of π employing $\rho_{n,t}^{(m)}$, and using this information to obtain $\{\boldsymbol{\mu}_{n,t+1}\}_{n=1}^N$ [24], [11], [25].

Table IV
MORE SOPHISTICATED ADAPTATION PROCEDURES.

P3 Adaptation by using MCMC transitions to obtain $\{\boldsymbol{\mu}_{n,t+1}\}_{n=1}^N$ given $\{\boldsymbol{\mu}_{n,t}\}_{n=1}^N$, as in [14], [23], [20], [21], [22].

P4 Adaptation by using stochastic gradient search of π for moving $\{\boldsymbol{\mu}_{n,t}\}_{n=1}^N$ to $\{\boldsymbol{\mu}_{n,t+1}\}_{n=1}^N$ [17], [18].

IV. THE DYNAMICS IN AIS

Henceforth, we consider $M = 1$ for the sake of simplicity. The consistency of the AIS estimators depends on a proper choice of $\Phi_{n,t}$ but also on the employed adaptation procedure. In many AIS methods [8], [24], [11], [25], the set of means at the $(t+1)$-th iteration, $\boldsymbol{\mu}_{1:N,t+1} = \{\boldsymbol{\mu}_{n,t+1}\}_{n=1}^N$, is chosen among the previous samples $\mathbf{x}_{1:N,1:t} = \{\mathbf{x}_{n,t}\}_{n,t=1}^{N,T}$, according to some probabilities. The next set of samples $\mathbf{x}_{1:N,t+1}$ are then drawn from $q_{n,t+1}(\mathbf{x}|\boldsymbol{\mu}_{n,t+1}, \mathbf{C}_n)$. Namely, the generating procedure of $(\boldsymbol{\mu}_{1:N,t+1}, \mathbf{x}_{1:N,t+1})$ can be expressed in terms of two conditional probabilities, i.e.,

$$\begin{cases} \boldsymbol{\mu}_{1:N,t+1} \sim p(\boldsymbol{\mu}_{1:N}|\mathbf{x}_{1:N,1:t}), \\ \mathbf{x}_{1:N,t+1} \sim p(\mathbf{x}_{1:N}|\boldsymbol{\mu}_{1:N,t+1}), \end{cases} \quad (11)$$

where $p(\boldsymbol{\mu}_{1:N}|\mathbf{x}_{1:N,1:t})$ depends on the specific adaptation procedure applied by the algorithm, and

$$p(\mathbf{x}_{1:N}|\boldsymbol{\mu}_{1:N,t+1}) = \prod_{n=1}^N q_{n,t+1}(\mathbf{x}|\boldsymbol{\mu}_{n,t+1}, \mathbf{C}_n), \quad (12)$$

is common for all the methods encompassed by the proposed framework. In some techniques, like PMC [8], [10], we have $p(\boldsymbol{\mu}_{1:N}|\mathbf{x}_{1:N,1:t}) = p(\boldsymbol{\mu}_{1:N}|\mathbf{x}_{1:N,t})$, i.e., the means a t the following iteration only depend on $\mathbf{x}_{1:N,t}$ at the previous iteration. Another interesting special case, often used in standard adaptation approaches [7], [24], [11], [25], occurs when $p(\boldsymbol{\mu}_{1:N}|\mathbf{x}_{1:N,1:t})$ is a singular density [26], i.e., $\boldsymbol{\mu}_{1:N,t+1}$ and $\mathbf{x}_{1:N,1:t}$ are linked by a deterministic function φ_t:

$$\boldsymbol{\mu}_{1:N,t+1} = \varphi_t(\mathbf{x}_{1:N,1:t}), \quad (13)$$

where φ_t is typically an estimator of $E_\pi[\mathbf{X}]$ with $\mathbf{X} \sim \pi(\mathbf{x})$ [11] or related values [24]. In other schemes, such as LAIS [14], the selection of $\boldsymbol{\mu}_{1:N,t+1}$ is performed independently from the generation of $\mathbf{x}_{1:N,t}$. In this case, the generating procedure for $(\boldsymbol{\mu}_{1:N,t+1}, \mathbf{x}_{1:N,t+1})$ becomes

$$\begin{cases} \boldsymbol{\mu}_{1:N,t+1} \sim p(\boldsymbol{\mu}_{1:N}|\boldsymbol{\mu}_{1:N,t}), \\ \mathbf{x}_{1:N,t+1} \sim p(\mathbf{x}_{1:N}|\boldsymbol{\mu}_{1:N,t+1}), \end{cases} \quad (14)$$

where $p(\mathbf{x}_{1:N}|\boldsymbol{\mu}_{1:N,t+1})$ is still given by Eq. (12). Figures 1(a)-(b) depict the graphical models corresponding to Eq. (11) and Eq. (14), respectively (with $N = 1$ for simplicity).

A. Consistency of the Estimators

On the one hand, note that the methods which apply the scheme in Figure 1(b) (i.e., those following Eq. (14)) can be converted to a static (i.e., non-adaptive) IS sampler by choosing all the means, $\{\boldsymbol{\mu}_{n,t}\}$ for $n = 1, \dots, N$ and $t = 1, \dots, T$, in advance, and then sampling all the x's. Therefore, if all the $q_{n,t}$'s have heavier tails than π and $\Phi_{n,t}$ is properly chosen, the consistency is always ensured using standard IS arguments. Namely, increasing the number of samples guarantees that the IS estimators converge to the integral value I.

On the other hand, for the methods which apply the scheme in Figure 1(a) (i.e., those following Eq. (11)) the consistency must be studied carefully on a case by case basis. Indeed, the adaptation could jeopardize the consistency in this case (yielding a systematic bias, for instance). However, the special case in Eq. (13) is simpler and consistency is guaranteed if $\boldsymbol{\mu}_{1:N,t}$ converges to a fixed vector as $t \to \infty$. In this case the adaptation is virtually stopped after a certain number of iterations, and the method then becomes asymptotically equivalent to a non adaptive IS algorithm. More generally,

important information about the dynamics of the algorithm is contained in the limiting distributions as $t \to \infty$ and $N \to \infty$, which are analyzed in the following section.

B. Limiting distributions

In the previous section, we have described two general approaches for designing adaptation procedures in AIS. In both cases, a joint densities is implicitly defined, i.e., $p(\boldsymbol{\mu}_{1:N,1:t}, \mathbf{x}_{1:N,1:t})$, for $t = 1, \ldots, T$. The marginal pdfs

$$h_{n,t}(\boldsymbol{\mu}) = p(\boldsymbol{\mu}_{n,t}), \quad \text{and} \quad g_{n,t}(\mathbf{x}) = p(\mathbf{x}_{n,t}), \quad (15)$$

are particularly interesting for studying the dynamics induced by the algorithm. In all the algorithms (that we are aware), $h_{n,t}(\boldsymbol{\mu})$ are $g_{n,t}(\mathbf{x})$ not dependent on n but to the total number of proposal pdfs N, i.e., we can write $h_t^{(N)}(\boldsymbol{\mu})$ and $g_t^{(N)}(\mathbf{x})$. The *limiting densities* (if exist)

$$h_\infty^{(N)}(\boldsymbol{\mu}) = \lim_{t \to \infty} h_t^{(N)}(\boldsymbol{\mu}), \quad h_t^{(\infty)}(\boldsymbol{\mu}) = \lim_{N \to \infty} h_t^{(N)}(\boldsymbol{\mu}),$$

$$g_\infty^{(N)}(\mathbf{x}) = \lim_{t \to \infty} g_t^{(N)}(\mathbf{x}), \quad g_t^{(\infty)}(\mathbf{x}) = \lim_{N \to \infty} g_t^{(N)}(\mathbf{x}),$$

are particularly important from a theoretical point of view. In LAIS [14], the density $h_\infty^{(N)}(\boldsymbol{\mu})$ is pre-established in advance by the user, like a *prior* pdf in a hierarchical Bayesian procedure. The information provided by $h_\infty^{(N)}(\boldsymbol{\mu})$, is crucial for ensuring the consistency of the IS estimators (controlling the dynamics of the AIS method). In LAIS [14], $h_\infty^{(N)}$ coincides to the target $\bar{\pi}$, i.e.,

$$h_\infty^{(N)}(\boldsymbol{\mu}) = \bar{\pi}(\boldsymbol{\mu}). \quad (16)$$

The pdf $g_\infty^{(N)}(\mathbf{x})$ can be considered as an *equivalent proposal density* [14]. Namely, after a certain number of iterations τ, i.e., for $t \le \tau$, one can consider

$$\mathbf{x}_{n,t} \sim g_\infty^{(N)}(\mathbf{x}), \quad t \ge \tau, \quad (17)$$

$\forall n = 1, \ldots, N$. For reducing the variance of the importance weights $g_\infty^{(N)}(\mathbf{x})$ has to be as close as possible to the target $\bar{\pi}(\mathbf{x})$. In PMC-type methods [8], [10] that apply resampling steps in the adaptation, we have

$$g_t^{(\infty)}(\mathbf{x}) = \bar{\pi}(\mathbf{x}), \quad (18)$$

and additionally $h_t^{(\infty)}(\boldsymbol{\mu}) = g_t^{(\infty)}(\boldsymbol{\mu})$. Finally, when the adaptation follows Eq. (13) [11], [25], we have $h_\infty^N(\boldsymbol{\mu}) = \delta(\boldsymbol{\mu} - E_\pi[\mathbf{X}])$, where $E_\pi[\mathbf{X}])$ is expected value of $\mathbf{X} \sim \bar{\pi}(\mathbf{x})$ and φ_t in Eq. (13) is an estimator of $E_\pi[\mathbf{X}]$. A similar strategies is given in [24]. Table V summarizes the limiting distribution $h_\infty^N(\boldsymbol{\mu})$ of different AIS schemes.

Table V
LIMITING DISTRIBUTION $h_\infty^N(\boldsymbol{\mu})$ WITH DIFFERENT AIS SCHEMES.

Stand. AIS approach [11]	PMC [8], APIS [13]	LAIS [14]
$\delta(\boldsymbol{\mu} - E_\pi[\mathbf{X}])$	unknown (if/when exists)	$\bar{\pi}(\boldsymbol{\mu})$

Table VI
MSE IN THE ESTIMATION OF THE MEAN OF THE TARGET. THE BEST RESULTS FOR EACH VALUE OF σ ARE HIGHLIGHTED IN BOLD-FACE.

Algorithm	$\sigma = 1$	$\sigma = 5$	$\sigma = 10$	$\sigma_{i,j} \sim \mathcal{U}([1,10])$
PMC [8]	107.58	0.6731	0.0744	0.0732
AMIS [11]	121.21	0.8640	**0.0121**	0.7328
APIS [13]	2.45	0.2424	0.0185	0.0045
LAIS [14]	**0.0021**	**0.0075**	0.0121	0.0049
GAPIS [17]	0.0023	0.0104	0.0143	**0.0040**

V. NUMERICAL RESULTS

Let us consider a bivariate multimodal target pdf formed by a mixture of 5 Gaussians, i.e.,

$$\bar{\pi}(\mathbf{x}) = \frac{1}{5} \sum_{i=1}^{5} \mathcal{N}(\mathbf{x}; \boldsymbol{\nu}_i, \boldsymbol{\Sigma}_i), \quad \mathbf{x} \in \mathbb{R}^2, \quad (19)$$

with means $\boldsymbol{\nu}_1 = [-10, -10]^\top$, $\boldsymbol{\nu}_2 = [0, 16]^\top$, $\boldsymbol{\nu}_3 = [13, 8]^\top$, $\boldsymbol{\nu}_4 = [-9, 7]^\top$, and $\boldsymbol{\nu}_5 = [14, -14]^\top$, and covariance matrices $\boldsymbol{\Sigma}_1 = [2, 0.6; 0.6, 1]$, $\boldsymbol{\Sigma}_2 = [2, -0.4; -0.4, 2]$, $\boldsymbol{\Sigma}_3 = [2, 0.8; 0.8, 2]$, $\boldsymbol{\Sigma}_4 = [3, 0; 0, 0.5]$, and $\boldsymbol{\Sigma}_5 = [2, -0.1; -0.1, 2]$. The goal is to estimate, via Monte Carlo, the mean of $\bar{\pi}$ (the true value is $[1.6, 1.4]^\top$), using different adaptive algorithms. For the sake of simplicity, we also consider Gaussian proposals.

We have tested APIS, AMIS, standard PMC, gradient APIS (GAPIS), and LAIS. At each run, the initial means of the proposals are selected uniformly within a square $\boldsymbol{\mu}_i^{(0)} \sim \mathcal{U}([-4, 4] \times [-4, 4])$. Initially, the same isotropic covariance matrix, $\mathbf{C}_i^{(0)} = \sigma^2 \mathbf{I}_2$, is used for every proposal. We test different values of $\sigma \in \{1, 5, 10\}$. Then, different non-isotropic diagonal covariance matrices, $\mathbf{C}_i^{(0)} = \text{diag}(\sigma_{i,1}^2, \sigma_{i,2}^2)$ with $\sigma_{i,j} \sim \mathcal{U}([1, 10])$ for $i = 1, \ldots N$ and $j \in \{1, 2\}$, are also tested. In this example, all the algorithms adapt only the means of the proposals, except for AMIS, that also adapts the covariance matrix. We use $N = 100$ for all the methods, except for AMIS (we recall that $N = 1$ in this case). In order to obtain a fair comparison, for each method T is selected in such a way that the total number of evaluations of the target is always fixed to $E = 2 \cdot 10^5$. Table VI shows the mean squared error (MSE) of the estimation of the first component of the mean (averaged over 500 runs). We have considered the best performance of each method after testing several combinations of parameters, as described in [13], [14], [17]. Note that GAPIS and LAIS provide the best results in general, whereas AMIS and APIS are also competitive when $\sigma = 10$ and $\sigma_{i,j} \sim \mathcal{U}([1, 10])$, respectively.

VI. CONCLUSION

In this paper, we have introduced a unified generic framework for adaptive importance sampling (AIS) that encompasses many well-known methods available in the literature. Several AIS algorithms have been described, their dynamics have been discussed, their consistency and limiting distributions analyzed, and their performance compared on a multimodal bivariate example.

REFERENCES

[1] A. Doucet and X. Wang, "Monte Carlo methods for signal processing," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 152–170, Nov. 2005.

[2] X. Wang, R. Chen, and J. S. Liu, "Monte Carlo Bayesian signal processing for wireless communications," *Journal of VLSI Signal Processing*, vol. 30, pp. 89–105, 2002.

[3] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, pp. 5–43, 2003.

[4] L. Martino, J. Read, and D. Luengo, "Independent doubly adaptive rejection Metropolis sampling within Gibbs sampling," *IEEE Transactions on Signal Processing*, vol. 63, no. 12, pp. 3123–3138, 2015.

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

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

[7] M. F. Bugallo, L. Martino, and J. Corander, "Adaptive importance sampling in signal processing," *Digital Signal Processing*, vol. 47, pp. 36–49, 2015.

[8] O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert, "Population Monte Carlo," *Journal of Computational and Graphical Statistics*, vol. 13, no. 4, pp. 907–929, 2004.

[9] E. Koblents and J. Miguez, "A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models," *Statistics and Computing*, 2014.

[10] V. Elvira, L. Martino, D. Luengo, and M.F. Bugallo, "Improving population monte carlo: Alternative weighting and resampling schemes," *viXra:1601.0174*, 2015.

[11] J. M. Cornuet, J. M. Marin, A. Mira, and C. P. Robert, "Adaptive multiple importance sampling," *Scandinavian Journal of Statistics*, vol. 39, no. 4, pp. 798–812, December 2012.

[12] L. Martino, V. Elvira, D. Luengo, and J. Corander, "An adaptive population importance sampler," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 8088–8092, 2014.

[13] L. Martino, V. Elvira, D. Luengo, and J. Corander, "An adaptive population importance sampler: Learning from the uncertanity," *IEEE Transactions on Signal Processing*, vol. 63, no. 16, pp. 4422–4437, 2015.

[14] L. Martino, V. Elvira, D. Luengo, and J. Corander, "Layered adaptive importance sampling," *Statistics and Computing, to appear*, 2016.

[15] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Generalized multiple importance sampling," *arXiv preprint arXiv:1511.03095*, 2015.

[16] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," *Handbook of Nonlinear Filtering*, vol. 12, no. 656-704, pp. 3, 2009.

[17] V. Elvira, L. Martino, D. Luengo, and J. Corander, "A gradient adaptive population importance sampler," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4075–4079.

[18] I. Schuster, "Gradient importance sampling," *arXiv:1507.05781*, pp. 313–316, 2015.

[19] L. Martino, V. Elvira, D. Luengo, and J. Corander, "MCMC-driven adaptive multiple importance sampling," *Interdisciplinary Bayesian Statistics Springer Proceedings in Mathematics & Statistics (Chapter 8)*, vol. 118, pp. 97–109, 2015.

[20] A. L. Nguyen, "Regenerative Markov chain importance sampling," *Communications in Statistics - Simulation and Computation*, vol. DOI: 10.1080/03610918.2015.1043383, 2015.

[21] Z. I. Botev, P. LEcuyer, and B. Tuffin, "Markov chain importance sampling with applications to rare event probability estimation," *Statistics and Computing*, vol. 23, pp. 271–285, 2013.

[22] X. Yuan, Z. Lu, and C. Z. Yue, "A novel adaptive importance sampling algorithm based on Markov chain and low-discrepancy sequence," *Aerospace Science and Technology*, vol. 29, pp. 253–261, 2013.

[23] P. Del Moral, A. Doucet, and A. Jasra, "Sequential Monte Carlo samplers," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 3, pp. 411–436, 2006.

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

[25] M.-S. Oh and J. Berger, "Adaptive importance sampling in Monte Carlo integration," *Journal of Statistical Computation and Simulation*, vol. 41, pp. 142–168, 1992.

[26] L. H. Koopmans, "Teaching singular distributions to undergraduates," *The American Statistician*, vol. 37, no. 4, pp. 313–316, 1983.