Optimized Population Monte Carlo

Víctor Elvira, Emilie Chouzenoux

Optimized Population Monte Carlo

Víctor Elvira, *Senior Member, IEEE,* and Émilie Chouzenoux, *Senior Member, IEEE*

***Abstract*—Adaptive importance sampling (AIS) methods are increasingly used for the approximation of distributions and related intractable integrals in the context of Bayesian inference. Population Monte Carlo (PMC) algorithms are a subclass of AIS methods, widely used due to their ease in the adaptation. In this paper, we propose a novel algorithm that exploits the benefits of the PMC framework and includes more efficient adaptive mechanisms, exploiting geometric information of the target distribution. In particular, the novel algorithm adapts the location and scale parameters of a set of importance densities (proposals). At each iteration, the location parameters are adapted by combining a versatile resampling strategy (i.e., using the information of previous weighted samples) with an advanced optimization-based scheme. Local second-order information of the target distribution is incorporated through a preconditioning matrix acting as a scaling metric onto a gradient direction. A damped Newton approach is adopted to ensure robustness of the scheme. The resulting metric is also used to update the scale parameters of the proposals. We discuss several key theoretical foundations for the proposed approach. Finally, we show the successful performance of the proposed method in three numerical examples, involving challenging distributions.**

***Index Terms*—Importance sampling, Monte Carlo methods, population Monte Carlo, Newton algorithm, covariance adaptation, stochastic optimization, Langevin dynamics.**

I. Introduction

Intractable integrals appear in countless problems of science and engineering. For instance, in Bayesian inference the interest is in estimating a posterior distribution of an unknown parameter given a set of related data. For most realistic models, the posterior distribution cannot be obtained in a closed form, and even more, it is not possible to simulate samples from it. Therefore, obtaining moments of interests (e.g., the mean, the variance, the probability of a certain event) is unfeasible either via an exact closed form or through approximations involving direct sampling. Importance sampling (IS) is a popular type of Monte Carlo methods [1], [2], [3] for the approximation of intractable distributions and related integrals. In its standard procedure, IS requires the simulation of samples from another distribution (called proposal). The samples receive an importance weight that takes into account the mismatch between target and proposal distributions. IS is a theoretically solid mechanism with strong guarantees, such as consistency, central limit theorems, and explicit error bounds [3], [4]. The performance in IS strongly depends on the adequacy of the proposal distribution. Intuitively, a proposal is good when it is *close* to the integrand in the targeted integral. However, it is usually impossible to know in advance where the probability mass of the target distribution is located (e.g., in Bayesian inference, one only has access to the evaluation of an unnormalized version of the posterior distribution). Therefore, advanced strategies must be employed, usually involving more than one proposal, which is called multiple importance sampling (MIS) [5], combined with the adaptation of the multiple proposals, leading to adaptive importance sampling (AIS) schemes [6].

The literature of AIS is vast, including methods based on sequential moment matching such as AMIS [7], [8], that comprises a Rao-Blackwellization of the temporal estimators, and APIS that incorporates multiple proposals [9]. Other recent methods have introduced Markov chain Monte Carlo (MCMC) mechanisms for the adaptation of the IS proposals [10], [11], [12]. The family of population Monte Carlo (PMC) methods also falls within AIS. Its key feature is arguably the use of resampling steps in the adaptation of the location parameters of the proposals [13], [14]. The seminal paper [15] introduced the PMC framework. Since then, other PMC algorithms have been proposed, increasing the resulting performance by the incorporation of stochastic expectation-maximization mechanisms [16], non-linear transformation of the importance weights [17], or better weighting and resampling schemes [18]. The method we propose in this paper falls within the PMC framework.

The state-of-the-art AIS methods, and particularly those belonging to the PMC family, suffer from several limitations that prevent a wider application of IS to more challenging problems. First, in the case of PMC, the resampling step provokes the well-known path degeneracy (see for instance [18, Fig. 4]), endangering the diversity of the proposals in the subsequent iterations. Some attempts have been recently done to attenuate this problem, e.g., the LR-PMC in [18] first forms a partition of the samples and then performs independent resampling step in each subset. However, this is at the expense of worsening the local exploration, since each partition approximates the target with less samples (see more details in [18]). The second limitation, not only in PMC but also in AIS in general, is that most existing methods only adapt the location parameters of the proposals, while the scale parameter remains fixed from the beginning. This is a clear limitation, since it is well known that the scale parameters of the proposals can make a significant difference in the efficiency of the AIS algorithm. Moreover a fine manual tuning requires a prior knowledge about the scale of the posterior distribution. Finally, even if such optimal fine tuning was possible, there is a clear advantage in adapting the scale

V. Elvira is with the School of Mathematics at the University of Edinburgh (UK) and with The Alan Turing Institute (UK). É Chouzenoux is with Université Paris-Saclay, Inria, CentraleSupélec, Centre de Vision Numérique (France).

V.E. and É.C. acknowledge support from the *Agence Nationale de la Recherche* of France under PISCES (ANR-17-CE40-0031-01) and MAJIC (ANR-17-CE40-0004-01) projects. V.E. acknowledges support from ARL/ARO under grant W911NF-20-1-0126. É.C. acknowledges support from the European Research Council Starting Grant MAJORIS ERC-2019-STG-850925.

parameter over the iterations, depending where the proposals are placed. Moreover, this represents an extra challenge when the dimensions of the posterior are of different order of magnitude and/or present strong correlations.

Some families of AIS methods use geometric information about the target for the adaptation of the location parameters, yielding to optimization-based adaptation schemes. For example, the GAPIS algorithm [19] is an AIS method that exploits the gradient and the Hessian of the logarithm of the target, and also introduces an artificial repulsion among proposals to promote the diversity (without any resampling step). Other methods such as [20], [21] adapt the location parameters by performing at each sample several steps of the unadjusted Langevin algorithm (ULA) [22], which can also be seen as an instance of a stochastic gradient descent method. The covariance is also adapted in those methods by either computing the sample autocorrelation [20] or using second-order information [19], [21]. A covariance adaptation has been also explored via robust moment-matching mechanisms in [23], [24]. We refer the interested reader to the survey [6]. The use of optimization techniques within PMC framework remains however unexplored. It is worth mentioning that optimization inspired schemes have also shown to be an efficient strategy to improve practical convergence rate in MCMC algorithms (see the survey paper [25] and references therein). In particular, the works [26], [27], [28], [29], [30] fall in the framework of the so-called Metropolis adjusted Langevin algorithms (MALA), where the ULA scheme is combined with a Metropolis-Hastings step. The Langevin-based strategy yields proposed samples that are more likely drawn from a highly probable region, with the consequence of a larger acceptance probability. MALA can be further improved by rescaling the drift term by a preconditioning matrix encoding local curvature information about the target density, through the Fisher metric [31], the Hessian matrix [32], [33], [29] or a tractable approximation of it [34], [35], [36], [37]. Optimization-based methods for accelerating MCMC sampling of non-differentiable targets have also been considered, for instance in [27], [38].

In this work, we propose a new Optimized PMC (O-PMC) approach.[1] To the best of our knowledge, the proposed algorithm is the first within the relevant PMC family to incorporate explicit optimization steps in order to enhance the resampling-based adaptation by exploiting the geometry of the target. In O-PMC, the proposals are adjusted using a stochastic Newton-based step adapted to the sample values resulting from a suitable resampling strategy. In contrast to the aforementioned works, here the mean and covariance adaptation are performed jointly, with the advantage of fitting the proposal distributions locally, boosting the exploration and increasing the performance. A damped Newton strategy, incorporating two stabilization features is proposed for the mean adaptation, and the retained scale matrix is per-used for the covariance adaptation. We show on three sets of numerical examples that this novel methodology catalyzes the local adaptation without endangering the diversity of the proposals nor the stability of the trajectories.

The rest of the paper is structured as follows. Section 2 introduces the problem setting, the AIS framework, and optimization-based proposal adaptation rules. In Section 3, we present the proposed method. We discuss its rationale and theoretical foundations in Section 4, including also a toy example. Finally, we show three numerical examples in Section 5 and conclude in Section 6.

II. Bayesian inference via importance sampling

In this section, we describe the Bayesian inference framework, the generic importance sampling methodology, and the standard PMC, which is an adaptive IS (AIS) algorithm. Note that, as stated in the introduction, the range of applicability of O-PMC goes beyond Bayesian inference (e.g., in the first two examples presented in Section V, the target distribution is available in a closed form and not necessarily coming from a Bayesian inference problem).

A. Bayesian inference

We consider the estimation problem of a vector of unknowns $\mathbf{x} \in \mathbb{R}^{d_x}$ that is statistically connected through a probabilistic model to the vector $\mathbf{y} \in \mathbb{R}^{d_y}$ that contains the available data. The observation model is embedded into the likelihood function $\ell(\mathbf{y}|\mathbf{x})$. The Bayesian approach allows for the incorporation of available prior information about $\mathbf{x}$ in the so-called prior distribution $p_0(\mathbf{x})$. The so-called posterior distribution of the unknowns given the data (a.k.a. *target* distribution) can then be expressed thanks to the Bayes rule:

$$\widetilde{\pi}(\mathbf{x}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\mathbf{x})p_0(\mathbf{x})}{Z(\mathbf{y})}. \tag{1}$$

Very often, the interest lies in the computation of a specific moment of the posterior distribution which amounts to solving integrals under the generic form

$$I = \int h(\mathbf{x})\widetilde{\pi}(\mathbf{x})d\mathbf{x} = \frac{1}{Z}\int h(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}, \tag{2}$$

where h is any integrable function w.r.t. $\widetilde{\pi}(\mathbf{x})$. Unfortunately, in most cases of interest, Eq. (2) cannot be computed, either because the integral is intractable or because the posterior distribution is rarely available in a closed form, mostly because of the impossibility of computing the normalizing constant $Z(y) \triangleq \int \pi(\mathbf{x}|\mathbf{y})d\mathbf{x}$ (a.k.a. model evidence, marginal likelihood, or partition function). Hence, it is useful to define the non-negative function $\pi(\mathbf{x}|\mathbf{y}) \triangleq \ell(\mathbf{y}|\mathbf{x})p_0(\mathbf{x}) = Z(\mathbf{y})\widetilde{\pi}(\mathbf{x}|\mathbf{y})$. From now on, in order to ease the notation, we drop $\mathbf{y}$ in Z, $\pi(\mathbf{x})$, and $\widetilde{\pi}(\mathbf{x})$. In order to overcome this limitation, approximate methods must be employed.

B. Importance sampling

Importance sampling (IS) is a Monte Carlo methodology that allows for the approximation of distributions and integrals as those of previous section. Unlike the raw (or standard) Monte Carlo technique, the basic IS method simulates all samples from the so-called *proposal* distribution $q(\mathbf{x})$. The samples are weighted accordingly in such a way consistent

[1] A limited version of this work was presented by the authors in the conference paper [39].

estimators can be built. More precisely, IS is composed of the two following steps:

1) **Sampling.** Simulate K samples as
$$\mathbf{x}_k \sim q(\mathbf{x}), \qquad k = 1, ..., K.$$
2) **Weighting.** Assign an importance weight to each sample as
$$w_k = \frac{\pi(\mathbf{x}_k)}{q(\mathbf{x}_k)} \qquad k = 1, ..., K.$$
i.e., the ratio between the unnormalized target and the proposal distribution, evaluated at the specific sample.

This basic sampling-weighting procedure allows for the construction of the both next estimators:

- Unnormalized IS (UIS) estimator:
$$\hat{I} = \frac{1}{KZ} \sum_{k=1}^{K} w_k h(\mathbf{x}_k). \tag{3}$$
- Self-normalized IS (SNIS) estimator:
$$\widetilde{I} = \sum_{k=1}^{K} \bar{w}_k h(\mathbf{x}_k). \tag{4}$$

Both UIS and SNIS are consistent with K, while only UIS is unbiased. However, UIS can be used only if Z is known. The key of success in IS is an appropriate choice of the proposal q in such a way that the aforementioned estimators have a low variance. The variance of the UIS estimator is minimized when the proposal is $q(\mathbf{x}) \propto |h(\mathbf{x})|\pi(\mathbf{x})$, while the optimal proposal of the SNIS estimator is $q(\mathbf{x}) \propto |h(\mathbf{x})|\pi(\mathbf{x})$ [1], [2], [3]. However, in most of cases it is impossible to design such proposal because it does not have a known form where sampling is possible. Hence, adaptive methods are required in order to iteratively improve the proposal.

C. Multiple importance sampling

Multiple importance sampling (MIS) refers to the case where several proposals $\{q_n(\mathbf{x})\}_{n=1}^N$ are used instead of just one, as in the previous section. It is known that in MIS, many possible sampling and weighting schemes are possible, and we refer the interested reader to an exhaustive comparison and analysis in [5]. Let us consider the case where $K = N$ samples are simulated from the set of N proposals. One can proceed as follows:

1) **Sampling:** Each sample is simulated from each of the proposals as
$$\mathbf{x}_n \sim q_n(\mathbf{x}), \qquad n = 1, ..., N$$
2) **Weighting:** Among all possible weighting options, we describe two possibilities:
   - **Option 1**: Standard MIS (s-MIS):
$$w_n = \frac{\pi(\mathbf{x}_n)}{q_n(\mathbf{x}_n)}, \quad n = 1, \dots, N$$
   - **Option 2**: Deterministic mixture MIS (DM-MIS):
$$w_n = \frac{\pi(\mathbf{x}_n)}{\psi(\mathbf{x}_n)} = \frac{\pi(\mathbf{x}_n)}{\frac{1}{N}\sum_{j=1}^{N} q_j(\mathbf{x}_n)}, \quad n = 1, \dots, N,$$
where $\psi(\mathbf{x}) = \frac{1}{N}\sum_{j=1}^{N} q_j(\mathbf{x})$ is the mixture pdf.

We recall that more sampling options are also possible. In the two MIS schemes presented below, it is possible to build the UIS estimator and also to normalize the weights to create the SNIS estimator. It is important to note that, while Option 1 (s-MIS) seems a natural extension of IS to MIS, it has been shown to provide always worse performance than Option 2 (DM-MIS), quantified in the variance of the UIS estimator. In very simple examples, the difference of this variance in both cases can be of several orders of magnitude (see [5] for more details).

TABLE I
STANDARD PMC ALGORITHM.

1) **[Initialization]**: Set $\sigma > 0$, $(N, T) \in \mathbb{N}^+$. For $n = 1, \dots, N$, select the initial adaptive parameters $\boldsymbol{\mu}_n^{(1)} \in \mathbb{R}^{d_x}$ and $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_{d_x}$.
2) **[For $t = 1$ to T]**:
   a) Draw one sample per each proposal pdf,
$$\mathbf{x}_n^{(t)} \sim q_n^{(t)}(\mathbf{x}; \boldsymbol{\mu}_n^{(t)}, \boldsymbol{\Sigma}) \tag{5}$$
   with $n = 1, \dots, N$.
   b) Compute the importance weights,
$$w_n^{(t)} = \frac{\pi(\mathbf{x}_n^{(t)})}{q_n^{(t)}(\mathbf{x}_n^{(t)})}. \tag{6}$$
   with $n = 1, \dots, N$.
   c) Resample N location parameters $\{\boldsymbol{\mu}_n^{(t+1)}\}_{n=1}^N$ from the set of N weighted samples of iteration t.
3) **[Output, $t = T$]**: Return the pairs $\{\mathbf{x}_n^{(t)}, w_n^{(t)}\}$, for $n = 1, \dots, N$ and $t = 1, \dots, T$.

D. Adaptive Importance Sampling and Population Monte Carlo

Adaptive importance sampling (AIS) is an iterative procedure for the adaptation of one or several proposals. The literature of AIS is vast, specially in the last decade, and we refer to [6] for an exhaustive review. Here we focus on population Monte Carlo (PMC), a family of AIS algorithms where the adaptation is based on resampling previous weighted particles. Table I describes the standard PMC algorithm [15].

In Step 1), the algorithm is initialized with N proposals where the location parameter is set to $\boldsymbol{\mu}_n^{(1)}$ (or could be chosen randomly) and the scale parameter is also set to $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_{d_x}$, with $\sigma > 0$. Then, the algorithm runs for T iterations as follows. In Step 2a), exactly one sample is simulated from each proposal. An importance weight is assigned to each sample in Step 2b). In Step 2c), the location parameters of next iteration are chosen by resampling the population of samples with probability proportional to the importance weight. Finally, the set of NT weighted samples is returned so the classical

unnormalized or self-normalized IS estimators can be built similarly to Eqs. (3)-(4).

Several PMC-based algorithms have been proposed since the publication of [15]. For instance, the M-PMC in [16] adapts a mixture proposal, including the weight, location, and scale parameter of each kernel of the mixture. The recent DM-PMC [18], adapts the location parameters of N proposals while the scale parameters remain static (e.g., with Gaussian proposals, the means are adapted but not the covariance matrices). The algorithm runs also over T iterations, where at each of them, K samples per proposal are simulated and weighted. The N adapted location parameters are resampled from the population of the NK samples at time t.

E. Optimization-based samplers

The choice of a suitable proposal distribution is a key challenge in sampling algorithms and have major consequences in their performance. While algorithms where the proposals are parametrized by static parameters might be easier to set up, this is often suboptimal. The reason is that the properties of the sought target are rarely known *a priori*, particularly in challenging applications. Many iterative schemes for proposal adaptation have been proposed in the literature of MC samplers, with the aim of a better and faster target exploration, especially at large dimension. One of the most relevant strategies within this recent trend is the Langevin-based sampling methods. Let us assume that $\log \pi$ is continuously differentiable on $\mathbb{R}^{d_x}$. Langevin samplers are derived from discrete approximations of the continuous diffusion initially introduced in [40]. They use a gradient descent step to move the samples location in the direction of a local increase of the target. This leads to an iterative strategy called unadjusted Langevin algorithm (ULA) [22]:

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \frac{\epsilon^2}{2} \nabla \log \pi(\mathbf{x}^{(t)}) + \epsilon \boldsymbol{\omega}^{(t)}, \quad (7)$$

where, for every $t \in \mathbb{N}$, $\boldsymbol{\omega}^{(t)} \in \mathbb{R}^{d_x}$ is a realization of a standard Gaussian distribution and $\epsilon > 0$ is the discretization stepsize. Note that the above scheme can also be interpreted as a gradient descent method perturbed with an i.i.d. stochastic error. Convergence analysis of the ULA sampler can be found for instance in [41], [22]. As emphasized in the aforementioned works, except in very specific situations, the Markov chain generated by the ULA scheme has a unique stationary distribution which differs from the target π (see in particular [42] for a quantification of this discrepancy). This undesirable effect is a consequence of the discretization procedure, as it is not present for the continuous Langevin diffusion [43]. To overcome this limitation, the ULA can be combined with a Metropolis-Hasting (MH) strategy, based on an acceptance-reject procedure, leading to the so-called Metropolis adjusted Langevin algorithm (MALA) [26]. The latter method has proved ergodic convergence, under milder assumptions on π. Moreover, its nice stability opens the door to the introduction of acceleration strategies. In particular, more sophisticated scale matrices, integrating more information (e.g., curvature) about the target [34], [31], [30], [29], [44], [26], [45], can be adopted. Let us in particular mention the Newton MH strategy [30], [29], which consists in combining an MH procedure with the stochastic Newton update:

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \mathbf{A}(\mathbf{x}^{(t)}) \nabla \log \pi(\mathbf{x}^{(t)}) + \mathbf{A}^{1/2}(\mathbf{x}^{(t)}) \boldsymbol{\omega}^{(t)}, \quad (8)$$

where $\mathbf{A}(\mathbf{x}^{(t)})$ is the inverse (or an approximation of it, when undefined or too complex) of the Hessian matrix $\nabla^2 \log \pi(\mathbf{x}^{(t)})$. In the next section, we present our main contribution, that is a new adaptive importance sampling algorithm that integrates such Newton-based strategy within the proposal adaptation rule of an advanced PMC scheme.

TABLE II
O-PMC ALGORITHM.

1) **[Initialization]**: Set $\sigma > 0$, $(N, K, T) \in \mathbb{N}^+$, $\{\boldsymbol{\nu}_n\}_{n=1}^N$. For $n = 1, \dots, N$, select the initial adaptive parameters $\boldsymbol{\mu}_n^{(1)} \in \mathbb{R}^{d_x}$ and $\boldsymbol{\Sigma}_n^{(1)} = \sigma^2 \mathbf{I}_{d_x}$.
2) **[For $t = 1$ to T]**:
   a) **[Sampling]**: Simulate NK samples as
   $$\mathbf{x}_{n,k}^{(t)} \sim q_n^{(t)}(\mathbf{x}; \boldsymbol{\mu}_n^{(t)}, \boldsymbol{\Sigma}_n^{(t)}, \boldsymbol{\nu}_n) \quad (9)$$
   with $n = 1, \dots, N$, and $k = 1, \dots, K$.
   b) **[Weighting]**: Calculate the normalized IS weights as
   $$w_{n,k}^{(t)} = \frac{\pi(\mathbf{x}_{n,k}^{(t)})}{\frac{1}{N}\sum_{i=1}^N q_i^{(t)}(\mathbf{x}_{n,k}^{(t)})}. \quad (10)$$
   c) **[Adaptation]**: Adapt the location and scale parameters of the proposal
      i) **[Resampling step]** Resample N proposals densities from the pool of NK weighted samples at the iteration t. The means and scales of the resampled proposals are denoted as $\widetilde{\boldsymbol{\mu}}_n^{(t)}$ and $\widetilde{\boldsymbol{\Sigma}}_n^{(t)}$, respectively. See Section III-C for explicit definitions of the notations.
      ii) **[Optimization step]** Adapt the proposal parameters $\{(\boldsymbol{\mu}_n^{(t+1)}, \boldsymbol{\Sigma}_n^{(t+1)})\}_{n=1}^N$ according to (11)-(12).
3) **[Output, $t = T$]**: Return the pairs $\{\mathbf{x}_{n,k}^{(t)}, w_{n,k}^{(t)}\}$, for $n = 1, \dots, N$, $k = 1, \dots, K$ and $t = 1, \dots, T$.

III. NEWTON POPULATION MONTE CARLO

In this section, we present the novel algorithm *optimized population Monte Carlo (O-PMC)*, an AIS algorithm that belongs to the family of PMC algorithms (see Table I). O-PMC incorporates several features for an efficient adaptation of the IS proposals with the goal of approximating Eq. (2). The O-PMC is presented in Table II. The algorithm is initialized with N proposals whose location and scale parameters are denoted with $\boldsymbol{\mu}_n^{(1)} \in \mathbb{R}^{d_x}$ and $\boldsymbol{\Sigma}_n^{(1)} = \sigma^2 \mathbf{I}_{d_x}$, respectively, with $\sigma > 0$, i.e., the initial scale matrices are isotropic. We denote the static parameters of the N proposals as $\{\boldsymbol{\nu}_n\}_{n=1}^N$. Then, the algorithm runs for T iterations, each of them divided in three steps: sampling in Step 2a), weighting in Step 2b), and adaptation in Step 2c). Finally the set of weighted samples is returned, so IS estimators can be built. In the following, we detail the steps.

A. Sampling and weighting

In Step 2a), at iteration t, each proposal is used to simulate exactly K samples from it. Note that it would be possible to have a different number of samples per proposal, K_n, although this variation should be accordingly done with the resampling step of previous iteration $t-1$. For simplicity in the description of the algorithm, we stick to a fixed K.

The weighting scheme is applied in Step 2b) according to Eq. (10), and in particular, those are based on the deterministic mixture weighting scheme (DM-MIS) of Section II-C. Note that these weights have been shown to provide a lower variance in UIS estimator compared to those of Eq. (6) of the original PMC method. We can call these weights *spatial* DM-MIS weights, since the proposals involved in the mixture of the denominator belong to the iteration t (see other options with temporal or spatial-temporal mixtures in [6], [10]).

B. Resampling

The resampling step is the first adaptive procedure (Step 2c) in Table II) which is then followed by the optimization step. In the resampling step, we select a set of N proposals, including a set of new location parameters $\{\widetilde{\boldsymbol{\mu}}_n^{(t)}\}_{n=1}^N$ and the associated (inherited) scale parameters $\{\widetilde{\boldsymbol{\Sigma}}_n^{(t)}\}_{n=1}^N$. The resampling can be interpreted as a sampling procedure from one or several particle approximations of the target distribution.

Here we present a novel resampling framework, for which existing resampling schemes are particular cases, while novel schemes can be derived (we propose one new scheme). Note that in all existing resampling schemes, only the location parameters are resampled, while here we also resample the associated scale parameters (which is equivalent to resampling the proposals). In our novel framework, the set of N resampled proposals with location parameters $\widetilde{\boldsymbol{\mu}}_n^{(t)} \triangleq \mathbf{x}_{i_n^{(t)},j_n^{(t)}}^{(t)}$ and scale parameters $\widetilde{\boldsymbol{\Sigma}}_n^{(t)} \triangleq \boldsymbol{\Sigma}_{i_n^{(t)}}^{(t)}$ are obtained by sampling (randomly) or choosing (deterministically) N pairs of indexes $\{i_n^{(t)}, j_n^{(t)}\}_{n=1}^N$. The index $i_n^{(t)} \in \{1,...,N\}$ points to the ancestor proposal which generated the sample that has been resampled, while the index $j_n^{(t)} \in \{1,...,K\}$ selects the specific sample in the set $\{\mathbf{x}_{i_n^{(t)},k}\}_{k=1}^K$. Note that the resampled scale parameter is selected by using only the ancestor index $i_n^{(t)}$. Let us now propose three particular and interesting choices for the resampling strategies, encompassed within our versatile framework.

Global resampling (GR): The N location parameters are simulated i.i.d. from a single particle approximation $\widehat{\pi}_t^{NK}(\mathbf{x}) = \sum_{n=1}^N \sum_{k=1}^K \overline{w}_{n,k} \delta(\mathbf{x} - \mathbf{x}_{n,k})$, constructed by the set of NK weighted samples $\mathbf{x}_{n,k}$ obtained from Step 2a), and the normalized weights $\overline{w}_{n,k}^{(t)} = \frac{w_{n,k}^{(t)}}{\sum_{i=1}^N \sum_{j=1}^K w_{i,j}^{(t)}}$. Therefore the two indexes are simulated jointly (but each pair is independent from other pairs), leading to n-th pair of indexes $\{i_n^{(t)}, j_n^{(t)}\} = \{\ell, k\}$ with associated probabilities $\overline{w}_{\ell,k}^{(t)}$, $\ell = 1,...,N$ and $k = 1,...,K$. Note that, for such choice, the resampled particles are strongly correlated (e.g., if one weight $\overline{w}_{n,k}$ dominates, all resampled particles can be identical). This scheme is closely related to the resampling step in the seminal PMC method [15] and it has been recently proposed in [18], although in both aforementioned works it only applied to the location parameters.

Local resampling (LR): An alternative strategy consists in simulating exactly one sample per ancestor (i.e., proposal). In this case, alternative re-normalized weights $\widetilde{w}_{n,k} = \frac{w_{n,k}}{\sum_{j=1}^K w_{n,j}}$ are required, in such a way that $\sum_{k=1}^K \widetilde{w}_{n,k} = 1$ for all $n = 1,\dots,N$. Then, the index of the n-th resampled proposal, $i_n^{(t)} = n$, is chosen deterministically, while the index $j_n^{(t)} = k$ is sampled with probability $\widetilde{w}_{n,k}$, for each $k = 1,...,K$. The advantage of the LR scheme is that the N resampled particles are different, preserving the diversity in the exploration through N paths that interact only due to the denominator in (10). The drawback is that it also preserves paths that are in non-relevant parts of the space. The limitations of both GR and LR strategies are closely linked to the tradeoff between particle degeneracy and path degeneracy, which is well-known in particle filtering [46], [47].

'Glocal' resampling (GLR): We introduce an original hybrid resampling approach, particularly tailored for the optimization-based adaptation that follows after the resampling step. The resampling step is done by following an LR step (i.e., using the $\widetilde{w}_{n,k}$ weights and preserving the diversity), except for the iterations with t multiple of a given period parameter $\Delta \in \mathbb{N}^*$, where a GR step is performed instead. The rationale of this novel scheme is explained in Section IV-B.

Finally, note that other existing schemes, such as the independent resampling (IR) of [47], are also encompassed in this framework.

C. Optimization

1) General rule: The second adaptive feature of our algorithm lies in Step 2c)ii). Here, in order to improve the exploration performance, we propose to adopt a Newton-based strategy for the construction of the proposal used to draw the next KN samples. The proposal density for iteration $t+1$, is modified, with a new adapted mean, given by

$$\boldsymbol{\mu}_n^{(t+1)} = \widetilde{\boldsymbol{\mu}}_n^{(t)} + \mathbf{A}(\widetilde{\boldsymbol{\mu}}_n^{(t)}) \nabla \log \pi(\widetilde{\boldsymbol{\mu}}_n^{(t)}), \tag{11}$$

where $\mathbf{A}(\widetilde{\boldsymbol{\mu}}_n^{(t)})$ is an SDP matrix of $\mathbb{R}^{d_x \times d_x}$. The scale matrix of the proposal is also adapted, in order to be consistent with the above location update, i.e.,

$$\boldsymbol{\Sigma}_n^{(t+1)} = \mathbf{A}(\widetilde{\boldsymbol{\mu}}_n^{(t)}). \tag{12}$$

As can been seen from (11)-(12), the scale matrix parameter $\mathbf{A}(\cdot)$ plays an important role in our scheme, since it drives the direction and length of the adapted jump. In the following, we present our simple yet efficient strategy for the setting of this parameter.

2) Scaling matrix: Newton-based strategy amounts to integrating information of the inverse of the Hessian of $\log \pi$, in the update rule for $\mathbf{A}(\widetilde{\boldsymbol{\mu}}_n^{(t)})$. However, in general cases, $\widetilde{\pi}$ may not be log-concave so that numerical issues can arise in the inversion of the Hessian matrix. Furthermore, even when the inversion is well defined, one Newton iteration with unit stepsize does not necessarily yield an increase of $\log \pi$

[48]. We thus propose to overcome those issues by adopting a damped Newton strategy, incorporating two specific features that aim at enforcing the numerical stability of the proposed scheme. The scaling matrix is defined as:

$$\mathbf{A}(\widetilde{\boldsymbol{\mu}}_n^{(t)}) = \theta_n^{(t)} \boldsymbol{\Gamma}(\widetilde{\boldsymbol{\mu}}_n^{(t)}), \tag{13}$$

with

$$\boldsymbol{\Gamma}(\widetilde{\boldsymbol{\mu}}_n^{(t)}) = \begin{cases} \left(-\nabla^2 \log \pi(\widetilde{\boldsymbol{\mu}}_n^{(t)})\right)^{-1}, & \text{if } \nabla^2 \log \pi(\widetilde{\boldsymbol{\mu}}_n^{(t)}) \succ 0, \\ \widetilde{\boldsymbol{\Sigma}}_n^{(t)}, & \text{otherwise.} \end{cases} \tag{14}$$

Otherwise stated, the covariance of the n-th proposal is set by using second order information when it yields to a definite positive matrix; otherwise, it inherits the covariance of the $i_n^{(t)}$-th proposal that generated the sample. Moreover, we introduced $\theta_n^{(t)} \in (0, 1]$, which is a stepsize tuned according to a backtracking scheme in order to avoid the degeneracy of the Newton iteration, and thus of our adaptation scheme. Starting with unit stepsize value, we reduce it by factor $\tau = 1/2$ until the condition below is met:

$$\pi\left(\widetilde{\boldsymbol{\mu}}_n^{(t)} + \mathbf{A}(\widetilde{\boldsymbol{\mu}}_n^{(t)}) \nabla \log \pi(\widetilde{\boldsymbol{\mu}}_n^{(t)})\right) \geq \pi\left(\widetilde{\boldsymbol{\mu}}_n^{(t)}\right). \tag{15}$$

D. Related works in the literature

The PMC algorithms perform the adaptation of the proposals via a resampling scheme. This step can be viewed as a stochastic procedure, since it is based on a multinomial resampling with replacement. However, since the set of proposals in the next iteration is parametrized by the resampled particles, this procedure can be alternatively seen as an implicit optimization procedure (in general, this perspective is not mentioned in the literature). In this paper, we propose, for the first time up to our knowledge, an explicit optimization procedure incorporated within the adaptation part of the algorithm, more precisely after the resampling step is done. Moreover, we design a suitable resampling step that allows the optimization step to exploit the benefits of stochastic and deterministic adaptation.

The introduction of optimization-based rules for improving the exploration properties of other AIS methods, not belonging to the PMC family, has been explored in the recent works [19], [20], [21], [49]. In [20], the authors propose a gradient descent with decreasing stepsize update for the location parameters, while the covariance update relies on the calculation of the empirical covariance of the past samples. Moreover, there is only one proposal. In GAPIS [19], the location parameters are updated according to a Newton step, with the stepsize remaining static, while the covariance is adapted using the Hessian of $-\log \pi$. Finally, in NIMIS and LIMIS [21], a temporal mixture is constructed, in the spirit of AMIS [7] but using a mixture that increases the number of components with the iterations (instead of remembering the whole mixture simply for the calculation of the importance weight). In LIMIS, the location parameters move along a gradient direction while the covariance adaptation relies on second-order approximation of the target, both updates being evaluated using Runge-Kutta numerical integration. Up to our knowledge, the use of a Newton-based adaptation for both location/covariance parameters, has never been considered in PMC literature.

IV. Discussion

In the following, we discuss the key elements of the novel O-PMC algorithm.

A. Importance weights

In O-PMC K samples are simulated from N proposals at each iteration. Then, the importance weights are computed in Eq. (10). First, note that these weights do not follow the standard functional consisting on the ratio between the target and the proposal distributions, e.g., the sample $\mathbf{x}_{n,k}^{(t)}$ is simulated from the n-th proposal but in the denominator of $w_{n,k}^{(t)}$, the whole mixture $\sum_{i=1}^{N} q_i^{(t)}(\mathbf{x})$ is evaluated (instead of just $q_n^{(t)}$). This alternative choice for the weights, called balance heuristic [50] or deterministic weights [51], has been shown to be unbiased and even more, to provide IS estimators with reduced variance [5]. The benefits of such alternative weights go beyond the superior performance of the estimators, and provide advantages in the resampling adaptation, compared to the standard weights. The reason is that, when evaluating the whole mixture in the denominator, the importance weight captures the mismatch between the target and the whole mixture of proposals at the iteration t (and not only the particular proposal that generated the sample). The resampling stage done with these weights allows to over-sample regions that are under-represented (see next section).

Finally, note that a mixture with the whole set of proposals $\{q_n^{(\tau)}\}_{1 \leq n \leq N, 1 \leq \tau \leq t}$, could be placed in the denominator of the importance weight, in the spirit of the AMIS algorithm [7]. We have however discarded this option as it would increase the computational complexity, particularly when T is large [6].

B. Resampling schemes

PMC algorithms adapt the set of proposals via a resampling step. In the seminal PMC algorithm from [15], the resampling is done at each time step using the standard weights. New resampling schemes have been proposed in [18], [47]. Note that by resampling scheme, we do not only refer to the way the sampling of the indexes that will be replicated is done (as in [13], [14]). In PMC algorithms, the samples are not i.i.d., and hence it is possible to enforce different adaptation behaviors. It is important to note that unlike in adaptive MCMC methods, where modifying the adaptation can endanger the convergence of the method, in AIS the needed assumptions are milder, since the validity of the estimators is ensured by the importance weights.

In O-PMC we propose two possible resampling schemes that are designed so as to exploit the *optimization step* (see the step 2)c)ii) in Table II). The local resampling (LR), proposed in [18] is particularly suitable for the optimization step that follows the resampling. As described in Section III-B, the LR scheme ensures that one (and only one) sample simulated

from the n-th proposal survives. This is advantageous for the Newton-based step that follows the resampling, since in practice there are N *chains* or threads, corresponding to the resampled particle among the set $\{\mathbf{x}_{n,k}^{(t)}\}_{k=1}^{K}$ for each $n = 1, ..., N$, that is later adapted using the geometry of the target. It is interesting to see that these N *chains* interact only in the importance weights calculation (because the whole mixture is in the denominator, as explained in the previous section). This interaction is very effective in combination with the other features of O-PMC, since it can give a lower weight to specific samples that are in over-populated areas (i.e., other proposals are covering that region), even if this part of the target has high probability mass. Therefore, even if the location parameter of each proposal is independently displaced to a region of higher probability mass via the Newton step with the risk of overpopulating the region around the mode(s), the resampling step re-balances the proposals at each iteration to approximate the target as a mixture. In O-PMC, we have discarded the use of the standard global resampling (GR) [15], [18] since it is known to reduce the diversity, endangering the exploration of the target (after the GR step, all resampled particles can come from the same proposal, and even more, can be exactly the same).

Moreover, in O-PMC we present another variant called *glocal* resampling (GLR). The GLR scheme can be seen as a modified LR scheme, where, at every $\Delta \in \mathbb{N}^*$ iterations, a GR step is performed instead of an LR one. The rationale is to preserve for most iterations the diversity in the exploration of each proposal (with a mild interaction in the denominator of the importance weights). Periodically, every Δ iterations, the GR step enforces a stronger cooperation among paths, killing those who are in irrelevant parts of the space, and replicating those who are more promising (which allows for a more exhaustive local exploration in the next iterations). The GLR strategy keeps clear ties with the adaptive resampling [52], allowing to find a good balance between an accurate local approximation of the target and a global exploration.

C. Newton-based adaptation

1) Improvement w.r.t. Newton scheme: In the optimization step, a straightforward approach may be simply choosing the scaling metric by relying on the information of the Hessian of $\log \pi$. In this approach, we might set $\mathbf{A}(\widetilde{\boldsymbol{\mu}}_n^{(t)})$ as the invert of $\nabla^2 \log \pi(\widetilde{\boldsymbol{\mu}}_n^{(t)})$. In such a way, (11) would read as one Newton iteration applied to the maximization of function $\log \pi$, and initialized at $\widetilde{\boldsymbol{\mu}}_n^{(t)}$. However, there are two drawbacks for the Newton optimization method, namely (i) the requirement for convexity of $-\log \pi$ for proper definition of the iteration, (ii) the local convergence behavior, i.e., convergence only when initialized "sufficiently close" to a mode. We thus propose two controlling rules, to avoid these difficulties. First, our proposed scheme in (14) introduces a test, taking into account the fact that $\log \pi$ might not be necessarily log-concave at $\widetilde{\boldsymbol{\mu}}_n^{(t)}$, so that $\nabla^2 \log \pi(\widetilde{\boldsymbol{\mu}}_n^{(t)})$ might be non invertible. We take here advantage of the trajectory tracking inherent to the AIS method, by re-using the past scaling matrix from the particle ancestors. This is particularly advantageous, since the samples are in general close from the location parameter of the density where they were simulated. An alternative to this approach would be inheriting the scale parameter of the proposal (at iteration t) that is closer to the sample, as it is done for instance in [21]. This may increase the performance without increasing the complexity if the proposals are Gaussian pdfs and the criterion is based on the Mahalanobis distance (since these distances are implicitly computed in the denominator of the importance weights). Another more ambitious scheme would be doing the same, but considering the set of whole set of Nt present and past proposals. While these alternative may capture better the second order information of the target in the neighborhood of the sample, in general they would translate into an increase of complexity. This is highly related to the well-known increase of complexity in AMIS algorithm when the number of iterations grow, with efficient versions of the algorithm trying to alleviate this issue, e.g., the EAMIS [53]. Hence, O-PMC either accounts for the second-order information at the sample location or inherits a more stable one from a close location. Second, a stepsize is introduced in (13), which is computed following a simple backtracking procedure. The idea is to constrain the Newton step within a region in which we believe that the second order model, inherent from the Newton approximation on $\log \pi$ is reliable, using iterative trials for the stepsize. If a notable increase of $\log \pi$ is gained, then the model is believed to be a good representation of the original objective function. If there is not significant improvement, the model is considered invalid, and a new step is tried. It is worth noting that the fulfillment of the descent condition (15) in finite time is ensured under mild assumptions on $\log \pi$ (e.g., Lipschitz differentiability, see [48]). Moreover, under the same assumptions, the unit stepsize satisfies (15) as soon as $\widetilde{\boldsymbol{\mu}}_n^{(t)}$ is sufficiently close to a local maximum of $\log \pi$ [48]. Otherwise stated, the classical (fast) Newton move of the location parameters is retrieved as soon as the particles get close to the modes of the target. More sophisticated approximations for the Hessian (or its inverse) may be desirable when exploring very large scale multimodal distributions. Low-rank [32] or majorizing [34] approximations, proposed in the context of Langevin-based MCMC, appear as appealing alternatives. However, it is not straightforward to incorporate those approaches in the proposed O-PMC scheme, while keeping the versatility of the algorithm. The exploration in high-dimensional problems may be improved by imposing an isotropic/diagonal structure in the covariance matrix [54], and fitting it through a particle approximation via importance sampling, at the expense of reducing the efficiency of the estimators in highly correlated target distributions.

2) Connection with scaled Langevin dynamics: Our scaled gradient adaptation scheme (11)-(12) keeps interesting connections with the discretized version of the scaled Langevin diffusion, mentioned for instance in [26], [34] in the context of MCMC sampling. This discretized Langevin diffusion can be expressed as (using similar notations as in (7)):

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \epsilon^2 \mathbf{b}(\mathbf{x}^{(t)}) + \epsilon \mathbf{A}^{1/2}(\mathbf{x}^{(t)}) \boldsymbol{\omega}^{(t)}. \quad (16)$$

Hereabove, $\mathbf{b} : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ is the so-called drift term such that, for every $1 \leq i \leq d_x$ and every $\mathbf{x} \in \mathbb{R}^{d_x}$,

$$b_i(\mathbf{x}) = \frac{1}{2}\sum_{j=1}^{d_x} A_{i,j}(\mathbf{x})\frac{\partial \log \pi(\mathbf{x})}{\partial x_j} + |\mathbf{A}(\mathbf{x})|^{\frac{1}{2}} \sum_{j=1}^{Q} \frac{\partial}{\partial x_j}\left(A_{i,j}(\mathbf{x})|\mathbf{A}(\mathbf{x})|^{-\frac{1}{2}}\right), \quad (17)$$

with $\mathbf{A}(\mathbf{x}) \in \mathbb{R}^{d_x \times d_x}$ a symmetric definite positive (SDP) matrix with determinant $|\mathbf{A}(\mathbf{x})|$. A typical approximation, adopted in [34], consists in ignoring the second term in (17), leading to the simplified sampling scheme:

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \frac{\epsilon^2}{2}\mathbf{A}(\mathbf{x}^{(t)})\nabla \log \pi(\mathbf{x}^{(t)}) + \epsilon \mathbf{A}^{1/2}(\mathbf{x}^{(t)})\boldsymbol{\omega}^{(t)}, \quad (18)$$

where ϵ is a positive stepsize and $\{\mathbf{A}(\mathbf{x}^{(t)})\}_{t\geq 0}$ are positive symmetric definite positive scale parameters, possibly varying over the discrete time iterates indexed by $t \in \mathbb{N}$. Ergodicity of the chain generated by (18), combined with a Metropolis-Hasting step, was established in [34], for a large class of choices for $\{\mathbf{A}(\mathbf{x}^{(t)})\}_{t\geq 0}$ and ϵ. It is noticeable that our optimization-based adaptation scheme in Eq. (11)-(12) is closely related to (18), identifying $\theta^{(t)}$ with $\frac{\epsilon^2}{2}$. Note that no factor 2 is present in our covariance adaptation rule in Eq. (12), in a similar fashion as in the Newton MH sampler from [30], [29]. In this way, the scale parameter of the proposal adapts, in a robust way, to the curvature of the target distribution, as we show in the next toy example.

D. Toy example

We illustrate the behavior of O-PMC along iterations on a simple example where the target is a mixture of bivariate Gaussian distributions, with means $[-5, -5]^\top$ and $[6, 4]^\top$, covariances $[0.25, 0; 0, 0.25]$ and $[0.52, 0.48; 0.48, 0.52]$, and mixture weights 0.7 and 0.3, respectively. We run $T = 10$ O-PMC iterations with $(N, K) = (50, 20)$, and resampling schemes LR and GLR (with period $\Delta = 5$). We initialize the location parameters of the proposals randomly in the square $[-5, 5] \times [-5, 5]$, and the initial covariance is set to $\mathbf{\Sigma}_n^{(1)} = \mathbf{I}_2$. We also run the GR-PMC and LR-PMC algorithms [18], the AMIS algorithm [7], and the GAPIS algorithm[19] algorithms, all with the same parameters for a fair comparison. We display in Fig. 1 the evolution along the iterations of the proposals, including the location parameters (black dots) and scale parameters (green ellipses). We also show the two marginal pdfs of the target distribution (blue line) and the equally weighted mixture of proposals (red line). We notice that O-PMC algorithm moves rapidly the proposal locations to the two modes. Moreover, it fits the scale parameters of the proposals to the scale parameters of each mode (depending on the part of the space where the proposal is located). The target is thus very accurately estimated, in few iterations, as can be seen in the 2D plots as well as in the marginals. In contrast, both GR-PMC and LR-PMC schemes struggle to reach a reasonably good target approximation. We also observe the benefits offered by GLR, our novel resampling scheme, as it can be easily noticed the improved performance of O-PMC w.r.t. the LR variant. This is particularly visible between $t = 5$ and $t = 6$, as $t = 6$ corresponds to the first (periodic) callback to the GR resampling (see Sec. III-B for more details). GAPIS discovers both modes, but since the step-size is not adapted in this algorithm, after $t = 10$ iterations the proposals are still in between the initialization and the asymptotic value (that would be close to the modes of the targets). Interestingly, GAPIS incorporates a repulsion behavior between proposals that has certain parallelism with the resampling step in O-PMC (due to the DM-MIS weights). The advantage of O-PMC is that this implicit repulsion does not require extra parameters. In this example, AMIS fails to discover one of the modes due to an initialization that was designed to be challenging for all algorithms. A different initialization (or a much higher initial variance of the proposals) may help the algorithm at the expense of being more inefficient and taking more iterations to converge.

V. Numerical Results

In this section, we present several sets of experiments, in order to assess the performance of the proposed O-PMC algorithm. Three examples will be considered for the target, namely (i) two-dimensional Gaussian mixture, (ii) multi-dimensional banana-shaped distribution, (iii) posterior distribution arising in a spectral analysis problem. These examples are representative as they incorporate challenging features related to multi-modality and high dimensionality. In all examples, we compare with competitive state-of-the-art adaptive importance sampling techniques, namely GR-PMC and LR-PMC that are two variants of the DM-PMC algorithm [18] (GR-PMC and LR-PMC), AMIS [7] and GAPIS [19]. Let us notice that, for the retained settings, the time spent by the methods are actually very similar, which confirms the fairness of our comparisons.

A. Mixture of Gaussians

In this first example, we consider a multimodal target which is a mixture of five bivariate Gaussian pdfs (i.e. $d_x = 2$):

$$(\forall \mathbf{x} \in \mathbb{R}^2) \quad \widetilde{\pi}(\mathbf{x}) = \frac{1}{5}\sum_{i=1}^{5} \mathcal{N}(\mathbf{x}; \gamma_i, \mathbf{C}_i). \quad (19)$$

Here, we set the means $\gamma_1 = [-10, -10]^\top$, $\gamma_2 = [0, 16]^\top$, $\gamma_3 = [13, 8]^\top$, $\gamma_4 = [-9, 7]^\top$, $\gamma_5 = [14, -4]^\top$, and the covariance matrices $\mathbf{C}_1 = [5,\ 2; 2,\ 5]$, $\mathbf{C}_2 = [2,\ -1.3; -1.3,\ 2]$, $\mathbf{C}_3 = [2,\ 0.8; 0.8,\ 2]$, $\mathbf{C}_4 = [3,\ 1.2; 1.2,\ 0.5]$ and $\mathbf{C}_5 = [0.2,\ -0.1; -0.1,\ 0.2]$. The main challenge in this example is the ability in discovering the 5 different modes of $\bar{\pi}(\mathbf{x}) \propto \pi(\mathbf{x})$. We focus in our tests on the approximation of three quantities, namely the target mean $\mathrm{E}_{\widetilde{\pi}}[\mathbf{X}] = [2.4, 3.4]^\top$, the second moment $\mathrm{E}_{\widetilde{\pi}}[\mathbf{X}^2] = [101.04, 98.94]^\top$, and the normalizing constant $Z = 1$. Since we know the ground truth for these quantities, we can easily assess qualitatively the performance of the different techniques. Furthermore, since the problem is low dimensional, it is possible to approximate the

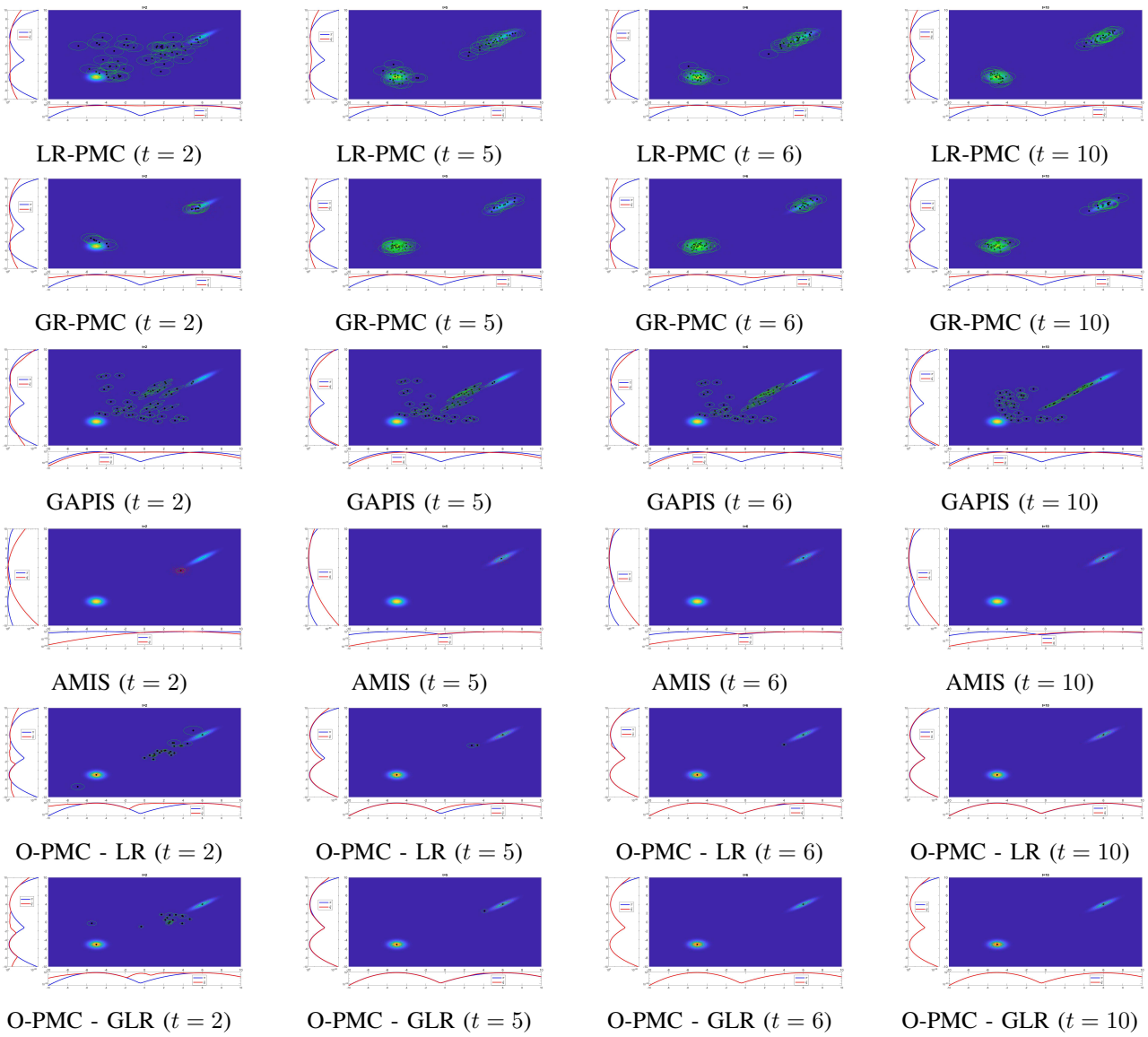

Fig. 1. **Toy example.** Evolution of the reconstructed target along iterations for LR-PMC, GR-PMC, GAPIS, AMIS, and O-PMC for both LR and GLR (with $\Delta = 5$). One can notice the fast convergence of the proposed O-PMC. The great impact of GLR can be seen, by comparing both O-PMC variants (LR / GLR) between time $t = 5$ to time $t = 6$, i.e. after and before applying the GR step in the GLR approach.

posterior with a very thin grid, allowing to compare visually the performance of the different sampling schemes.

Except for AMIS, in all other algorithms we set $N = 50$ proposals (randomly initialized in the square $[-15, 15] \times [-15, 15]$), $T = 20$ iterations, and $K = 20$ samples per proposal and iteration. Since AMIS has a single proposal, we set $N = 1$, $T = 500$ and $K = 40$, i.e., keeping the same number of target evaluations for a fair comparison. For all algorithms we use isotropic Gaussian proposals with standard deviation $\sigma \in \{1, 3, 5\}$, except for O-PMC, where the proposals are initialized using $\mathbf{\Sigma}_n^{(1)} = \sigma^2 \mathbf{I}_2$, with $\sigma = 5$ and then adapted over the iterations. In the GLR version of O-PMC, we set the period $\Delta = 5$. In Table III we display the relative mean square error (RMSE) of the AIS estimators. We build the estimators by averaging all the weighted samples of the second half of the iterations, which allows to better determine the adaptive capabilities of each algorithm. We see that the novel O-PMC outperforms all other algorithms, in most cases by several orders of magnitude.

B. High-dimensional banana-shaped distribution

The second example focuses on the banana-shaped distribution [55], [56]. This target shape has been widely used in the past for assessing the performance of sampling methods, as it is particularly challenging to approximate precisely, especially when the dimension of the problem increases. Let us consider a d_x-dimensional multivariate Gaussian r.v. $\bar{\mathbf{X}} \sim \mathcal{N}(\mathbf{x}; \mathbf{0}_{d_x}, \mathbf{C})$ with $\mathbf{C} = \text{diag}(c^2, 1, ..., 1)$. The banana-shaped distribution is defined as the pdf of the transformed multivariate variable $(X_j)_{1 \leq j \leq d_x}$ such that $X_j = \bar{X}_j$ for $j \in \{1, ..., d_x\} \setminus 2$, and $X_2 = \bar{X}_2 - b(\bar{X}_1^2 - c^2)$. Hereabove, b and c are shape parameters set in the sequel to be equal to $c = 1$ and $b = 3$. We evaluate the performance of different AIS methods in estimating $\mathrm{E}_{\widetilde{\pi}}[\mathbf{X}]$, for different dimensions

	GR-PMC			LR-PMC			GAPIS			AMIS			O-PMC	
	$\sigma=1$	$\sigma=3$	$\sigma=5$	$\sigma=1$	$\sigma=3$	$\sigma=5$	$\sigma=1$	$\sigma=3$	$\sigma=5$	$\sigma=1$	$\sigma=3$	$\sigma=5$	LR	GLR
Z	0.0432	0.0066	0.0073	0.0025	0.0031	0.0161	0.4882	0.0409	0.0481	0.9836	0.9814	0.9487	**4** $\cdot 10^{-4}$	**4** $\cdot 10^{-4}$
$\mathrm{E}_{\tilde{\pi}}[\mathbf{X}]$	2.4280	0.4846	0.3599	0.2229	0.2291	0.6367	2.5397	1.7318	1.2595	54.5381	51.0631	23.4267	**0.03532**	0.03583
$\mathrm{E}_{\tilde{\pi}}[\mathbf{X}^2]$	4.4581	0.4571	0.5014	0.2244	0.2203	0.7778	2.7414	1.4743	2.1444	31.9803	30.1377	21.4783	**0.0426**	0.0434

TABLE III

EXAMPLE V-A. RELATIVE MSE IN THE ESTIMATION OF Z, $\mathrm{E}_{\tilde{\pi}}[\mathbf{X}]$, AND $\mathrm{E}_{\tilde{\pi}}[\mathbf{X}^2]$ IN GM2D EXAMPLE. FOR O-PMC, WE SET THE INITIAL PROPOSAL VARIANCE TO $\sigma = 5$. THE PERIOD FOR GLR IS SET TO $\Delta = 5$. IN ALL PMC-BASED METHODS, $(N, K, T) = (50, 20, 20)$ WHILE $(N, K, T) = (1, 500, 40)$ FOR AMIS.

$d_x \in \{2, 5, 10, 15, 20, 30, 40, 50\}$. All algorithms initialize the location parameters of the proposals randomly and uniformly within the square $[-4, 4] \times [-4, 4]$, and 1000 independent runs are performed. In all algorithms, except AMIS, we set $N = 50$, $K = 20$, and $T = 20$. In AMIS, we set $N = 1$, $K = 500$ and $T = 40$, for a fair comparison in terms of total number of target evaluations (we recall AMIS imposes a unique proposal). In O-PMC, the initial proposal covariances are isotropic, $\mathbf{\Sigma}_n^{(1)} = \sigma^2 \mathbf{I}_{d_x}$, with $\sigma = 3$, and we implement the resampling strategies LR and GLR (with $\Delta = 5$). The other algorithms are initialized also with isotropic covariances with $\sigma \in \{1, 3, 5\}$. In Table IV we show the MSE of the proposed O-PMC and its competitors in the estimation of the target mean for dimensions $d_x \in \{5, 20, 50\}$. We also display in Fig. 2 the performance of O-PMC, LR-PMC and GR-PMC, measured in terms of MSE averaged across dimensions. In this example, the best performance is reached with the LR version of the O-PMC, followed by the GLR version of the same algorithm. AMIS is the second best algorithm in most cases, possibly due to the covariance adaptation that it incorporates (unlike the GR-PMC and LR-PMC schemes). The competitors GR-PMC and LR-PMC degrade when the dimension decreases. Note that the MSE of our O-PMC decreases with the dimension. This can be explained by the particular structure of the target, which is conditionally Gaussian in all dimensions except one, and hence it represents a challenge for O-PMC only in that particular dimension.

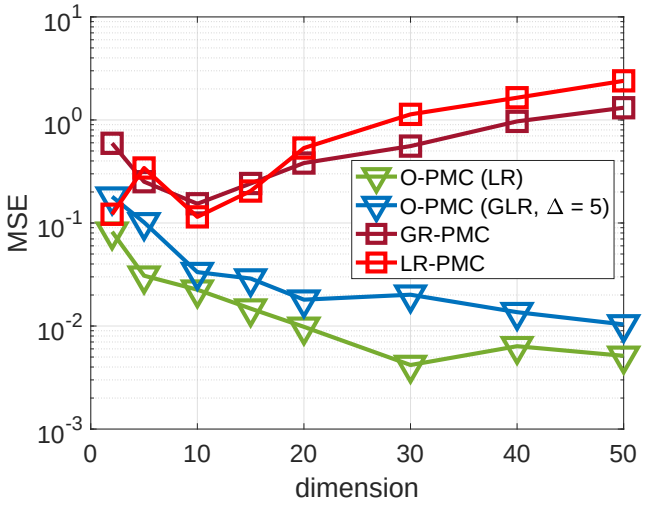

Fig. 2. **Example V-B.** MSE in the estimation of $\mathrm{E}_{\tilde{\pi}}[\mathbf{X}]$ of the banana-shaped distribution versus the dimension d_x, with GR-PMC, LR-PMC with $\sigma = 1$ and the proposed O-PMC method.

C. Spectral analysis

Our last example addresses the problem of estimated the parameters of a multi-sinusoidal signal from noisy and undersampled acquisitions of it. We consider the following observation model:

$$(\forall j \in \{1, \ldots, d_y\}) \quad y_j = \sum_{s=1}^{S} a_s \sin(2\pi\omega_s \tau_j + \varphi_s) + n_j, \tag{20}$$

where $(\tau_j)_{1 \le j \le d_y}$ defines a discrete uniform time grid, $(n_j)_{1 \le j \le d_y}$ a noise assumed to be i.i.d. Gaussian with known variance σ_n^2, and $(a_s, \omega_s, \varphi_s)_{1 \le s \le S}$ the amplitude, frequency and phase parameters, respectively, of S sinusoidal components. We focus on the problem of identifying the unknown frequencies and amplitudes, i.e. $d_x = 2S$, and for all $i \in \{1 \le i \le d_x\}$, $x_i = \omega_s$ and $x_{i+S} = a_s$.

Given the Gaussian model on the noise, the posterior distribution of $\mathbf{x}$ given $\mathbf{y}$ reads $\pi(\mathbf{x}) \propto \exp(-f(\mathbf{x}))$ with

$$f(\mathbf{x}) = \frac{1}{2\sigma_n^2} \sum_{j=1}^{J} \left(y_j - \sum_{s=1}^{S} a_s \sin(2\pi\omega_s \tau_j + \varphi_s) \right)^2 - \log(p_0(\mathbf{x})), \tag{21}$$

with p_0 the prior distribution on $\mathbf{x}$. This prior factorizes as $p_0(\mathbf{x}) = p_\omega(\mathbf{x}_{1:S}) p_a(\mathbf{x}_{S+1:2S})$, where $\mathbf{x}_{1:S}$ contains the first S dimensions of $\mathbf{x}$ (i.e., corresponding to the S unknown frequencies), and $\mathbf{x}_{S+1:2S}$ corresponds to the S unknown amplitudes. The prior p_ω is uniform in the support $\{\mathbf{x}_{1:S} : 0 \le \mathbf{x}_1 \le \mathbf{x}_2 ... \le \mathbf{x}_S \le 0.5\}$, i.e., we restrict the frequencies to be defined in increasing order. The prior p_a factorizes across all dimensions and is a uniform distribution in $[0, +\infty[$. The data is generated by simulating $d_y = 30\,S$ points regularly spaced over $[1, d_y]$. We explore the case with $S \in \{2, 3, 4, 5\}$ (i.e., $d_x \in \{4, 6, 8, 10\}$). We set the observation noise variance to $\sigma_n^2 = 0.5^2$, and the phases $\varphi_s = 0$, for $s = 1, ..., S$.

All algorithms simulate the initial location parameters as the prior p_0, and the initial covariance matrices are chosen to be isotropic with $\sigma \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, except O-PMC , where the initialization is done only with $\sigma = 10^{-2}$ since the covariance is adapted. Table V shows the median squared error (medianSE) in the estimation of the target mean, considering as ground truth the true frequencies and amplitudes that we have set to generate the data. Note that the target mean can be significantly different from those parameters, and for this reason, we also displayed in Table VI the averaged MSE between the signal reconstructed with the estimated parameters w.r.t. the noiseless sequence generated with the true parameters. We observe that with both figures of merit, O-PMC obtains the best results for most setups. While

	GR-PMC			LR-PMC			GAPIS			AMIS			O-PMC	
	$\sigma=1$	$\sigma=3$	$\sigma=5$	$\sigma=1$	$\sigma=3$	$\sigma=5$	$\sigma=1$	$\sigma=3$	$\sigma=5$	$\sigma=1$	$\sigma=3$	$\sigma=5$	LR	GLR
$d_x=5$	0.2515	0.1350	0.2299	0.3418	0.5289	0.5925	0.3007	0.3631	0.7790	0.1758	0.1783	0.1572	**0.0308**	0.1014
$d_x=20$	0.3818	3.1430	11.1921	0.5340	6.4936	23.3693	1.5299	1.6555	1.5640	0.1901	0.1574	0.2673	**0.0098**	0.0180
$d_x=50$	1.3134	9.6571	42.6815	2.3963	21.7097	6.3350	2.5524	2.5632	2.8486	0.6074	0.7992	1.5334	**0.0051**	0.0104

TABLE IV

EXAMPLE V-B. MSE IN THE ESTIMATION OF $\mathrm{E}_{\widetilde{\pi}}[\mathbf{X}]$ OF THE BANANA-SHAPED DISTRIBUTION FOR DIMENSIONS $d_x=5$, 20 AND 50. FOR O-PMC, WE SET THE INITIAL PROPOSAL VARIANCE TO $\sigma=3$. THE PERIOD FOR GLR IS SET TO 5. IN ALL PMC-BASED METHODS, $(N,K,T)=(50,20,20)$ WHILE $(N,K,T)=(1,500,40)$ FOR AMIS.

all methods obtain reasonable results for $S=2$ ($d_x=4$), their performance degrade faster than in O-PMC when the dimension of the problem is increased.

In Fig. 3, we display the evolution with the number of iterations of the medianSE in the target mean estimator for GR-PMC, LR-PMC, and O-PMC (LR, GLR with $\Delta=2$, and GLR with $\Delta=5$) algorithms. At each iteration t, we compute the estimator with all simulated samples from the beginning, re-normalizing the importance weights to build a unique estimator as it is done, for instance, in [57]. We use the same parameters as those of Tables V and VI, setting $\sigma=10^{-2}$. We observe that all algorithms improve when the number of iterations grows, and that all versions of our proposed O-PMC algorithm adapt faster than the competitors. We observe that the GLR version of O-PMC with $\Delta=5$ adapts faster than the case with $\Delta=2$, while the best adaptation for this particular setup is obtained by the LR version of O-PMC.

Finally, Fig. 4 displays the ground truth and the estimators obtained by GR-PMC and LR-PMC (left subplots) and the LR and GLR versions of O-PMC(right subplots). We explore the cases with $S=2$ ($d_x=4$; top subplot), $S=3$ ($d_x=6$; middle subplot), and $S=4$ ($d_x=8$; bottom subplot). The vertical bars represent the median estimate $\pm$ the mean absolute deviation (MAD). We observe that in all dimensions, the O-PMC obtains closer estimates to the ground truth, both in the frequencies and in the amplitudes. We note that when the dimension is increased (bottom subplots), the problem becomes more challenging but O-PMC still performs successfully unlike GR-PMC and LR-PMC.

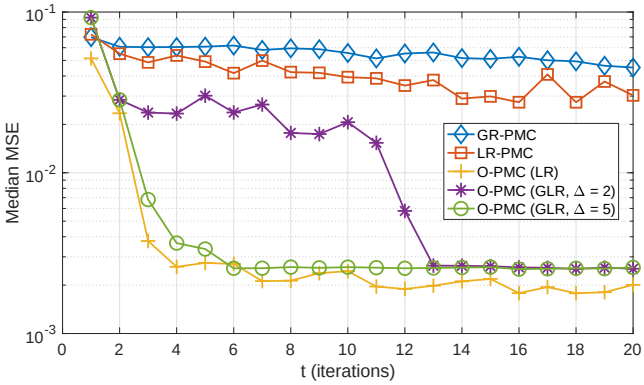

Fig. 3. **Example V-C.** Evolution of the medianSE with respect to ground truth amplitudes and frequencies for dimension $d_x=4$ as function of the number of iterations of GR-PMC, LR-PMC, and O-PMC.

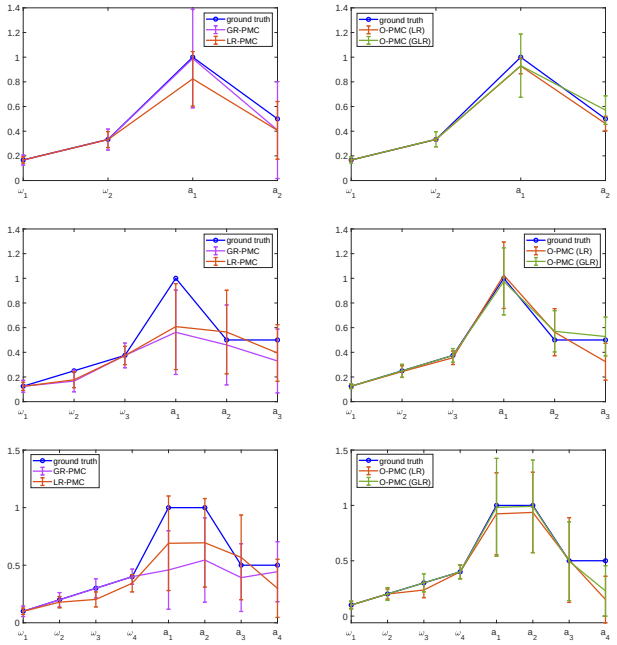

Fig. 4. **Example V-C.** Ground truth (blue) and estimated values (median $\pm$ MAD of the mean estimator) for frequencies and amplitudes in dimension 4 (top), 6 (middle) and 8 (bottom), for GR-PMC, LR-PMC, and O-PMC using either LR or GLR scheme.

VI. CONCLUSION

We have proposed the O-PMC algorithm, an AIS sampler of the family of PMC algorithms that incorporates geometric information of the target distribution. O-PMC exploits the benefits of the PMC framework, incorporates suitable resampling schemes, and includes efficient adaptive mechanisms. In particular, the novel algorithm adapts the location and scale parameters of a set of proposals. At each iteration, the location parameters are adapted through a suitable resampling strategy combined with an advanced optimization-based scheme. The local second-order information of the target is exploited through a preconditioning matrix that acts as a scaling metric onto a gradient direction. This metric is also used in order to adapt the scale parameters of the proposals. We have discussed the choice of parameters, included an illustrative toy example, and evaluated numerically the performance of the novel algorithm in three challenging problems, comparing the results with state-of-the-art competitive methods. As a future work, we may explore the implementation of low-complexity approximations of the Hessian to adapt the scale parameters of the proposals.

	GR-PMC			LR-PMC			GAPIS			AMIS			O-PMC	
	$\sigma = 10^{-3}$	$\sigma = 10^{-2}$	$\sigma = 10^{-1}$	$\sigma = 10^{-3}$	$\sigma = 10^{-2}$	$\sigma = 10^{-1}$	$\sigma = 10^{-3}$	$\sigma = 10^{-2}$	$\sigma = 10^{-1}$	$\sigma = 10^{-3}$	$\sigma = 10^{-2}$	$\sigma = 10^{-1}$	LR	GLR
$d_x = 4$	0.1083	0.0479	0.0249	0.0516	0.0185	0.0299	0.8422	0.4176	0.3342	0.0623	0.0384	0.0504	**0.0017**	0.0024
$d_x = 6$	0.0929	0.0568	0.0544	0.0808	0.0598	0.0621	3.9936	4.9897	4.0805	0.0881	0.0956	0.0806	0.0076	**0.0014**
$d_x = 8$	0.1163	0.0906	0.1022	0.1041	0.0718	0.1128	13.0336	10.4020	7.3938	0.1837	0.1459	0.1261	0.0418	**0.0343**
$d_x = 10$	0.0671	0.0804	0.0671	0.0609	0.0933	0.0757	18.7525	18.5906	14.8404	0.1279	0.1284	0.1811	0.1027	**0.0560**

TABLE V

EXAMPLE V-C. MEDIAN MSE WITH RESPECT TO GROUND TRUTH AMPLITUDES AND FREQUENCIES PARAMETERS, FOR DIMENSIONS $d_x = 4$, 6, 8 AND 10. FOR O-PMC, WE SET THE INITIAL PROPOSAL VARIANCE TO $\sigma = 10^{-2}$. THE PERIOD FOR GLR IS SET TO $\Delta = 5$. IN ALL PMC-BASED METHODS, $(N, K, T) = (50, 20, 20)$ WHILE $(N, K, T) = (1, 500, 40)$ FOR AMIS.

	GR-PMC			LR-PMC			GAPIS			AMIS			O-PMC	
	$\sigma = 10^{-3}$	$\sigma = 10^{-2}$	$\sigma = 10^{-1}$	$\sigma = 10^{-3}$	$\sigma = 10^{-2}$	$\sigma = 10^{-1}$	$\sigma = 10^{-3}$	$\sigma = 10^{-2}$	$\sigma = 10^{-1}$	$\sigma = 10^{-3}$	$\sigma = 10^{-2}$	$\sigma = 10^{-1}$	LR	GLR
$d_x = 4$	0.4206	0.1715	**0.1122**	0.3670	0.2453	0.5379	0.5538	0.3475	0.3810	0.2956	0.2697	0.3185	0.2081	0.3043
$d_x = 6$	0.7301	0.3572	0.2677	0.6632	0.3745	0.4739	0.8761	0.6995	0.6757	0.6480	0.6631	0.6811	0.3859	**0.1681**
$d_x = 8$	1.3138	0.6910	0.8157	1.2692	0.7238	1.0867	1.5086	1.2156	1.5075	3.9607	3.3690	3.2748	0.5733	**0.4259**
$d_x = 10$	2.6353	1.0050	2.8790	2.1628	1.1288	2.8382	1.4591	1.5012	1.5146	4.5971	4.5016	4.7863	1.1124	**0.7351**

TABLE VI

EXAMPLE V-C. RECONSTRUCTED MSE FOR DIMENSIONS $d_x = 4$, 6, 8 AND 10. FOR O-PMC, WE SET THE INITIAL PROPOSAL VARIANCE TO $\sigma = 10^{-2}$. THE PERIOD FOR GLR IS SET TO $\Delta = 5$. IN ALL PMC-BASED METHODS, $(N, K, T) = (50, 20, 20)$ WHILE $(N, K, T) = (1, 500, 40)$ FOR AMIS.

REFERENCES

[1] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer-Verlag New York, 2004.

[2] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag New York, 2004.

[3] A. Owen, *Monte Carlo Theory, Methods and Examples*. http://statweb.stanford.edu/∼owen/mc/, 2013.

[4] V. Elvira and L. Martino, "Advances in importance sampling," *Wiley StatsRef: Statistics Reference Online*, pp. 1–22, 2021.

[5] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Generalized multiple importance sampling," *Statistical Science*, vol. 34, no. 1, pp. 129–155, 2019.

[6] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Míguez, and P. M. Djuric, "Adaptive importance sampling: The past, the present, and the future," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 60–79, 2017.

[7] J. M. Cornuet, J. M. Marin, A. Mira, and C. P. Robert, "Adaptive multiple importance sampling," *Scandinavian Journal of Statistics*, vol. 39, pp. 798–812, December 2012.

[8] J.-M. Marin, P. Pudlo, and M. Sedki, "Consistency of adaptive importance sampling and recycling schemes," *Bernoulli*, vol. 25, no. 3, pp. 1977 – 1998, 2019.

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

[10] L. Martino, V. Elvira, D. Luengo, and J. Corander, "Layered adaptive importance sampling," *Statistics and Computing*, vol. 27, no. 3, pp. 599–623, 2015.

[11] I. Schuster and I. Klebanov, "Markov chain importance sampling - a highly efficient estimator for MCMC," *(to appear) Journal of Computational and Graphical Statistics*, 2021. https://arxiv.org/abs/1805.07179.

[12] D. Rudolf and B. Sprungk, "On a Metropolis–Hastings importance sampling estimator," *Electronic Journal of Statistics*, vol. 14, no. 1, pp. 857–889, 2020.

[13] R. Douc, O. Cappé, and E. Moulines, "Comparison of resampling schemes for particle filtering," in *Proceedings of the 4^{th} International Symposium on Image and Signal Processing and Analysis (ISPA 2005)*, (Zagreb, Croatia), pp. 64–69, 15-17 September 2005.

[14] T. Li, M. Bolic, and P. M. Djuric, "Resampling methods for particle filtering: Classification, implementation, and strategies," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 70–86, 2015.

[15] 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.

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

[17] E. Koblents and J. Miguez, "Robust mixture population Monte Carlo scheme with adaptation of the number of components," in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO 2013)*, (Marrakech, Morocco), pp. 1–5, 9-13 September 2013.

[18] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Improving Population Monte Carlo: Alternative weighting and resampling schemes," *Signal Processing*, vol. 131, no. 12, pp. 77–91, 2017.

[19] V. Elvira, L. Martino, L. Luengo, and J. Corander, "A gradient adaptive population importance sampler," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, (Brisbane, Australia), pp. 4075–4079, 19-24 April 2015.

[20] I. Schuster, "Gradient importance sampling," tech. rep., 2015. https://arxiv.org/abs/1507.05781.

[21] M. Fasiolo, F. E. de Melo, and S. Maskell, "Langevin incremental mixture importance sampling," *Statistical Computing*, vol. 28, no. 3, pp. 549–561, 2018.

[22] G. O. Roberts and L. R. Tweedie, "Exponential convergence of Langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, pp. 341–363, Dec. 1996.

[23] Y. El-Laham, V. Elvira, and M. F. Bugallo, "Robust covariance adaptation in adaptive importance sampling," *IEEE Signal Processing Letters*, 2018.

[24] Y. El-Laham, V. Elvira, and M. Bugallo, "Recursive shrinkage covariance learning in adaptive importance sampling," in *Proceedings of the 8th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP 2019)*, (Guadeloupe, France), pp. 624–628, 15-18 December 2019.

[25] M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. Hero, and S. McLaughlin, "A survey of stochastic simulation and optimization methods in signal processing," *IEEE Journal on Selected Topics in Signal Processing*, vol. 10, pp. 224–241, March 2016.

[26] G. O. Roberts and O. Stramer, "Langevin diffusions and Metropolis-Hastings algorithms," *Methodology and Computing in Applied Probability*, vol. 4, no. 4, pp. 337–357, 2002.

[27] A. Durmus, E. Moulines, and M. Pereyra, "Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau," *SIAM Journal on Imaging Sciences*, vol. 11, no. 1, pp. 473–506, 2018.

[28] A. Schreck, G. Fort, S. L. Corff, and E. Moulines, "A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection," *IEEE Journal on Selected Topics in Signal Processing*, vol. 10, pp. 366–375, March 2016.

[29] Y. Qi and T. P. Minka, "Hessian-based Markov Chain Monte-Carlo algorithms," *Proceedings of the First Cape Cod Workshop on Monte Carlo Methods*, 2002. https://www.microsoft.com/en-us/research/publication/hessian-based-markov-chain-monte-carlo-algorithms/.

[30] C. Vacar, J.-F. Giovannelli, and Y. Berthoumieu, "Langevin and Hessian with Fisher approximation stochastic sampling for parameter estimation of structured covariance," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, (Prague, Czech Republic), pp. 3964–3967, 22-27 May 2011.

[31] M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," *Journal of the Royal Statistical*

Society Series B (Statistical Methodology), vol. 73, pp. 123–214, March 2011.

[32] J. Martin, C. L. Wilcox, C. Burstedde, and O. Ghattas, "A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion," *SIAM Journal on Scientific Computing*, vol. 34, pp. 1460–1487, January 2012.

[33] Y. Zhang and C. A. Sutton, "Quasi-Newton methods for Markov chain Monte Carlo," in *Proceedings of the Neural Information Processing Systems workshop (NIPS 2011)*, no. 24, (Granada, Spain), pp. 2393–2401, 12-17 December 2011.

[34] Y. Marnissi, E. Chouzenoux, A. Benazza-Benyahia, and J.-C. Pesquet, "Majorize-Minimize adapted Metropolis-Hastings algorithm," *IEEE Transactions on Signal Processing*, pp. 2356–2369, March 2020.

[35] Y. Marnissi, A. Benazza-Benyahia, E. Chouzenoux, and J.-C. Pesquet, "Majorize-Minimize adapted Metropolis Hastings algorithm. application to multichannel image recovery," in *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO 2014)*, (Lisboa, Portugal), pp. 1332–1336, 1-5 September 2014.

[36] Y. Marnissi, E. Chouzenoux, A. Benazza-Benyahia, and J.-C. Pesquet, "An auxiliary variable method for MCMC algorithms in high dimension," *Entropy*, vol. 20, no. 110, 2018.

[37] U. Simsekli, R. Badeau, A. T. Cemgil, and G. Richard, "Stochastic quasi-Newton Langevin Monte Carlo," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML 2016)*, vol. 48, p. 642–651, 2016.

[38] M. Vono, N. Dobigeon, and P. Chainais, "Split-and-augmented Gibbs sampler - Application to large-scale inference problems," *IEEE Transactions on Signal Processing*, vol. 67, pp. 1648–1661, March 2019.

[39] V. Elvira and E. Chouzenoux, "Langevin-based strategy for efficient proposal adaptation in population Monte Carlo," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2019)*, (Brighton, UK), pp. 5077–5081, 12-17 May 2019.

[40] P. Langevin, "On the theory of Brownian motion," *Comptes Rendus de l'Académie des Sciences*, vol. 146, pp. 530–533, 1908.

[41] D. Talay and L. Tubaro, "Expansion of the global error for numerical schemes solving stochastic differential equations," *Stochastic Analysis and Applications*, vol. 8, no. 4, pp. 483–509, 1991.

[42] A. Durmus and E. Moulines, "High-dimensional Bayesian inference via the unadjusted Langevin algorithm," *Bernoulli*, no. 4A, pp. 2854–2882, 2019.

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

[43] A. Lau and T. Lubensky, "State-dependent diffusion: Thermodynamic consistency and its path integral formulation," *Physical Review E*, no. 76, p. 011123, 2007.

[44] T. Xifara, C. Sherlock, S. Livingstone, S. Byrne, and M. Girolami, "Langevin diffusions and the Metropolis-adjusted Langevin algorithm," *Statistics and Probability Letters*, vol. 91, pp. 14 – 19, 2014.

[45] S. Sabanis and Y. Zhang, "Higher order Langevin Monte Carlo algorithm," *Electronic Journal of Statistics*, vol. 13, no. 2, pp. 3805–3850, 2019.

[46] S. Särkkä, *Bayesian filtering and smoothing*, vol. 3. Cambridge University Press, 2013.

[47] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Population Monte Carlo schemes with reduced path degeneracy," in *Proceedings of the IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP 2017)*, (Curacao, Dutch Antilles), pp. 1–5, 10-13 December 2017.

[48] J. Nocedal and S. Wright, *Numerical Optimization*. Springer, New York, NY, 2nd ed., 2006.

[49] O. D. Akyildiz and J. Míguez, "Convergence rates for optimised adaptive importance samplers," *Statistics and Computing*, vol. 31, no. 2, pp. 1–17, 2021.

[50] E. Veach and L. Guibas, "Optimally combining sampling techniques for Monte Carlo rendering," in *Proceedings of the 22nd International ACM Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1995)*, pp. 419–428, September 1995.

[51] A. Owen and Y. Zhou, "Safe and effective importance sampling," *Journal of the American Statistical Association*, vol. 95, no. 449, pp. 135–143, 2000.

[52] P. Del Moral, A. Doucet, and A. Jasra, "On adaptive resampling strategies for sequential Monte Carlo methods," *Bernoulli*, vol. 18, no. 1, pp. 252–278, 2012.

[54] J. Fernandez-Bes, V. Elvira, and S. Van Vaerenbergh, "A probabilistic least-mean-squares filter," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2199–2203, IEEE, 2015.

[55] H. Haario, E. Saksman, and J. Tamminen, "Adaptive proposal distribution for random walk Metropolis algorithm," *Computational Statistics*, vol. 14, no. 3, pp. 375–396, 1999.

[56] H. Haario, E. Saksman, and J. Tamminen, "An adaptive Metropolis algorithm," *Bernoulli*, vol. 7, pp. 223–242, April 2001.

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