Submitted to Bernoulli
arXiv: 1211.2548

Consistency of Adaptive Importance Sampling and Recycling Schemes

JEAN-MICHEL MARIN[1,*] PIERRE PUDLO[2,**] and MOHAMMED SEDKI[3,†]

[1] *Université de Montpellier, Institut Montpellirain Alexander Grothendieck IMAG, Place E. Bataillon, 34095 Montpellier CEDEX, France. E-mail:* [*]jean-michel.marin@umontpellier.fr

[2] *Aix Marseille Université, CNRS, Centrale Marseille, I2M UMR 7373, 13453, Marseille CEDEX 13, France. E-mail:* [**]pierre.pudlo@univ-amu.fr

[3] *Université Paris-Sud, U1181, Hôpital Paul Brousse, Bât. Inserm 15/16, 16 Avenue Paul Vaillant-Couturier 94807 Villejuif CEDEX, France. E-mail:* [†]mohammed.sedki@u-psud.fr

Among Monte Carlo techniques, the importance sampling requires fine tuning of a proposal distribution, which is now fluently resolved through iterative schemes. Sequential adaptive algorithms have been proposed to calibrate the sampling distribution. Cornuet et al. (2012) provides a significant improvement in stability and effective sample size by the introduction of a recycling procedure. However, the consistency of such algorithms have been rarely tackled because of their complexity. Moreover, the recycling strategy of the AMIS estimator adds another difficulty and its consistency remains largely open. In this work we prove the convergence of sequential adaptive sampling, with finite Monte Carlo sample size at each iteration, and consistency of recycling procedures. Contrary to Douc et al. (2007a), results are obtained here in the asymptotic regime where the number of iterations is going to infinity while the number of drawings per iteration is a fixed, but growing sequence of integers. Hence some of the results shed new light on adaptive population Monte Carlo algorithms in that last regime and give advices on how the sample sizes should be fixed.

Keywords: Monte Carlo methods, importance sampling, sequential Monte Carlo, population Monte Carlo, adaptive algorithms, triangular arrays.

1. Introduction

A customary aim of Monte Carlo techniques is to approximate a targeted distribution Π on some space $\mathscr{X}$ with a random sample. Popular examples of such schemes are Markov chain Monte Carlo (MCMC, see, e.g. Robert and Casella, 2004) as well as importance sampling (IS, see Hesterberg, 1988, 1995; Ripley, 1987). Both methods (MCMC and IS) target Π by sampling $\mathscr{X}$ with auxiliary distributions, either according to a Markov kernel whose stationary distribution is Π or according to another proposal distribution Q. If Q charges the support of the target, IS corrects the discrepancy between Q and Π by weighting the sample with the Radon-Nikodym derivative of Π with respect to Q (which can be computed as ratio of both densities according a reference measure). The shortcomings of IS are well known: the distribution of the weights generally deteriorates except if the proposal distribution Q is tuned properly, see e.g. Owen and Zhou (2000). Adaptive importance sampling (AIS) as proposed by Douc et al. (2007a,b) or Cappé et al. (2008) is popular example of a sequential scheme that adapts the proposal distribution gradually over time. These authors fit the proposals to the target among a given set of distribution, namely finite mixtures of kernels. And note that AIS can be seen as a particular case of the Population Monte Carlo (PMC) of Cappé et al. (2004) which borrows principles from both IS and MCMC.

Moreover, as with MCMC, we can collect all the simulations along time in sequential schemes such as AIS to built an approximation of the target. In many real problems where computing the Radon-Nikodym derivates, *i.e.*, the density of the target according to the reference measure, is time consuming, recycling the successive samples generated during the run of the algorithm is crucial. In the particular case of AIS, we end up with samples drawn according to the successive proposal distributions we have tried at each stage of the algorithm. But Veach and Guibas (1995) and Owen and Zhou (2000) have shown that considering different samples drawn from different proposals as a single sample drawn from the mixture of proposal and correcting the weights according to this remark can stabilize IS approximation by reducing the variance of the weights. Indeed, the Radon-Nikodym derivates of the target with respect to the mixture distribution becomes very large only on the part of the space which is of low probability for each component of the mixture of proposals. He and Owen (2014) extended this work and optimized the rates of the mixture to gain further efficiency.

Cornuet et al. (2012) relied on the clever recycling strategy of Owen and Zhou (2000) to propose adaptive multiple importance sampling (AMIS). The AMIS is a sequential scheme in the same vein as Cappé et al. (2008), combining multiple importance sampling methods and adaptive techniques: the novelty of the AMIS is the update of the weights of all past simulations following Veach and Guibas (1995) and Owen and Zhou (2000) which recycles all past samples to learn the new proposal at the current iteration of the sequential scheme. On various numerical experiments where the target is the posterior distribution of some population genetics data sets, Cornuet et al. (2012) show considerable improvements of the AMIS in effective sample size (denoted further ESS, see Liu, 2008, chapter 2), *i.e.*, AMIS manage to reduce the variance of the weights. In another genetical example where the target is some posterior distribution, Sirén, Marttinen and Corander (2010) claim that the AMIS was efficient while other sequential importance sampling scheme misses the target.

We study here the theoretical proporties of a Modified version of AMIS (MAMIS), which introduces a simpler recycling strategy than AMIS. The MAMIS strategy, that was proposed in a previous version of this work, has been considered in various works such as Martino et al. (2016); Schuster (2015a,b); Bugallo, Martino and Corander (2015); Cameron and Pettitt (2014); Feroz et al. (2013). And Martino et al. (2015) rely on the same adaptive scheme as MAMIS to build their Adaptive Population Importance Sampling, called APIS. Note also that Šmídl and Hofman (2014) consider the use of different Population Monte Carlo algorithms to estimate the parameter of a highly nonlinear state space model for which the evaluation of the likelihood function requires extensive numerical calculations, as an automated method of instantaneous radiation situation assessment that does not underestimate its uncertainty. They found that both AMIS and MAMIS algorithms have similar performances and give the best results among the set of considered algorithms. Finally, for Gaussian processes with latent variables, Xiong, Šmídl and Filippone (2016) study the application of AMIS when the likelihood is explicit, and propose to rely on a Pseudo-Marginal AMIS/MAMIS for non-Gaussian likelihoods, where the marginal likelihood is unbiasedly estimated. The results suggest that the proposed framework, relying on AMIS and MAMIS, outperforms MCMC-based inference of covariance parameters in a wide range of scenarios and remains competitive for moderately large dimensional parameter spaces.

However, no proof of convergence had yet been provided in the literature. It is worth noting that the update of the weights at each iteration according to all learnt importance distributions introduces long memory dependence between the samples, and a bias which is not controlled by theoretical results. Even in very simple settings where the target is Gaussian, while the family of proposals are decentered Student $t(3)$ distributions, asymptotic consistency of the AMIS/MAMIS remains an open problem. The main purpose of this paper is to fill in this gap, and to prove the consistency of the algorithm at the cost of a slight modification in the adaptive process.

Sequential Monte Carlo scheme are presented in Section 2, including the algorithm we intend to study in the present paper, namely MAMIS. We suggest running an adaptive importance sampling algorithm that learns the new parameter of the proposal on the last simulated sample of size N_t, weighted with the classical importance sampling weights. The only recycling procedure in our algorithm is in the final stage that merges all the previously generated samples in the spirit of Veach and Guibas (1995), Owen and Zhou (2000) or Cornuet et al. (2012). The final stage outputs a weighted empirical distribution composed of $\Omega_T = N_1 + \cdots + N_T$ particles weighted as if they were simulated from the mixture of learnt proposals.

In Douc et al. (2007a) for instance, the consistency of the adaptive population Monte Carlo schemes is proven assuming that the number of iterations, say T, is fixed and that the number of simulations within each iteration, $N = N_1 = N_2 = \cdots = N_T$, goes to infinity. We decided to adopt a more realistic asymptotic setting in this paper. Contrary to these last results, the convergence of Theorem 3.5 holds when $N_1, \ldots, N_T$ is a growing, but fixed sequence and T goes to infinity. Hence the proofs of Theorem 3.2 provide new insights on adaptive PMC in that last asymptotic regime.

The strong convergence of the learnt parameters to the suitable parameter θ^* one seeks is given in Theorem 3.2. Its proof relies on a clever application of the Chebyshev inequality to obtain the almost sure consistency. To identify the main ideas of the proof clearly, we assumed that the tuning parameter is a (generalized) moment of the targeted distribution Π, see the discussion in Section 2.2. Interestingly Theorem 3.2 assumes that the sample size grows fast enough to infinity so that $\sum_t 1/N_t$ is finite, a condition that occurred sometimes in the literature studying sequential Monte Carlo schemes, see e.g. Forbes and Fort (2007).

The final merging of our algorithm updates the weights as if all simulations were drawn from a mixture of all learnt proposals. The consistency of this final output is given in Theorem 3.5, proven in Section 4. The result is not a straightforward consequence of asymptotic theorems and requires the introduction of a new weighting according to a given proposal distribution that is more simple to study, although biased and non explicitly computable (because the suitable value θ^* of the parameter is unknown). Under the set of assumptions given below, this last weighting scheme is consistent (see Proposition 4.4) and is comparable to the clever weighting as mixture of all learnt proposals, which yields the consistency proven in Theorem 3.5.

To sum up, Section 2 presents the algorithms, including AMIS/MAMIS in details and introduces the main notations of the paper. The main results of the paper are given in Section 3. Their proofs are given in Section 4. The paper ends with a conclusion and a discussion in Section 5.

2. Adaptive importance sampling

Simulating according to a target distribution Π whose probability density is known, up to a normalizing constant, is an important issue of Monte Carlo methods, in the field of stochastic modeling. Specifically, when conducting a Bayesian analysis, these algorithms serve to sample the parameter space $\mathscr{X}$ with respect to the posterior distribution. In this case, the posterior distribution is the target distribution Π and is known up to a normalizing constant. Simulating with a simple algorithm from Π is impossible in most situations. Different strategies exist and can be divided into two large families of algorithms: Markov chain Monte Carlo (MCMC) methods and importance sampling (IS) methods. These days, a new generation of hydrid algorithms have emerged, such as particle MCMC (Andrieu, Doucet and Holenstein, 2010). The hybrid algorithms bypass the issue of MCMC methods of assessing the convergence to the stationary distribution by relying on IS and correct the use of the wrong distribution at each iteration by weighting the samples.

Here we focus on importance sampling and its theoretical properties. Such importance sampling methods rely on the definition of an instrumental (or importance sampling) distribution, that serves as sampling distribution to cover the support of the target. The discrepancy between the sampling and the target distribution is corrected by weighting each simulation. But the efficiency of the method depends heavily on the choice of the instrumental distribution, whose choice is known to be difficult. To solve the issue, iterative and adaptive algorithms have been proposed in the literature. These algorithms build the instrumental distribution gradually by learning features of the target at each stage.

More formally, we wish to simulate according to the target and approximate integrals $\Pi(\psi) = \int \psi(x)\Pi(\mathrm{d}x)$ for a large class of function ψ: polynomial functions to obtain approximations of the moments of the target, but also indicator functions to approximate probabilities of certain events,... If Π is absolutely continuous with respect to the instrumental distribution Q, and if both of them have densities $\pi(x)$ and $q(x)$ with respect to a common reference measure $\mathrm{d}x$, then

$$\Pi(\psi) = \int \psi(x)\pi(x)\mathrm{d}x = \int \psi(x)\frac{\pi(x)}{q(x)}q(x)\mathrm{d}x = \int \psi(x)\frac{\pi(x)}{q(x)}Q(\mathrm{d}x) = Q\left(\psi\frac{\pi}{q}\right).$$

Note that the ratio $\pi(x)/q(x)$ does not depend on the reference measure and is actually the Radon-Nikodym derivative of Π with respect to Q. It serves as a weight in importance sampling (IS) algorithm: if $X_1, \ldots, X_N$ are N iid particles simulated according to Q, IS algorithms approximate the integral $\Pi(\psi)$ with the weighted average

$$\widehat{\Pi}_N^{IS}(\psi) = \frac{1}{N} \sum_{i=1}^{N} \frac{\pi(X_i)}{q(X_i)} \psi(X_i)\,.$$

Obtaining a good sampling distribution Q is not an easy task, since the distribution Q must have queues as large as the target Π (otherwise the weights degenerate). A strategy is to set a well chosen parametric family of distributions and to gradually learn which distribution is the best sampling distribution Q from that family. This leads to algorithms with a time dimension over standard IS methods, *i.e.*, iterative algorithms. The first algorithm that has been proposed in this direction is Adaptive Importance Sampling (AIS, Douc et al., 2007a; Cappé et al., 2008), that can be seen as a particular case of Population Monte Carlo (PMC, Cappé et al., 2004; Douc et al., 2007b). But PMC is a larger family of algorithms since the instrumental distributions can be Markov kernels. Note also that PMC can be seen as a particular case of Sequential Monte Carlo (SMC, Del Moral, Doucet and Jasra, 2006). But the target of PMC can be a distributions of much larger dimension on a much larger product space.

Formally, at stage t of the AIS algorithm, see Algorithm 1, the X_i^t's are drawn independently from a common distribution Q^t. This distribution might be selected from a parametric family $Q(\theta)$ of distributions as $Q^t = Q(\widehat{\theta}_t)$ where $\widehat{\theta}_t$ depends on past samples. Each X_i^t is weighted with the Radon-Nikodym derivative $\mathrm{d}\Pi/\mathrm{d}Q^t$. This is the classical weight of importance sampling, see e.g. Robert and Casella (2004). Moreover, the weights do not need to be normalised if we can compute exactly the Radon-Nikodym derivatives $\mathrm{d}\Pi/\mathrm{d}Q^t$. Note that the number N_t of simulations X_i^t may vary from stage to stage. The resulting algorithm has been considered by Douc et al. (2007a) and Cappé et al. (2008) and is exposed in Algorithm 1.

Algorithm 1 Adaptive importance sampling

```
for $t = 1 \to T$ do
    select the importance distribution $Q^t = Q(\widehat{\theta}_t)$ by learning $\widehat{\theta}_t$ from past weighted samples
    for $i = 1 \to N_t$ do
        draw $X_i^t$ from $Q^t$
        set $\omega_i^t = (\mathrm{d}\Pi/\mathrm{d}Q^t)(X_i^t)$
    end for
end for
return the sample $(X_1^T, \ldots, X_{N_T}^T)$ with weights $(\omega_1^T, \ldots \omega_{N_T}^T)$
```

Here, whatever the value of $N_1, \ldots, N_T$, we can easily see that the empirical distribution

$$\widehat{\Pi}_T^{AIS} := \frac{1}{N_T} \sum_{i=1}^{N_T} \omega_i^T \delta_{X_i^T}$$

is an unbiased approximation of the target Π provided that Π is absolutely continuous with respect to each $Q(\theta)$. Indeed, conditionally on the past simulations, *i.e.*, on the σ-field $\mathscr{F}_T := \sigma\big(X_i^t;\ t < T, i < N_t\big)$,

$$\begin{aligned}\mathbb{E}\left(\int \psi(x)\widehat{\Pi}_T^{AIS}(\mathrm{d}x)\middle|\mathscr{F}_T\right) &= \frac{1}{N_T}\sum_{i=1}^{N_T}\mathbb{E}\left(\psi(X_i^T)\frac{\mathrm{d}\Pi}{\mathrm{d}Q(\widehat{\theta}_T)}(X_i^T)\middle|\mathscr{F}_T\right)\\ &= \frac{1}{N_T}\sum_{i=1}^{N_T}\int \psi(x)\frac{\mathrm{d}\Pi}{\mathrm{d}Q(\widehat{\theta}_T)}(x)\ Q(\widehat{\theta}_T,\mathrm{d}x)\\ &= \frac{1}{N_T}\sum_{i=1}^{N_T}\int \psi(x)\Pi(\mathrm{d}x) = \int \psi(x)\Pi(\mathrm{d}x).\end{aligned}$$

The above cited papers (Douc et al., 2007a,b; Cappé et al., 2008) only consider the important case where the sampling distribution Q^t as to be adapted from a parametric family of finite mixture distributions but the framework is more general. It might be tempting to minimise the Kullback-Leibler divergence between the target Π and the sampling distribution Q_t at each iteration of the algorithm. Indeed, with the above reasoning, each weighted empirical distribution Π_t^{AIS} is an unbiased approximation of the target Π and

$$\begin{aligned}\mathrm{KL}(Q(\theta)|\Pi) &= \int \log\left(\frac{\mathrm{d}\Pi}{\mathrm{d}Q}(x)\right)\Pi(\mathrm{d}x)\\ &\approx \int \log\left(\frac{\mathrm{d}\Pi}{\mathrm{d}Q(\theta)}(x)\right)\widehat{\Pi}_t^{AIS}(\mathrm{d}x)\\ &\approx \frac{1}{N_t}\sum_{i=1}^{N_t}\omega_i^t\log\left(\frac{\mathrm{d}\Pi}{\mathrm{d}Q(\theta)}(X_i^t)\right). \qquad (2.1)\end{aligned}$$

Thus, $\widehat{\theta}_{t+1}$ might be set as the minimizer (in θ) of the above empirical criterion. This is not what is proposed in Douc et al. (2007b) since fitting the parameter of a mixture distribution requires complex algorithms such as the EM algorithm (see, e.g. McLachlan and Krishnan, 2007, and the references therein). Instead, Douc et al. (2007b) learn the new value $\widehat{\theta}_{t+1}$ by resorting only to one step of the EM algorithm.

Douc et al. (2007b) proved that, when $N_1 = N_2 = \cdots = N_T = N$, and when $N \to \infty$, the random sequence $(\widehat{\theta}_1,\ldots,\widehat{\theta}_T)$ converges (in probability) to the deterministic sequence $(\theta_1,\ldots,\theta_T)$ where

$$\theta_{t+1} = \int h(x,\theta_t)\Pi(\mathrm{d}x) \qquad (2.2)$$

for some explicit bivariate function h. Due to the update of θ with one step of an EM algorithm, the sole guaranty we have is that the Kullback-Leibler divergence $\mathrm{KL}(Q(\theta)|\Pi)$ decreases along the deterministic sequence $(\theta_1,\ldots,\theta_T)$. And, moreover, the limit of this deterministic sequence, namely $\lim_{T\to\infty}\theta_T$, is the minimiser of the Kullback-Leibler divergence $\mathrm{KL}(Q(\theta)|\Pi)$. The major drawback of these asymptotical results is that they do not study how Monte Carlo errors might propagate along the iterations of the scheme. Actually, in the above asymptotical regime where each sample size N_t is sent to infinity, the Monte Carlo error of the previous iteration is sent to zero, hence $\widehat{\theta}_t$ is sent to the deterministic θ_t before entering the t iteration.

2.1. Recycling strategies

Since the seminal paper of Cappé et al. (2004), users were aware that one might built an approximation of the target relying on all the particles along the iterations of adaptative schemes such as Algorithm 1. In settings where calculating the importance weights is time consuming, such efficient recycling processes make sense. Indeed, each iteration provides an approximation $\widehat{\Pi}_t$ of the target Π with a weighted empirical distribution. Thus, the mixture weighting each $\widehat{\Pi}_t$ with the number of particles simulated during the t iteration, namely

$$\frac{1}{\Omega_T}\sum_{t=1}^{T} N_t \widehat{\Pi}_t$$

where $\Omega_T = N_1 + N_2 + \cdots + N_T$, is also an approximation of the target Π. With Algorithm 1, since each Π_t^{AIS} is unbiased, the resulting mixture, which is

$$\frac{1}{\Omega_T}\sum_{t=1}^{T}\sum_{i=1}^{N_t} \omega_i^t \delta_{X_i^t}, \quad \text{where } \omega_i^t = \frac{\mathrm{d}\Pi}{\mathrm{d}Q^t}(X_i^t) \tag{2.3}$$

is also unbiased.

This is a naive recycling procedure of the T samples drawn from different importance distributions:

$$\begin{aligned} X_1^1, \ldots, X_{N_1}^1 &\sim Q^1 = Q\big(\widehat{\theta}_1\big), \\ X_1^2, \ldots, X_{N_2}^2 &\sim Q^2 = Q\big(\widehat{\theta}_2\big), \\ &\vdots \\ X_1^T, \ldots, X_{N_T}^T &\sim Q^T = Q\big(\widehat{\theta}_T\big). \end{aligned}$$

Actually, if $Q^1, \ldots, Q^T$ are fixed before the first iteration of Algorithm 1, Veach and Guibas (1995) have shown that we can see the whole collection $(X_i^t; t \le T, i \le N_t)$ as drawn from the mixture

$$Q_{\text{mixt}} := \sum_{t=1}^{T} \alpha_t Q^t, \quad \text{where } \alpha_t = N_t/\Omega_T$$

hence that we can replace ω_i^t by the alternative weight

$$\widetilde{\omega}_i^t := \frac{\mathrm{d}\Pi}{\mathrm{d}Q_{\text{mixt}}}(X_i^t)$$

in the Equation (2.3) and still obtain an unbiased approximation of the target Π. This recycling strategy is more clever and Veach and Guibas (1995) have shown that this alternative weighting strategy stabilises the approximation by reducing the variance of the weights, as also emphasized by Owen and Zhou (2000). Indeed, the alternative weights, *i.e.*, the Radon-Nikodym derivate $\mathrm{d}\Pi/\mathrm{d}Q_{\mathrm{mixt}}(x)$ becomes very large only when x is in a part of space which is of low probability for each component Q^t of the mixture $Q_{\mathrm{mixt}} := \sum_{t=1}^{T} \alpha_t Q^t$.

The above stabilising strategy that recycles all past simulations can be used as well when the importance distributions Q^t are adapted as in the general Algorithm 1. But then we do not have any theoretical guarantee about

$$\frac{1}{\Omega_T} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \frac{\mathrm{d}\Pi}{\mathrm{d}Q_{\mathrm{mixt}}}(X_i^t)\delta_{X_i^t} \tag{2.4}$$

since the final weight of each particles depends on the whole collection $(X_i^t; t \leq T-1, i \leq N_t)$. Nevertheless, Cornuet et al. (2012) relied on the alternative recycling strategie (2.4) to update θ at each stage t of the scheme. The resulting algorithm, called adaptive multiple importance sampling or AMIS, was not proven consistent. But numerical examples given in the original article or in Sirén, Marttinen and Corander (2010), show considerable improvements in effective sampling size (denoted further ESS, see Liu, 2008, chapter 2), *i.e.*, AMIS manages to reduce the variance of importance weights.

2.2. An adaptive algorithm ending with a multiple recycling scheme

We end up with a precise description of MAMIS, the algorithm we intend to study in the rest of the paper. To avoid dealing with sequential M-estimators (see, e.g., Van der Vaart, 2000, Chapter 5) minimizing (2.1) at each stage of the algorithm, we rather assume that, whatever the criterion we believe in (Kullback-Leibler divergence, or distances based on moment differences), the suitable value of θ might be written as

$$\theta^* = \int h(x)\Pi(\mathrm{d}x) \tag{2.5}$$

where h is an explicitly known function. For instance, one can easily show that, when the family of proposals $Q(\theta)$ is composed of decentered Student $t(3)$ distributions with means θ, then the minimiser of the Kullback-Leibler divergence can be written as (2.5) with $h(x) = x$. Likewise Cornuet et al. (2012) discussed moment fitting at the end of Section 3 of their paper. The resulting algorithm, which is Algorithm 1 complemented with a multiple recycling strategy, is given in Algorithm 2. We hope improvements in the accuracy of the current update of θ against previous estimations by requiring that the sample size N_t grows at each iteration.

To simplify notations, we assume that the target Π and the proposal distributions $Q(\theta)$ have density $\pi(x)$ and $q(x, \theta)$ with respect to a common reference measure $\mathrm{d}x$. In such settings, the Radon-Nikodym derivates are simply ratio of densities.

Algorithm 2 The studied scheme: MAMIS

```
Require: an initial parameter $\widehat{\theta}_1$ and increasing sample sizes $N_1, \dots, N_T$.
 1: for $t = 1 \to T$ do
 2:     for $i = 1 \to N_t$ do
 3:         draw $X_i^t$ from $Q(\widehat{\theta}_t)$
 4:         compute $\omega_i^t = \pi(X_i^t) / q(X_i^t, \widehat{\theta}_t)$.
 5:     end for
 6:     compute
            $\widehat{\theta}_{t+1} = N_t^{-1} \sum_{i=1}^{N_t} \omega_i^t h(X_i^t)$        (2.6)
 7: end for
 8: set $\Omega_T = N_1 + \cdots + N_T$
 9: for $t = 1 \to T$ do
10:     for $i = 1 \to N_t$ do
11:         update $\omega_i^t = \pi(X_i^t) \Big/ \Omega_T^{-1} \sum_{k=1}^T N_k q(X_i^t, \widehat{\theta}_k)$
12:     end for
13: end for
14: return the empirical distribution
            $\widehat{\Pi}_T^{\text{MAMIS}} := \frac{1}{\Omega_T} \sum_{t=1}^T \sum_{i=1}^{N_t} \omega_i^t \delta_{X_i^t}$        (2.7)
```

The studied scheme is given in Algorithm 2. We name it modified adaptive multiple importance sampling or MAMIS. The learning process between lines 1 and 7 draws a sequence of samples from which it calibrates gradually the parameter θ. The new value of the proposal's parameter we compute at line 6 depends only on the last sample we have drawn. This is the only discrepancy from the AMIS algorithm of Cornuet et al. (2012): MAMIS updates the parameter θ with the last sample, while AMIS updates the parameter θ by taking into account all past simulations. More precisely, the formula replacing (2.6) in AMIS is

$$\widehat{\theta}_{t+1}^{\text{AMIS}} = \Omega_t^{-1} \sum_{s=1}^{t} \sum_{i=1}^{N_s} \tilde{\omega}_{s,i}^t h(X_i^s), \quad \text{with } \tilde{\omega}_{s,i}^t = \pi(X_i^s) \Big/ \Omega_t^{-1} \sum_{k=1}^{t} N_k q(X_i^s, \widehat{\theta}_k).$$

In MAMIS, the only recycling process occurs during the final stage after line 9. Finally, we should note that, if calculating $\pi(x)$ is time consuming, the value computed at line 4 should be stored in memory to perform more easily the update at line 11 during the recycling process.

Hence, the estimator of the integral $\Pi(\psi) = \int \psi(x)\Pi(\mathrm{d}x)$ is

$$\widehat{\Pi}_T^{\text{MAMIS}}(\psi) = \frac{1}{\Omega_T} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \left[\frac{\pi(X_i^t)}{\Omega_T^{-1} \sum_{k=1}^T N_k q(X_i^t, \widehat{\theta}_k)} \right] \psi(X_i^t), \tag{2.8}$$

based on the empirical distribution given in (2.7).

Finally we define the σ-fields $\mathcal{F}_t = \sigma\big(X_1^1, \dots, X_{N_1}^1, \dots, X_1^{t-1}, \dots, X_{N_{t-1}}^{t-1}\big)$ which form a filtration.

3. Consistency results

We state here our main results (on the learning process in Paragraph 3.2, and on the final output in Paragraph 3.3). For the sake of clarity, technical parts of the proof of the second results are postponed to Section 4. We begin with some hypothesis on the parametric family of proposals.

3.1. Assumptions on the family of proposals

We assume that the space Θ is a subset of the space $\mathbb{R}^d$ endowed with the Euclidean norm $\|\cdot\|$. The set $\mathscr{X}$ is a subset of a finite-dimensional vector space, equiped with a reference measure $\mathrm{d}x$. All $Q(\theta)$ for $\theta \in \Theta$ and Π are absolutely continuous with respect to the reference measure. They have densities $q(x,\theta)$ and $\pi(x)$ respectively. The minimal hypothesis for importance sampling schemes to provide consistent estimates is that Π is absolutely continuous with respect to all proposals: $\forall \theta \in \Theta$, $\Pi \ll Q(\theta)$. Hence we assume that $q(x,\theta) = 0$ implies $\pi(x) = 0$.

We can note that $\widehat{\theta}_{t+1}$ is defined in (2.6) as a linear combination of (random) values of h, and the only fact we can safely affirm on the coefficients of this combination is that they are positive. Therefore, to avoid any interruption of the algorithm, any positive linear combination of elements of Θ should fall into Θ. In particular, this implies that Θ cannot be a bounded subset of a Euclidean space.

We also impose some regularity conditions on the family of proposals $\{Q(\theta)\}_{\theta\in\Theta}$ which will ensure consistency of our procedure. For all $x \in \mathscr{X}$, $\theta \mapsto q(x,\theta)$ is continuous on Θ, and the joint function $(x,\theta) \mapsto q(x,\theta)$ is lower semicontinuous on $\mathscr{X} \times \Theta$. Moreover, when $\theta \to \theta^*$, $q(\cdot,\theta)$ converges to $q(\cdot,\theta^*)$ uniformly over compact sets, *i.e.*, for any compact subset K of $\mathscr{X}$,

$$\|q(\cdot,\theta) - q(\cdot,\theta^*)\|_{K,\infty} := \sup\big\{|q(x,\theta) - q(x,\theta^*)| \ :\ x \in K\big\}$$

converges to 0.

Finally, for all $\varepsilon > 0$ and all $x \in \mathscr{X}$, set

$$m_\varepsilon(x) := \inf\{q(x,\theta), \|\theta - \theta^*\| \le \varepsilon\}. \tag{3.1}$$

We assume that, for all ε small enough, and all x in the (possibly unknown) support of Π, the function $m_\varepsilon(x) > 0$.

3.2. Consistency of the learning process

We focus here on the learnt parameters $\widehat{\theta}_t$ defined in (2.6) and show convergence to the suitable value $\theta^* = \int h(x)\pi(x)\mathrm{d}x$. The MAMIS weight of a particle, see (2.8), is an average over the path $\widehat{\theta}_1, \dots, \widehat{\theta}_T$ in the parameter space Θ. Weak consistency, *i.e.*, convergence in probability, is not enough to control such averages, as there exists no Cesàro Lemma for the convergence in probability, see for instance Billingsley (1995), exercise 20.23 p. 272. Thus, we decided to rely on almost sure convergence. The challenge is to prove that the sequential algorithm do not accumulate Monte Carlo errors over iterations. Let us introduce the following class of functions.

Definition 3.1. *A function $\psi : \mathscr{X} \to \mathbb{R}^d$ belongs to the class $\mathscr{G}^2(\mathbb{R}^d)$ if and only if $\int \pi^2(x)\|\psi(x)\|^2/q(x,\theta)\mathrm{d}x$ is finite for all θ in Θ and depends continuously on θ.*

We can interpret the integrability condition in the above definition as follow: the classical importance sampling algorithm that estimates $\Pi(\psi)$ with the proposal $Q(\theta)$ have finite variance whatever the value of $\theta \in \Theta$. With such assumptions on the function h used to learn the suitable value of the parameter, and a condition on the sample sizes, we can show the following result.

Theorem 3.2. (i) *If $h \in \mathscr{G}^2(\mathbb{R}^d)$ and $N_t \to \infty$, the estimate $\widehat{\theta}_t$ tends to θ^* in probability when $t \to \infty$.* (ii) *If, additionally, $\sum_t 1/N_t < \infty$, then $\widehat{\theta}_t \to \theta^*$ almost surely.*

Proof of Theorem 3.2(i). Fix $\varepsilon > 0$ and $z > 0$. Recall that $\widehat{\theta}_t$ is defined in (2.6). By conditional independence, the trace of the conditional variance-covariance matrix of $\widehat{\theta}_{t+1}$ can be bounded by

$$\mathrm{tr}\,\mathrm{Var}\left(\widehat{\theta}_{t+1}\middle|\mathcal{F}_t\right) = \mathrm{tr}\,\mathrm{Var}\left(\widehat{\theta}_{t+1}\middle|\widehat{\theta}_t\right) \le \frac{1}{N_t}v(\widehat{\theta}_t) \tag{3.2}$$

where $v(\theta) = \int \pi(x)^2\|h(x)\|^2/q(x,\theta)\mathrm{d}x$. Thus, using conditional Chebyshev inequality,

$$\mathbb{P}\left(\|\widehat{\theta}_{t+1} - \theta^*\| > \varepsilon\middle|\widehat{\theta}_t = \theta\right) \le \frac{v(\theta)}{N_t\varepsilon^2} \quad \text{a.s.}$$

using the fact that $\mathbb{E}\|\widehat{\theta}_{t+1} - \theta^*\|^2|\widehat{\theta}_t) = \mathrm{tr}\,\mathrm{Var}(\widehat{\theta}_{t+1}|\widehat{\theta}_t)$. Multiplying by $\mathbf{1}\{\|\theta\| \le z\}$ on both sides of the above inequality and integrating over the distribution of $\widehat{\theta}_t$ leads to

$$\begin{aligned}\mathbb{P}\left(\|\widehat{\theta}_{t+1} - \theta^*\| > \varepsilon,\ \|\widehat{\theta}_t\| \le z\right) &\le \frac{\sup\{v(\theta), \|\theta\| \le z\}}{N_t\varepsilon^2}\,\mathbb{P}(\|\widehat{\theta}_t\| \le z) \\ &\le \frac{\sup\{v(\theta), \|\theta\| \le z\}}{N_t\varepsilon^2} \qquad (3.3)\end{aligned}$$

On the other hand, because $\mathbb{E}(\|\widehat{\theta}_t\|\,|\mathcal{F}_t) \le \Pi(\|h\|)$, we have $\mathbb{P}(\|\widehat{\theta}_t\| > z) \le \Pi(\|h\|)/z$. Combining the last inequality with (3.3) gives

$$\mathbb{P}\left(\|\widehat{\theta}_{t+1} - \theta^*\| > \varepsilon\right) \le \frac{\sup\{v(\theta), \|\theta\| \le z\}}{N_t\varepsilon^2} + \frac{\Pi(\|h\|)}{z}.$$

By assumption of Theorem 3.2, $h \in \mathscr{G}^2(\mathbb{R}^d)$, hence $v(\theta)$ is finite and continuous. In particular, its supremum over $\|\theta\| \le z$ is also finite. Thus,

$$\limsup_{t\to\infty} \mathbb{P}\left(\|\widehat{\theta}_{t+1} - \theta^*\| > \varepsilon \right) \le \frac{\Pi(\|h\|)}{z}$$

Since z can be arbitrarily large, we have proven that $\widehat{\theta}_{t+1} \to \theta^*$ in probability. □

Proof of Theorem3.2(ii). Set

$$C_\varepsilon := \sup\{v(\theta), \|\theta - \theta^*\| \le \varepsilon\}$$

which is finite because $h \in \mathscr{G}^2(\mathbb{R}^d)$. Using (3.2) as above, we obtain

$$\mathbb{P}\left(\left\|\widehat{\theta}_{t+1} - \theta^*\right\| > \varepsilon \middle| \|\widehat{\theta}_t - \theta^*\| \le \varepsilon \right) \le \frac{C_\varepsilon}{\varepsilon^2 N_t}. \tag{3.4}$$

Now, we recall that $\widehat{\theta}_t$ forms a (time-inhomogeneous) Markov chain. Thus, using (3.4),

$$\begin{aligned}\mathbb{P}\left(\bigcap_{t=T}^{T'} \|\widehat{\theta}_{t+1} - \theta^*\| \le \varepsilon \right) &= \mathbb{P}\left(\|\widehat{\theta}_{T+1} - \theta^*\| \le \varepsilon \right) \prod_{t=T+1}^{T'-1} \mathbb{P}\left(\|\widehat{\theta}_{t+1} - \theta^*\| \le \varepsilon \middle| \|\widehat{\theta}_t - \theta^*\| \le \varepsilon \right) \\ &\ge \mathbb{P}\left(\|\widehat{\theta}_{T+1} - \theta^*\| \le \varepsilon \right) \prod_{t=T+1}^{T'-1} \left(1 - \frac{C_\varepsilon}{\varepsilon^2 N_t} \right).\end{aligned}$$

And, when $T' \to \infty$, we obtain

$$\mathbb{P}\left(\bigcap_{t\ge T} \|\widehat{\theta}_{t+1} - \theta^*\| \le \varepsilon \right) \ge \mathbb{P}\left(\|\widehat{\theta}_{T+1} - \theta^*\| \le \varepsilon \right) \prod_{t\ge T+1} \left(1 - \frac{C_\varepsilon}{\varepsilon^2 N_t} \right).$$

Applying the logarithm on the product and classical results on series, because $\sum_t 1/N_t$ is finite, the infinite product $\prod_t (1 - C_\varepsilon/\varepsilon^2 N_t)$ converges (that is to say the limit is finite and strictly positive). In particular, the remainder of the infinite product in the right hand side of the above inequality tends to 1 when $T \to \infty$. Furthermore, because of the convergence in probability proven above, $\mathbb{P}\left(\|\widehat{\theta}_{T+1} - \theta^*\| \le \varepsilon \right)$ tends also to 1 and thus

$$\lim_{T\to\infty} \mathbb{P}\left(\bigcap_{t\ge T} \|\widehat{\theta}_{t+1} - \theta^*\| \le \varepsilon \right) = 1.$$

And we have proven that $\limsup_{T\to\infty} \|\widehat{\theta}_T - \theta^*\| \le \varepsilon$ almost surely. Since ε is arbitrary, the desired almost sure convergence holds true. □

Remark. Looking carefully at both parts of the proof of Theorem 3.2, we deal with Monte Carlo errors on θ^* through a conditional Chebyshev inequality. This inequality is based on the assumption that $h \in \mathscr{G}^2(\mathbb{R}^d)$, *i.e.*, a conditional L^2 assumption which we consider as a minimal assumption to have faith in the result of importance sampling schemes. The main drawback of this assumption is that the concentration rate given by the Chebyshev inequality is the pessimistic $1/N_t$, see, e.g., Equation (3.4). To control how the Monte Carlo error propagates along the iterations of the scheme, we added the additional assumption that the sum $\sum_t 1/N_t$ is finite.

But we can replace $h \in \mathscr{G}^2(\mathbb{R}^d)$ with a stronger assumption and thus weaken the assumption on N_t. For instance, if we assume that, for all θ, $\pi(\cdot)\|h(\cdot)\|/q(\cdot,\theta)$ is bounded by some $\gamma(\theta)$ and that γ is a continuous function of θ, then we can rely on the Hoeffding inequality instead of (3.4) to obtain a sharper conditional concentration rate which is exponentially decreasing in N_t. In this way, we can thus weaken the former assumption on N_t into $\sum_t \exp(-\lambda N_t)$ is finite for all $\lambda > 0$ and still obtain the almost sure consistency of the learnt parameters.

Nevertheless we believe that assumptions stronger than $h \in \mathscr{G}^2(\mathbb{R}^d)$ are difficult to check in concrete cases. Even this mild L^2 assumption might not hold in quite a few cases. Thus we have preferred to leave Theorem 3.2 as it. It has the merit of providing pratical guidance on how to allocate the computational effort between iterations of the scheme under mild assumptions on the family of proposals.

3.3. Consistency of the final recycling scheme

The remaining part of our results deals with the final output that merges all the samples. More precisely, Theorem 3.5 below says that the empirical sum of a function ψ on the merged, re-weighted sample given in (2.8) provides a consistent approximation of the integral $\Pi(\psi)$. The class of integrands $\psi \in \mathbb{L}^1(\Pi)$ for which the above holds is determined by the following class of functions.

Definition 3.3. *A function $\psi : \mathscr{X} \to \mathbb{R}$ belongs to the class $\mathscr{H}^2(\mathbb{R})$ if and only if the integral $\int \big[\pi(x)\psi(x)/q\big(x,\theta^*\big)\big]^2 q\big(x,\theta\big)\mathrm{d}x$ is finite for all $\theta \in \Theta$ and is a function of θ that is continuous at $\theta = \theta^*$.*

Likewise, the above class of functions might be interpreted in terms of quadratic moments. Note that, if ψ is in $\mathscr{H}^2(\mathbb{R})$, then ψ is in $\mathbb{L}^1(\Pi)$. Moreover, we have the following, which is a straightforward consequence of Lemma 4.1.

Proposition 3.4. *If φ is a measurable function, and if $|\varphi| \le \psi$ for some function $\psi \in \mathscr{H}^2(\mathbb{R})$, then $\varphi \in \mathscr{H}^2(\mathbb{R})$.*

Finally, recall from (3.1) that, by assumption, see Paragraph 3.1, $m_\varepsilon(x) := \inf\{q(x,\theta) : \|\theta - \theta^*\| \le \varepsilon\}$ is positive on the support of Π. We are now in a position to state the following strong consistency.

Theorem 3.5. *Assume that $h \in \mathscr{G}^2(\mathbb{R}^d)$ and $\sum_t 1/N_t < \infty$. Moreover, assume that, for some $\varepsilon > 0$, $\psi(\cdot)q(\cdot,\theta^*)/m_\varepsilon(\cdot)$ is in $\mathscr{H}^2(\mathbb{R})$. Then, the sum over the final weighted sample $\widehat{\Pi}_T^{MAMIS}(\psi)$ given in (2.8) tends almost surely to $\int \psi(x)\pi(x)\mathrm{d}x$ when $T \to \infty$.*

The proof of the above Theorem is detailed in the next Section: it is a straightforward consequence of Propositions 4.4(iii) and 4.6. But note that the function $q(\cdot,\theta^*)/m_\varepsilon(\cdot)$ is larger than 1 on $\mathscr{X}$, and goes to 1 as $\varepsilon \to 0$. Hence, because of Proposition 3.4, the assumption that $\psi(\cdot)q(\cdot,\theta^*)/m_\varepsilon(\cdot)$ is in $\mathscr{H}^2(\mathbb{R})$ implies that ψ is in $\mathscr{H}^2(\mathbb{R})$.

4. Proof of Theorem 3.5

Recall that the MAMIS estimator of the integral of ψ with respect to the target Π is given by

$$\widehat{\Pi}_T^{\mathrm{MAMIS}}(\psi) = \frac{1}{\Omega_T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\left[\frac{\pi(X_i^t)}{D_T(X_i^t)}\right]\psi(X_i^t),$$

where, for all $x \in \mathscr{X}$,

$$D_T(x) = \Omega_T^{-1}\sum_{k=1}^{T} N_k q(X_i^t, \widehat{\theta}_k). \tag{4.1}$$

And let us introduce a simpler estimator (but which cannot be computed in practice because θ^* is unknown), namely

$$\widehat{\Pi}_T^*(\psi) := \frac{1}{\Omega_T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{\pi(X_i^t)}{q\big(X_i^t,\theta^*\big)}\psi(X_i^t). \tag{4.2}$$

It differs from $\widehat{\Pi}_T^{\mathrm{MAMIS}}(\psi)$ by the fact that we have replace $D_T(X_i^t)$ by $q(X_i^t,\theta^*)$. Note that the auxiliary estimate defined in (4.2) is a (weighted) average of the random variables $\widehat{\pi}_t^*(\psi)$ given by

$$\widehat{\pi}_t^*(\psi) := \frac{1}{N_t}\sum_{i=1}^{N_t}\frac{\pi(X_i^t)}{q\big(X_i^t,\theta^*\big)}\psi(X_i^t). \tag{4.3}$$

Adapting the proof of Theorem 3.2, we can show that the random variables $\widehat{\pi}_t^*(\psi)$ tends to $\Pi(\psi)$ almost surely, see Proposition 4.4. Now, applying Cesàro Lemma (Lemma 4.3 with $b_t = N_t$) yields to the convergence of $\widehat{\Pi}_T^*(\psi) = \Omega_T^{-1}\sum_{t=1}^{T} N_t\widehat{\pi}_t^*(\psi)$.

To conclude the proof of Theorem 3.5, we show in Proposition 4.6 that the MAMIS estimator $\widehat{\Pi}_T^{\mathrm{MAMIS}}(\psi)$ has the same asymptotic behavior as $\widehat{\Pi}_T^*(\psi)$, namely that $\widehat{\Pi}_T^{\mathrm{MAMIS}}(\psi) - \widehat{\Pi}_T^*(\psi) \to 0$ almost surely.

4.1. On the functions of class $\mathscr{H}^{\mathbf{2}}(\mathbb{R})$

Lemma 4.1. *Assume that, for any $\theta \in \Theta$, the integral*

$$w_\psi(\theta) := \int \pi^2(x)\psi^2(x)\Big/ q^2\big(x, \theta^*\big) q(x, \theta) \mathrm{d}x$$

is finite. These conditions are equivalent: (i) *w_ψ is continuous at θ^*; and* (ii) *when $\theta \to \theta^*$,*

$$\int \frac{\pi^2(x)\psi^2(x)}{q^2\big(x, \theta^*\big)} \big| q(x, \theta) - q(x, \theta^*) \big| \mathrm{d}x \to 0.$$

Proof. Clearly, (ii) implies (i) because

$$|w_\psi(\theta) - w_\psi(\theta^*)| \leq \int \frac{\pi^2(x)\psi^2(x)}{q^2\big(x, \theta^*\big)} \big| q(x, \theta) - q(x, \theta^*) \big| \mathrm{d}x.$$

We now have to show that (i) implies (ii), that is to say, if θ_n is a sequence converging to θ, then $\mathbb{E}|Z_n - Z^*| \to 0$ where

$$Z_n = \frac{\pi(X)\psi^2(x)}{q^2(X, \theta_n)} q(X, \theta_n), \quad \text{and} \quad Z = \frac{\pi(X)\psi^2(x)}{q^2(X, \theta^*)} q(X, \theta^*)$$

for some random variable X with distribution Π. With the continuity assumptions on the family $Q(\theta)$, $Z_n \to Z$ almost surely. Because of (i), $w_\psi(\theta_n) = \mathbb{E}(Z_n) \to w_\psi(\theta^*) = \mathbb{E}(Z)$. And note that Z_n, Z are non negative. Now, fix $r > 0$. We have

$$\mathbb{E}\Big(Z_n \mathbf{1}\{Z_n \geq r\}\Big) \leq \mathbb{E}\Big(Z_n - Z_n \wedge (r - Z_n)_+\Big).$$

The last expected value goes to $\mathbb{E}(Z - Z \wedge (r - Z)_+)$ since both $\mathbb{E}(Z_n) \to \mathbb{E}(Z)$ (because of (i)) and $\mathbb{E}(Z_n \wedge (r - Z_n)_+) \to \mathbb{E}(Z \wedge (r - Z)_+)$ (by dominated convergence). Thus,

$$\limsup_n \mathbb{E}\Big(Z_n \mathbf{1}\{Z_n \geq r\}\Big) \leq \mathbb{E}\Big(Z - Z \wedge (r - Z)\Big).$$

Moreover, $|Z - Z \wedge (r - Z)_+| \leq Z$ and goes almost surely to 0 when $r \to \infty$. The dominated convergence theorem yields to

$$\lim_{r \to \infty} \limsup_n \mathbb{E}\Big(Z_n \mathbf{1}\{Z_n \geq r\}\Big) = 0,$$

that is to say the sequence $\{Z_n\}_n$ is uniformly integrable. Hence $|Z_n - Z|$, which is bounded by $Z_n + Z$, is also uniformly integrable and we get

$$\lim_n \mathbb{E}|Z_n - Z| = \mathbb{E} \lim_n |Z_n - Z| = 0. \qquad \square$$

We also need the following to control the conditional expected value of $\pi_t^*(\psi)$ knowing that $\widehat{\theta}_t = \theta$, which is

$$I_\psi^*(\theta) := \int \left[\frac{\pi(x)\psi(x)}{q(x,\theta^*)}\right] q(x,\theta)\mathrm{d}x. \tag{4.4}$$

Lemma 4.2. *If $\psi \in \mathscr{H}^2(\mathbb{R})$, then the integrals $I_\psi^*(\theta)$ defined in (4.4) are well defined for all θ and the map I_ψ^* is continuous at $\theta = \theta^*$.*

Proof. Fix any $\theta \in \Theta$. If X_θ is a random variable with distribution $Q(\theta)$, then $Y_\theta = \pi(X_\theta)\psi(X_\theta)/q(X_\theta,\theta^*)$ is square integrable since ψ is in $\mathscr{H}^2(\mathbb{R})$. Hence Y_θ is a L^1-random variable and its expected value, namely $I_\psi(\theta)$ is well defined.

Now, set $g(x) = \pi(x)\psi(x)/q(x,\theta^*)$. We have $|g(x)| \le \max(1, g^2(x))$, thus

$$\int |g(x)|\big|q(x,\theta) - q(x,\theta^*)\big|\mathrm{d}x \le \int \big|q(x,\theta) - q(x,\theta^*)\big|\mathrm{d}x + \int g^2(x)\big|q(x,\theta) - q(x,\theta^*)\big|\mathrm{d}x$$

The first integral in this bound goes to 0 because of Scheffé's Theorem, see e.g. Billingsley (1995), Theorem 16.12 p. 215. The second integral goes also to 0, because ψ is in $\mathscr{H}^2(\mathbb{R})$ and because of Lemma 4.1. Whence

$$|I_\psi^*(\theta) - I_\psi^*(\theta^*)| \le \int |g(x)|\big|q(x,\theta) - q(x,\theta^*)\big|\mathrm{d}x \to 0. \qquad \square$$

4.2. Convergence of some auxiliary variables

Let us begin by recalling the following Lemma, whose proof is obvious, using Cesàro Lemma on sequence of (non random) vectors.

Lemma 4.3. *Let $\{U_t\}$ be a sequence of random vectors and U another random vector. If $\{b_t\}$ is a sequence of positive real numbers such that $B_t = b_1 + \ldots + b_t$ tends to infinity, then the event $\Big\{U_t \to U\Big\}$ is included in the event $\Big\{B_t^{-1}\sum_{k=1}^t b_k U_k \to U_\infty\Big\}$.*

Using the technic of proof of Theorem 3.2 as detailed below, we have the following.

Proposition 4.4. *Assume that $h \in \mathscr{G}^2(\mathbb{R}^d)$, $\sum_t 1/N_t$ is finite and $\psi \in \mathscr{H}^2(\mathbb{R})$.*

(i) *When $t \to \infty$, $\big(\widehat{\pi}_t^*(\psi) - I_\psi^*(\widehat{\theta}_{t-1})\big)$ tends to 0 almost surely.*

(ii) *Moreover, under those assumptions, $I_\psi^*(\widehat{\theta}_t) \to I_\psi^*(\theta^*) = \Pi(\psi)$ almost surely.*

(iii) *Finally, $\widehat{\Pi}_T^*(\psi)$ tends to $\Pi(\psi)$ almost surely when $T \to \infty$*

Proof. To prove *(i)*, we first prove that $\big(\widehat{\pi}_t^*(\psi) - I_\psi^*(\widehat{\theta}_{t-1})\big) \to 0$ in probability. To this aim, fix $\varepsilon > 0$ and $z > 0$. We have $\mathbb{E}\big(\widehat{\pi}_t^*(\psi)\big|\mathcal{F}_t\big) = I_\psi^*(\widehat{\theta}_t)$ and, by conditional independence,

$$\mathrm{Var}\left(\widehat{\pi}_t^*(\psi)\Big|\mathcal{F}_t\right) \le w_\psi(\widehat{\theta}_t)/N_t.$$

Thus, with a conditional Chebyshev inequality,

$$\mathbb{P}\left(\left|\widehat{\pi}_t^*(\psi)-I_\psi^*(\widehat{\theta}_{t-1})\right|\geq\varepsilon\middle|\mathcal{F}_t\right)\leq w_\psi(\widehat{\theta}_t)\Big/(N_t\varepsilon^2).$$

By integrating over the event $\{\|\theta-\theta^*\|\leq z\}$, we obtain

$$\mathbb{P}\left(\left|\widehat{\pi}_t^*(\psi)-I_\psi^*(\widehat{\theta}_{t-1})\right|\geq\varepsilon,\ \|\widehat{\theta}_t-\theta^*\|\leq z\right)\leq C_z'\Big/(N_t\varepsilon^2),$$

where $C_z'=\sup\{w_\psi(\theta),\|\theta-\theta^*\|\leq z\}$ is finite because $\psi\in\mathscr{H}^2(\mathbb{R})$. On the other hand,

$$\mathbb{P}\left(\|\widehat{\theta}_t-\theta^*\|>z\right)\leq 2\Pi(\|h\|)/z.$$

Combining the last two inequalities yields

$$\mathbb{P}\left(\left|\widehat{\pi}_t^*(\psi)-I_\psi^*(\widehat{\theta}_{t-1})\right|\geq\varepsilon\right)\leq\frac{C_z'}{N_t\varepsilon^2}+\frac{2\Pi(\|h\|)}{z}.$$

Thus

$$\limsup_t\mathbb{P}\left(\left|\widehat{\pi}_t^*(\psi)-I_\psi^*(\widehat{\theta}_{t-1})\right|\geq\varepsilon\right)\leq\frac{2\Pi(\|h\|)}{z}.$$

Finally, because z can be arbitrarily large, we have proven that $|\widehat{\pi}_t^*(\psi)-I_\psi^*(\widehat{\theta}_{t-1})|\to 0$ in probability.

To obtain *(i)* it remains to show that the above convergence toward 0 is actually an almost sure convergence. Fix $\varepsilon>0$. Set $\Delta_{t+1}=\widehat{\pi}_{t+1}^*(\psi)-I_\psi^*(\widehat{\theta}_t)$, $A_{t+1}=\Big\{|\Delta_{t+1}|\leq\varepsilon\Big\}\cap\Big\{\|\widehat{\theta}_{t+1}-\theta^*\|\leq\varepsilon\Big\}$ and $\bar{A}_{t+1}$ the complementary event. We have

$$\mathbb{E}\Big\{(\Delta_{t+1})^2\Big|\widehat{\theta}_t=\theta\Big\}=\frac{1}{N_t}\left[\int\left(\frac{\pi(x)\psi(x)}{q\big(x,\theta^*\big)}\right)^2q\big(x,\theta\big)\mathrm{d}x-I^*(\theta)^2\right].$$

and the term between brackets is bounded from above by $C_\varepsilon'=\sup\big\{w_\psi(\theta),\|\theta-\theta^*\|\leq\varepsilon\big\}$ for any $\theta\in\bar{\mathcal{B}}(\theta^*,\varepsilon)$, where $\bar{\mathcal{B}}(\theta^*,\varepsilon)=\{\theta\in\Theta\ :\ \|\theta-\theta^*\|\leq\varepsilon\}$. Thus, for all $\theta\in\bar{\mathcal{B}}(\theta^*,\varepsilon)$,

$$\begin{aligned}\mathbb{P}\left(\bar{A}_{t+1}\middle|\widehat{\theta}_t=\theta\right)&\leq\mathbb{P}\left(|\Delta_{t+1}|>\varepsilon\ \middle|\widehat{\theta}_t=\theta\right)+\mathbb{P}\left(\|\widehat{\theta}_{t+1}-\theta^*\|\geq\varepsilon\middle|\widehat{\theta}_t=\theta\right)\\&\leq\frac{C_\varepsilon'+C_\varepsilon}{\varepsilon^2N_t},\end{aligned}$$

where we have used the Chebyshev inequality. The suprema

$$\begin{aligned}C_\varepsilon&=\sup\big\{v(\theta),\|\theta-\theta^*\|\leq\varepsilon\big\},\\C_\varepsilon'&=\sup\big\{w_\psi(\theta),\|\theta-\theta^*\|\leq\varepsilon\big\}\end{aligned}$$

are both finite because $h \in \mathcal{G}^2(\mathbb{R}^d)$ and $\psi \in \mathcal{H}(\mathbb{R})$. Hence $\mathbb{P}\left(\bar{A}_{t+1}\middle|A_t\right) \le C''_\varepsilon/N_t$ for some finite C''_ε. Now, since $(\Delta_t, \widehat{\theta}_t)$ is a (time-inhomogeneous) Markov chain, we have

$$\mathbb{P}\left(\bigcap_{t=T}^{\infty} A_{t+1}\right) \ge \mathbb{P}\left(A_{T+1}\right) \prod_{T=t}^{\infty} \left(1 - \frac{C''_\varepsilon}{N_t}\right).$$

The above infinite product tends to 1 since $\sum_t 1/N_t$ is finite. And $\mathbb{P}\left(A_{T+1}\right) \to 1$ because both Δ_{T+1} and $\|\widehat{\theta}_{T+1} - \theta^*\|$ tends to 0 in probability. Hence, Claim *(i)* is proven.

Claim *(ii)* follows from Theorem 3.2 and continuity of I^*_ψ at θ^* proven in Lemma 4.2. Combining *(i)* and *(ii)*, we obtain that $\widehat{\pi}^*_t(\psi)$ tends to $\Pi(\psi)$ almost surely. Thus, $\widehat{\Pi}^*_T(\psi)$, which is given by

$$\widehat{\Pi}^*_T(\psi) = \frac{1}{\Omega_T} \sum_{t=1}^{T} N_t \widehat{\pi}^*_t(\psi). \qquad \square$$

tends also to $\Pi(\psi)$ almost surely because of Lemma 4.3.

4.3. Controlling the discrepency between the MAMIS estimator and the auxiliary variable

The convergence of $\widehat{\Pi}^{\text{MAMIS}}_T(\psi) - \widehat{\Pi}^*_T(\psi)$ towards 0 almost surely is proven in Proposition 4.6 below, whose proof relies on some preliminary result given in Lemma 4.5. To this end, we define the function $D_T\big(\,\cdot\,\big) : \mathscr{X} \mapsto \mathbb{R}_+$ by

$$D_T\big(x\big) = \Omega_T^{-1} \sum_{k=1}^{T} N_k q\left(x, \widehat{\theta}_k\right)$$

which appears in the denominator of the MAMIS weights in (2.8). Because of the consistency of the learning scheme proven in Theorem 3.2, we are able to show in the following lemma that this denominator resembles the denominator of the classical importance sampling weight, when the proposal distribution is $Q(\theta^*)$.

Lemma 4.5. *Let K be a compact subset of Θ. The event $\left\{\widehat{\theta}_t \to \theta^*\right\}$ is included in the event where*

$$\lim_{T\to+\infty} \left\| \frac{q(\cdot,\theta^*)}{D_T(\cdot)} - 1 \right\|_{K,\infty} = 0.$$

Proof. Denote by $m_{\varepsilon,K}$ the infimum of $m_\varepsilon(x)$ on K, where $m_\varepsilon(\cdot)$ is the function defined in (3.1). Actually, $m_{\varepsilon,K}$ is the infimum of the function $(x,\theta) \mapsto q(x,\theta)$ on the compact set $K \times \bar{\mathcal{B}}(\theta^*,\varepsilon)$, $\bar{\mathcal{B}}(\theta^*,\varepsilon) = \{\theta \in \Theta \ : \ \|\theta - \theta^*\| \le \varepsilon\}$. By assumption, see Paragraph 3.1, this function is lower semicontinuous. Since a lower semicontinuous function attains its lower bound on any compact set, and $q(x,\theta) > 0$ for all x and θ, the infimum $m_{\varepsilon,K}$ is positive.

Now fix a point of the probability space in the event $\left\{\widehat{\theta}_t \to \theta^*\right\}$. There, there exists some t_ε such that, for all $t > t_\varepsilon$, $\|\widehat{\theta}_t - \theta^*\| < \varepsilon$. Hence, for all $T > t_\varepsilon$, and all $x \in \mathscr{X}$,

$$D_T(x) \geq \frac{1}{\Omega_T} \sum_{k=t_\varepsilon+1}^{T} N_k q(x, \theta_k) \geq \frac{\Omega_T - \Omega_\varepsilon}{\Omega_T} m_\varepsilon(x) \quad \text{where } \Omega_\varepsilon = \sum_{k=1}^{t_\varepsilon} N_k.$$

Therefore

$$\begin{aligned} \left|\frac{q(x,\theta^*)}{D_T(x)} - 1\right| &\leq \frac{\Omega_T}{(\Omega_T - \Omega_\varepsilon) m_\varepsilon(x)} \left|q(x,\theta^*) - D_T(x)\right| \\ &\leq \frac{\Omega_T}{(\Omega_T - \Omega_\varepsilon) m_{\varepsilon,K}} \Omega_T^{-1} \sum_{t=1}^{T} N_k \left\| q(\cdot,\theta^*) - q(\cdot,\widehat{\theta}_t)\right\|_{K,\infty}. \end{aligned} \tag{4.5}$$

The bound in (4.5) is uniform on K and goes to 0 using Lemma 4.3, which leads to the desired result. □

We can now state and prove the result controlling the difference between the MAMIS estimator and the auxiliary variable $\widehat{\Pi}_T^*(\psi)$.

Proposition 4.6. *Assume that* $h \in \mathscr{G}^2(\mathbb{R}^d)$ *and* $\sum_t 1/N_t < \infty$. *Moreover, assume that, for some* $\varepsilon > 0$, $\psi(\cdot) q(\cdot,\theta^*)/m_\varepsilon(\cdot)$ *is in* $\mathscr{H}^2(\mathbb{R})$. *Then*

$$\lim_{T\to+\infty} \widehat{\Pi}_T^{MAMIS}(\psi) - \widehat{\Pi}_T^*(\psi) = 0 \quad \textit{almost surely.}$$

Proof. Fix $\alpha > 0$. The integral

$$\int \left|\psi(x)\right| \frac{q(x,\theta^*)}{m_\varepsilon(x)} \pi(x) \mathrm{d}x$$

is finite because $|\psi(\cdot)| q^*(\cdot)/m_\varepsilon(\cdot) \in \mathscr{H}^2(\mathbb{R})$. Therefore we can find some compact subset K of $\mathscr{X}$ such that

$$\int_{\mathscr{X}\setminus K} \left|\psi(x)\right| \frac{q(x,\theta^*)}{m_\varepsilon(x)} \pi(x) \mathrm{d}x < \alpha.$$

Now, set $\psi_1(x) := \psi(x)\mathbf{1}\{x \in K\}$, $\psi_2(x) := \psi(x)\mathbf{1}\{x \notin K\}$ so that $\psi(x) = \psi_1(x) + \psi_2(x)$. And consider the event

$$E = \left\{\widehat{\theta}_t \to \theta^*\right\} \cap \left\{\widehat{\Pi}_T^*(|\psi_1|) \to \Pi(|\psi_1|)\right\} \cap \left\{\widehat{\Pi}_T^*(|\psi_2|) \to \Pi(|\psi_2|)\right\} \cap \left\{\widehat{\Pi}_T^*(\varphi) \to \Pi(\varphi)\right\}.$$

where $\varphi(x) := |\psi_2(x)| q(x,\theta^*) \Big/ m_\varepsilon(x)$. With Theorem 3.2, Proposition 4.4(iii) and Proposition 3.4, this event is of probability 1. Moreover note that, because of (3.1), $q(x,\theta^*)/m_\varepsilon(x) \geq 1$ and thus

$$\Pi(|\psi_2|) \leq \Pi(\varphi) = \int_{\mathscr{X}\setminus K} |\psi_2(x)| \frac{q(x,\theta^*)}{m_\varepsilon(x)} \pi(x) \mathrm{d}x < \alpha. \tag{4.6}$$

Then, using linearity of the operators $\widehat{\Pi}_T^{\mathrm{MAMIS}}$ and $\widehat{\Pi}_T^*$, we have

$$\begin{aligned}\left|\widehat{\Pi}_T^{\mathrm{MAMIS}}(\psi)-\widehat{\Pi}_T^*(\psi)\right| &\le \left|\widehat{\Pi}_T^{\mathrm{MAMIS}}(\psi_1)-\widehat{\Pi}_T^*(\psi_1)\right|+\left|\widehat{\Pi}_T^{\mathrm{MAMIS}}(\psi_2)-\widehat{\Pi}_T^*(\psi_2)\right|\\ &\le \left|\widehat{\Pi}_T^{\mathrm{MAMIS}}(\psi_1)-\widehat{\Pi}_T^*(\psi_1)\right|+\widehat{\Pi}_T^{\mathrm{MAMIS}}(|\psi_2|)+\widehat{\Pi}_T^*(|\psi_2|). \qquad (4.7)\end{aligned}$$

The first term in the right hand side of (4.7) can be controlled as follow:

$$\begin{aligned}\Delta_T := \left|\widehat{\Pi}_T^{\mathrm{MAMIS}}(\psi_1)-\widehat{\Pi}_T^*(\psi_1)\right| &\le \frac{1}{\Omega_T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{\pi(X_i^t)\left|\psi_1(X_i^t)\right|}{q(X_i^t,\theta^*)}\left\|\frac{q(\cdot,\theta^*)}{D_T(\cdot)}-1\right\|_{K,\infty}\\ &\le \left\|\frac{q(\cdot,\theta^*)}{D_T(\cdot)}-1\right\|_{K,\infty}\widehat{\Pi}_T^*(|\psi_1|).\end{aligned}$$

On the event E, using Lemma 4.5, the first term of the last bound goes to 0, and $\widehat{\Pi}_T^*(|\psi_1|)\to\Pi(|\psi_1|)$. Hence $\lim_T\Delta_T=0$ on E.

On the event E, the second term of the right hand side of (4.7) can be bounded by

$$\begin{aligned}\widehat{\Pi}_T^{\mathrm{MAMIS}}(|\psi_2|) &\le \frac{1}{\Omega_T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{\pi(X_i^t)|\psi_2(X_i^t)|}{q(X_i^t,\theta^*)}\ \frac{q(X_i^t,\theta^*)}{D_T(X_i^t)}\\ &\le \frac{\Omega_T}{\Omega_T-\Omega_\varepsilon}\widehat{\Pi}_T^*\left(|\psi_2(\cdot)|\frac{q(\cdot,\theta^*)}{m_\varepsilon(\cdot)}\right)=\frac{\Omega_T}{\Omega_T-\Omega_\varepsilon}\widehat{\Pi}_T^*(\varphi) \qquad (4.8)\end{aligned}$$

using the fact that, on E, there exists a t_ε such that, for all $t>t_\varepsilon$, $\|\widehat{\theta}_t-\theta^*\|<\varepsilon$. Hence, for all $T>t_\varepsilon$, and all $x\in\mathscr{X}$,

$$D_T(x)\ge\frac{1}{\Omega_T}\sum_{k=t_\varepsilon+1}^{T}N_kq(x,\theta_k)\ge\frac{\Omega_T-\Omega_\varepsilon}{\Omega_T}m_\varepsilon(x)\quad\text{where }\Omega_\varepsilon=\sum_{k=1}^{t_\varepsilon}N_k.$$

On the event E, $\widehat{\Pi}_t^*(\varphi)$ converges to $\Pi(\varphi)$ which is smaller than α because of (4.6). Moreover, $(\Omega_T-\Omega_\varepsilon)/\Omega_T\to 1$. Hence, on the event E,

$$\limsup_T\widehat{\Pi}_T^{\mathrm{MAMIS}}(|\psi_2|)\le\alpha$$

And, finally, on the event E, the third term of the right hand side of (4.7) converges to $\Pi(|\psi_2|)$ which is smaller than α using (4.6). Hence, on E,

$$\limsup_T\widehat{\Pi}_T^*(|\psi_2|)\le\alpha$$

Reporting in (4.7), we obtain that, on the event E of probability 1,

$$\limsup_T\left|\widehat{\Pi}_T^{\mathrm{MAMIS}}(\psi)-\widehat{\Pi}_T^*(\psi)\right|\le 2\alpha.$$

Because α is arbitrary small, we have proven the desired result. □

5. Conclusion and discussion

For a certain class of functions, we derived strong consistency of Algorithm 2. Apart from the restrictive assumption on θ^*, the algorithm encompasses many sequential adaptive importance sampling schemes. And the main novelty of the above results is in the asymptotic design we have considered: finite and fixed sample size at each iteration of the sequential scheme. The asymptotic framework we have adopted is very different from the common framework on adaptive importance sampling scheme, namely the number of iterations is fixed and at each iteration, and the sample size $N_1 = N_2 = \cdots = N_T = N$ goes to infinity. Indeed, if we want to derive guidelines to the user from the common, past results in the literature, we obtain the following, helpless advice: if you are dissatisfied by the result of the algorithm, you should throw away all your simulations, and restart the algorithm from the very first iteration with a larger sample size.

We proved a strong law of large numbers for a large class of integrands characterized by regularity conditions and for a general family of proposals. One of the clear benefit of the asymptotic regime we considered is that it reveals an important condition on the design of the algorithm in terms of the sizes of the samples at each stage. Indeed, to prove the strong consistency of the tuning parameter we assumed that N_t tend to infinity quickly enough with t so that $\sum_t 1/N_t$ is finite. This assumption is intriguing. It provides guidelines to set these sample sizes when running adaptive algorithms, namely that the major parts of the computational effort should lies in the last iterations of the algorithm, when the sampling distribution $Q(\theta)$ has been tuned to the target. While this is not a surprising guideline, it should be offset by the initialization issues of adaptive algorithms which have already been discussed in the literature. Very often, asymptotical results do not provide guidelines on how to initialize a sequential scheme. The original paper of Cornuet et al. (2012) proposed an initialization of the AMIS based on a logistic sample when nothing is known on the target. We stress here that the starting distribution is of great practical consequence: for instance, if the first sample misses a mode of the target distribution, we have almost no chance to see it during the whole process. That was summed up by Cornuet et al. (2012) as the "what-you-get-is-what-you-see" nature of the AMIS, but the advice is also true for any adaptive importance sampling scheme.

At least theoretically the intriguing assumption on N_t might be alleviate. It is due to the fact that we only assumed that $\pi(X)\|h(X)\|/q(X,\theta)$ has a finite quadratic moment when $X \sim Q(\theta)$. Assuming the above random variable has finite exponential moments, and using Chernoff inequalities instead of the Chebyshev bound might lead also to a strong law of large number on tuned θ. But assuming that the above random variable has exponential moments is a very strong assumption on the target and the family of sampling distributions used in the scheme. Hence we have left this road since we were not convinced that finite exponential moments is a reasonable assumption.

The present paper proposes a road to study adaptive, recycling scheme and prove their consistency in a relevant asymptotic framework. Since it is the first paper in this direction, some assumptions were limiting, in particular the assumption on θ^* in Equation (2.5). The next step is certainly to weaken the assumption, and maybe study algorithms relying on (2.2) where the suitable tuning parameters are defined sequentially.

Acknowledgments

This research was financially supported by the French "*Agence Nationale de la Recherche*" through the ANR project EMILE (ANR-09-BLAN-0145-04). All three authors thank Christian P. Robert for useful discussions on the AMIS. The first author was financially supported by the *Labex* NUMEV (*Solutions Numériques,Matérielles et Modélisation pour l'Environnement et le Vivant*); and the second author by the *Labex* Archimède in Marseille.

References

Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistiscal Society, Series B* **72** 269–342.

Billingsley, P. (1995). *Probability and measure*, third ed. *Wiley Series in Probability and Mathematical Statistics*. John Wiley & Sons Inc., New York.

Bugallo, M., Martino, L. and Corander, J. (2015). Adaptive importance sampling in signal processing. *Digital Signal Processing* **47** 36 - 49. Special Issue in Honour of William J. (Bill) Fitzgerald.

Cameron, E. and Pettitt, A. (2014). Recursive pathways to marginal likelihood estimation with prior-sensitivity analysis. *Statistical Science* **29** 397–419.

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

Cappé, O., Guillin, A., Marin, J.-M. and Robert, C. P. (2008). Adaptive importance sampling in general mixture classes. *Statistics and Computing* **18** 587–600.

Cornuet, J.-M., Marin, J.-M., Mira, A. and Robert, C. P. (2012). Adaptive Multiple Importance Sampling. *Scandinavian Journal of Statistics* **to appear**.

Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo Samplers. *Journal of the Royal Statistiscal Society, Series B* **68** 411-436.

Douc, R., Guillin, A., Marin, J. M. and Robert, C. P. (2007a). Convergences of adaptive mixtures of importance sampling schemes. *The Annals of Statistics* **35** 420–448.

Douc, R., Guillin, A., Marin, J.-M. and Robert, C. P. (2007b). Minimum variance importance sampling via Population Monte Carlo. *ESAIM: Probability and Statistics* **11** 427–447.

Feroz, F., Hobson, M., Cameron, E. and Pettitt, A. (2013). Importance Nested Sampling and the MultiNest Algorithm. *arXiv preprint arXiv:1306.2144*.

Forbes, F. and Fort, G. (2007). Combining Monte Carlo and mean-field-like methods for inference in hidden Markov random fields. *Image Processing, IEEE Transactions on* **16** 824–837.

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

Hesterberg, T. (1988). Advances in Importance Sampling PhD thesis, Stanford University.

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

Liu, J. S. (2008). *Monte Carlo Strategies in Scientific Computing. Series in Statistics.* Springer.

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

Martino, L., Elvira, V., Luengo, D. and Corander, J. (2016). Layered adaptive importance sampling. *Statistics and Computing* 1–25.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions* **382**. John Wiley & Sons.

Owen, A. and Zhou, Y. (2000). Safe and Effective Importance Sampling. *Journal of the American Statistical Association* **95** 135-143.

Ripley, B. D. (1987). *Stochastic Simulation.* John Wiley & Sons Inc.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*, second edition ed. Springer-Verlag.

Schuster, I. (2015a). Gradient Importance Sampling. *arXiv preprint arXiv:1507.05781.*

Schuster, I. (2015b). Consistency of Importance Sampling estimates based on dependent sample sets and an application to models with factorizing likelihoods. *arXiv preprint arXiv:1503.00357.*

Sirén, J., Marttinen, P. and Corander, J. (2010). Reconstructing population histories from single-nucleotide polymorphism data. *Molecular Biology and Evolution* **28** 673–683.

Van der Vaart, A. W. (2000). *Asymptotic statistics.* Cambridge university press.

Veach, E. and Guibas, L. J. (1995). Optimally Comabining Sampling Techniques For Monte Carlo Rendering. In *SIGGRAPH'95 Proceeding* 419–428. Addison-Wesley.

Šmídl, V. and Hofman, R. (2014). Efficient sequential Monte Carlo sampling for continuous monitoring of a radiation situation. *Technometrics* **56** 514–528.

Xiong, X., Šmídl, V. and Filippone, M. (2016). Adaptive Multiple Importance Sampling for Gaussian Processes. *arXiv preprint arXiv:1508.01050.*