Gradient-based Adaptive Importance Samplers

Víctor Elvira, Emilie Chouzenoux, Ömer Deniz Akyildiz, Luca Martino

Gradient-based Adaptive Importance Samplers

Víctor Elvira[a,*], Émilie Chouzenoux[b], Ömer Deniz Akyildiz[c], Luca Martino[d]

[a] *School of Mathematics, University of Edinburgh, UK.*
[b] *CVN, INRIA Saclay, CentraleSupélec, France.*
[c] *Dept. of Mathematics, Imperial College London, UK.*
[d] *Universidad Rey Juan Carlos, Spain.*

Abstract

Importance sampling (IS) is a powerful Monte Carlo methodology for the approximation of intractable integrals, very often involving a target probability density function. The performance of IS heavily depends on the appropriate selection of the proposal distributions where the samples are simulated from. In this paper, we propose an adaptive importance sampler, called GRAMIS, that iteratively improves the set of proposals. The algorithm exploits geometric information of the target to adapt the location and scale parameters of those proposals. Moreover, in order to allow for a cooperative adaptation, a repulsion term is introduced that favors a coordinated exploration of the state space. This translates into a more diverse exploration and a better approximation of the target via the mixture of proposals. Moreover, we provide a theoretical justification of the repulsion term. We show the good performance of GRAMIS in two problems where the target has a challenging shape and cannot be easily approximated by a standard uni-modal proposal.

Keywords: Adaptive importance sampling, Monte Carlo, Bayesian inference, Langevin adaptation, Poisson field, Gaussian mixture

1. Introduction

The approximation of intractable integrals is a common statistical task in many applications of science and engineering. A relevant example is the case of Bayesian

*Corresponding author
Email address: victor.elvira@ed.ac.uk (Víctor Elvira)

inference arising for instance in statistical machine learning. A posterior distribution of unknown parameters is constructed by combining data (through a likelihood model) and previous information (through the prior distribution). Unfortunately, the posterior is often unavailable, typically due to the intractability of the marginal likelihood.

Monte Carlo methods have proved to be effective in those relevant problems, producing a set of random samples that can be used to approximate a target probability distribution and integrals related to it [1, 2, 3]. Importance sampling (IS) is one of the main Monte Carlo families, with solid theoretical guarantees [3, 4]. The vanilla version of IS simulates samples from the so-called proposal distribution. Then, each sample receives an associated weight which is computed by taking into account the mismatch between this proposal probability density function (pdf) and the target pdf. While the IS mechanism is valid under very mild assumptions [3, 4], the efficiency of the estimators is driven by the choice of the proposal distribution. Unfortunately this choice is a difficult task, even more in contexts where one has access to the evaluation of an unnormalized version of the posterior distribution, as it is the case in Bayesian inference. The two main methodological directions to overcome the limitations in IS are the usage of several proposals, which is known as multiple IS (MIS) [5], and the iterative adaptation of those proposals, known as adaptive importance sampling (AIS) [6].

There exist a plethora of AIS algorithms, and we refer the interested reader to [6]. There are also strong theoretical guarantees for some subclasses of AIS algorithms, see, e.g., [7, 8, 9]. These AIS methods can be arguably divided in three main categories. The first category is based on sequential moment matching and includes algorithms that implement Rao-Blackwellization in the temporal estimators ([10, 11]), extend to the MIS setting [12, 13], or are based on a sequence of transformations that can be interpreted as a change of proposal [14]. A second AIS family comprises the population Monte Carlo (PMC) methods which use resampling mechanisms to adapt the proposals [15, 16]. The PMC framework was first introduced in [17] and then extended by the incorporation of stochastic expectation-maximization mechanisms [18], clipping of the importance weights [19], improving the weighting and resampling mechanisms [20, 21], targeting the estimation of rare-event probabilities [22], or introducing opti-

mization schemes [23].

Finally, a third category contains AIS methods with a hierarchical or layered structure. Examples of these algorithms are those that adapt the location parameters of the proposals using a Markov chain Monte Carlo (MCMC) mechanism [24, 25, 26, 27]. In this category, we also include methods that exploit geometric information about the target for the adaptation of the location parameters, yielding to optimization-based adaptation schemes. In the layered mechanism, the past samples do not affect the proposal adaptation which is rather driven by the geometric properties of the targeted density. However, there also exists hybrid mechanisms, e.g., the O-PMC framework which implements resampling and also incorporates geometric information [23].

1.1. Contribution within the state of the art

In this paper, we propose the gradient-based adaptive multiple importance sampling (GRAMIS) method, which falls within the layered family of AIS algorithms. Its main feature is the exploitation of geometric information about the target by incorporating an optimization approach. It has been shown that geometric-based optimization mechanisms improve the convergence rate in MCMC algorithms (see the review in [28]) and in AIS methods (e.g., [23]). In the context of MCMC, the methods in [29, 30, 31] are called Metropolis adjusted Langevin algorithms (MALA). The Langevin-based adaptation included in their MCMC proposal updates reads as a noisy gradient descent (called drift term) that favors the exploration of the target in areas with larger probability mass, resulting in a larger acceptance probability in the Metropolis-Hastings (MH) step. Preconditioning can be added for a further improvement of the exploration. To do so, local curvature information about the target density is used to build a matrix scaling the drift term. Fisher metric [32], Hessian matrix [33, 34, 35], or a tractable approximation of it [36, 37, 38, 39] have been considered for that purpose. Within AIS, the algorithms in [40, 41] adapt the location parameters via several steps of the unadjusted Langevin algorithm (ULA) [42].

A limitation that is present in most AIS algorithms is the lack of adaptation of the scale parameters of the proposals, e.g., the covariance matrices in case of Gaussian proposals. However, suitable scale parameters are essential for an adequate perfor-

mance of the AIS algorithms, and the alternative to their adaptation is setting them a priori, which is in general a difficult task. The inefficiency of AIS algorithms that do not implement adaptation in the scale parameters is particularly damaging in high dimensions and where the target presents strong correlations that are unknown a priori. A covariance adaptation has been explored via robust moment-matching mechanisms in [43, 44], second-order information in [41, 23], and sample autocorrelation [40]. The proposed GRAMIS algorithm implements an adaptation of the covariance by using second-order information (when it yields to a definite positive matrix). In particular, the covariance adaptation of each proposal is adapted by using the Hessian of the logarithm of the target, evaluated at the updated location parameter. The second-order information is also used to pre-condition the gradient in the adaptation of the location parameters.

Another limitation in AIS algorithms is the lack of cooperation (or insufficient cooperation) between the multiple proposals at the adaptation stage. Some of the few algorithms that implement a type of cooperation can be found in [18] through a probabilistic clustering of all samples, and in [20] through a resampling that use the deterministic mixture weights. In the paper, we implement an explicit repulsion between the proposals during the adaptation stage in order to improve the exploration of the algorithm.

Finally, GRAMIS implements the balance-heuristic (also called deterministic mixture) importance weights [45, 46], which have been shown a theoretical superiority (in the unnormalized IS estimators) [5] and a superior performance in other types of AIS algorithms (e.g., in [13, 20, 23].

1.2. Structure of the paper

The rest of the paper is structured as follows. Section 2 introduces the problem and relevant background. In Section 3, we describe the GRAMIS algorithm. We provide numerical examples in Section 4. Section 5 closes the paper with some conclusion.

2. Background in importance sampling

Let us consider the posterior distribution

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

where

- $\mathbf{x} \in \mathbb{R}^{d_x}$ is the variable associated to the r.v. $\mathbf{X}$ of the vector of unknowns to be estimated;
- $\mathbf{y} \in \mathbb{R}^{d_y}$ represents the available data;
- $\ell(\mathbf{y}|\mathbf{x})$ is the likelihood function; and
- $p_0(\mathbf{x})$ is the prior distribution.

We consider the problem where one must compute the integral

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

where h is any integrable function w.r.t. $\widetilde{\pi}(\mathbf{x})$. Such problem can arise for instance in the field of Bayesian learning, when $\mathbf{y}$ gathers the available data to train a model described by vector $\mathbf{x}$ [47, 48].

In most cases, (2) is intractable, either because one does not have access to the pdf (typically the marginal likelihood, $Z(\mathbf{y}) \triangleq \int \pi(\mathbf{x}|\mathbf{y})d\mathbf{x}$, is intractable in Bayesian inference) or because the integral cannot be computed analytically. Thus, in many cases one has access only to the non-negative function $\pi(\mathbf{x}|\mathbf{y}) \triangleq \ell(\mathbf{y}|\mathbf{x})p_0(\mathbf{x}) = Z(\mathbf{y})\widetilde{\pi}(\mathbf{x}|\mathbf{y})$. In the rest of the paper, we consider the data $\mathbf{y}$ to be fixed, and we alleviate the notation by denoting Z, $\pi(\mathbf{x})$, and $\widetilde{\pi}(\mathbf{x})$.

2.1. Importance sampling

Importance sampling (IS) is one of the main Monte Carlo methodologies for the approximation of distributions and related integrals. In IS, K samples $\mathbf{x}_k$ are simulated from an alternative distribution $q(\mathbf{x})$, called proposal, and receive and importance weight w_k. The procedure comprises two basic steps:

1. **Sampling.** K samples are drawn as

$$\mathbf{x}_k \sim q(\mathbf{x}), \qquad k = 1, ..., K.$$

2. **Weighting.** Each sample is associated to an importance sampling

$$w_k = \frac{\pi(\mathbf{x}_k)}{q(\mathbf{x}_k)}, \qquad k = 1, ..., K.$$

The resulting sets of samples $\{\mathbf{x}_k\}_{k=1}^K$ and weights $\{w_k\}_{k=1}^K$ are used in order to produce estimators that approximate I in Eq. (2). When Z is available, it is possible to produce the unnormalized IS (UIS) estimator

$$\widehat{I} = \frac{1}{KZ} \sum_{k=1}^{K} w_k h(\mathbf{x}_k). \tag{3}$$

If Z is not available, the alternative is to use the self-normalized IS (SNIS) estimator,

$$\widetilde{I} = \sum_{k=1}^{K} \overline{w}_k h(\mathbf{x}_k), \tag{4}$$

where $\overline{w}_k = w_k / \sum_{j=1}^K w_j$, $k = 1, ..., K$ are the importance weights.

The UIS estimator is unbiased and consistent. The SNIS estimator is consistent and has a bias which vanishes at a faster rate than the variance when K grows. The optimal proposal of the UIS estimator is $q(\mathbf{x}) \propto |h(\mathbf{x})|\pi(\mathbf{x})$ [1, 2], while the optimal proposal of the SNIS estimator is (approximately) $q(\mathbf{x}) \propto |h(\mathbf{x})|\pi(\mathbf{x})$ [3].

2.2. Multiple importance sampling

One of the most common strategies is to use several proposals, $\{q_n(\mathbf{x})\}_{n=1}^N$ [49, 46, 3]. The last years have witnessed and increased of attention in MIS [50, 51, 52, 53, 54] (see a generic framework with theoretical analysis in [5]). It has been shown that several weighting and sampling schemes are possible, i.e., that lead to consistent UIS and SNIS estimators [5]. We consider the simplified example where we simulate $K = N$ samples from the set of proposals. Then, one possibility is to simulate exactly one sample per proposal as $\mathbf{x}_n \sim q_n(\mathbf{x})$, $n = 1, ..., N$. Then, two popular weighting approaches are the standard MIS (s-MIS),

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

and the deterministic mixture MIS (DM-MIS),

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

In [5], it is proved that DM-MIS weights provide an UIS estimator with less variance compared to s-MIS, for any function h, any target $\widetilde{\pi}$, and any set of proposals.

2.3. Optimization-based samplers

The performance of sampling algorithms depends greatly on the choice of the proposal distribution. Proposals parametrized with static parameters are easier to implement and manipulate, as they require minimal self tuning, but this simplicity comes at the price of a suboptimal (since not adaptative) target exploration. To cope with this issue, several methods have been proposed to iteratively update the proposal, along the sampling algorithm iterations, so as to improve and accelerate the target exploration. The most common technique is to resort to a Langevin-based approach, where gradient descent steps (assuming differentiability of $\log \pi$) are performed to adapt the proposal mean (i.e., location). The discretization of the Langevin dynamics leads to the unadjusted Langevin algorithm (ULA) [42], which can also be viewed as a gradient descent algorithm perturbed with an independent and identically distributed (i.i.d.) stochastic error. The convergence of ULA is discussed in [55, 42]. However, in most situations, the stationary distribution of the samples produced by ULA differs from the target π [56], due to the discretization of the Langevin dynamic. MALA [29] tackles this issue by introducing an MH strategy, hence guaranteeing ergodic convergence to the sought target law. Accelerated variants of MALA have been investigated, based on preconditioning techniques to account for more information (e.g., curvature) about the target [36, 32, 57, 35, 58, 29, 59]. For instance, the Newton MH strategy [57, 35] consists in combining an MH procedure with a stochastic Newton update involving the inverse (or an approximation of it, when undefined or too complex) of the Hessian matrix of $\log \pi(\mathbf{x})$. This approach will serve as starting point for introducing a proposal adaptation within our novel approach GRAMIS.

3. The GRAMIS algorithm

We now describe GRAMIS, the proposed AIS algorithm, in Table 1. The algorithm runs over T iterations, adapting N proposals, and simulating K sampler per proposal and iteration.

First, the location parameters are updated in (9) following a gradient step that includes an optimized stepsize $\theta_n^{(t-1)}$, and first and second-order information of the log-target at the previous location parameter $\boldsymbol{\mu}_n^{(t-1)}$ of each proposal (see Section 3.1 for more details). A repulsion term between each pair of proposals (i.e., between j-th and i-th proposals), $\mathbf{r}_{i,j}^{(t-1)}$, is introduced. This repulsion force which inversely proportional to the (Euclidean) distance $||\mathbf{d}_{i,j}|| = ||\boldsymbol{\mu}_i^{(t-1)} - \boldsymbol{\mu}_j^{(t-1)}||$. More practical information about the choice of repulsion strategy can be found in Section 3.3. Second, the scale parameters are updated by using second-order information of the log-target (see Section 3.2). Third, K samples are simulated from each proposal. The importance weights are computed in (11). Note that we implement the DM-MIS version, which presents an advantage as discussed in Section 2.2. GRAMIS returns KNT weighted samples that can be used to estimate both I and Z (in the case it is unknown). The simplest version of those estimators is given below.

- UIS estimator:
$$\widehat{I} = \frac{1}{KTNZ} \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{k=1}^{K} w_{n,k}^{(t)} h(\mathbf{x}_{i,k}^{(t)}). \tag{5}$$

- SNIS estimator:
$$\widetilde{I} = \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{k=1}^{K} \widetilde{w}_{n,k}^{(t)} h(\mathbf{x}_{i,k}^{(t)}), \tag{6}$$
where
$$\widetilde{w}_{n,k}^{(t)} = \frac{w_{n,k}^{(t)} h(\mathbf{x}_{i,k}^{(t)})}{\sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{k=1}^{K} w_{n,k}^{(t)}} \tag{7}$$
are the re-normalized weights.

- estimator of Z:
$$\widehat{Z} = \frac{1}{KTNZ} \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{k=1}^{K} w_{n,k}^{(t)}. \tag{8}$$

Table 1: GRAMIS.

1. **[Initialization]**: Initialize the proposal means $\boldsymbol{\mu}_n^{(0)}$ and the non-adapted parameters $\boldsymbol{\nu}_n$. Compute the scale parameter matrix $\boldsymbol{\Sigma}_n^{(0)}$ using (13).

2. **[For $t = 1$ to T]**:

   (a) **Mean adaptation:**

      i. Compute the stepsize $\theta_n^{(t-1)}$ using the backtracking procedure so as to satisfy (12).

      ii. The mean of the n-th proposal is adapted as

$$\boldsymbol{\mu}_n^{(t)} = \boldsymbol{\mu}_n^{(t-1)} + \theta_n^{(t-1)} \boldsymbol{\Sigma}_n^{(t-1)} \nabla \log \left(\pi(\boldsymbol{\mu}_n^{(t-1)}) \right) + \sum_{j=1, j\neq n}^{N} \mathbf{r}_{n,j}^{(t-1)}, \tag{9}$$

      with

$$\mathbf{r}_{n,j}^{(t-1)} = G_t \frac{m_n m_j}{\|\mathbf{d}_{n,j}^{(t-1)}\|^{d_x}} \mathbf{d}_{n,j}^{(t-1)}, \tag{10}$$

      where $\|\cdot\|$ represents the norm operator, $\mathbf{d}_{n,j}^{(t-1)} = \boldsymbol{\mu}_n^{(t-1)} - \boldsymbol{\mu}_j^{(t-1)}$, and $m_n, m_j > 0$ are two positive terms that depend on the n-th and j-th proposals respectively.

   (b) **Covariance adaptation:** The covariance matrix of the n-th proposal $\boldsymbol{\Sigma}_n^{(t)}$ is adapted using (13).

   (c) **Sampling steps:**

      i. Draw K independent samples from each proposal, i.e., $\mathbf{x}_{n,k}^{(t)} \sim q_n^{(t)}(\mathbf{x}; \boldsymbol{\mu}_n^{(t)}, \boldsymbol{\Sigma}_n^{(t)}, \boldsymbol{\nu}_n)$ for $k = 1 \ldots, K$ and $n = 1, \ldots, N$.

      ii. Compute the importance weights,

$$w_{n,k}^{(t)} = \frac{\pi(\mathbf{x}_{n,k}^{(t)})}{\frac{1}{N}\sum_{j=1}^{N} q_j^{(t)}(\mathbf{x}_{n,k}^{(t)}; \boldsymbol{\mu}_n^{(t)}, \boldsymbol{\Sigma}_n^{(t)})}, \tag{11}$$

      for $n = 1, \ldots, N$, and $k = 1, \ldots, K$.

3. **[Output]**: Return the pairs $\{\mathbf{x}_{n,k}^{(t)}, w_{n,k}^{(t)}\}$, for $n = 1, \ldots, N$, $k = 1, \ldots, K$, and $t = 1, \ldots, T$.

3.1. Adaptation of location parameters

The location parameters are adapted as in Eq. (9). The adaptation process implements a Newton ascent on $\log \pi$. The gradient of the log target is evaluated at the previous location parameter and pre-conditioned by $\boldsymbol{\Sigma}_n^{(t-1)}$ where $\boldsymbol{\Sigma}_n^{(t-1)}$ is the same that we use in the previous section in order to update the covariance. We furthermore introduced $\theta_n^{(t-1)} \in (0, 1]$, 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, for non log-concave distributions. Starting with unit stepsize value, we reduce it by factor $\tau = 1/2$ until the condition below is met:

$$\pi\left(\boldsymbol{\mu}_n^{(t-1)} + \theta_n^{(t-1)}\boldsymbol{\Sigma}_n^{(t-1)}\nabla \log\left(\pi(\boldsymbol{\mu}_n^{(t-1)})\right)\right) \geq \pi\left(\boldsymbol{\mu}_n^{(t-1)}\right). \tag{12}$$

The update in Eq. (9) also incorporates an innovative repulsion term among proposals. The purpose is to efficiently explore the space in a cooperative manner. This repulsion term admits several interpretations. It can be seen as an over-spreading of the mixture proposal, i.e., a safer choice of mixture that will overweight the tails of the target [45]. Also, it can be interpreted as a negative coupling among proposals. It shares connections with MCMC algorithms that implement interacting parallel chains, with a similar spirit as it is done in MCMC [60, 61]. In Section 3.3, we discuss the practical repulsion schemes, and the rationale of the adaptation is discussed in Section 3.4.

3.2. Adaptation of scale parameters

We implement a Newton-based strategy to exploit the the Hessian of $\log \pi$ in the update of the scale parameter. In general scenarios, the convexity of $-\log \pi$ is not ensured, and numerical issues might arise when computing the inverse of its Hessian. We thus propose to introduce a safe rule in our adaptation method, so that

$$\boldsymbol{\Sigma}_n^{(t)} = \begin{cases} \left(-\nabla^2 \log \pi(\boldsymbol{\mu}_n^{(t)})\right)^{-1}, & \text{if } \nabla^2 \log \pi(\boldsymbol{\mu}_n^{(t)}) \succ 0, \\ \boldsymbol{\Sigma}_n^{(t-1)}, & \text{otherwise.} \end{cases} \tag{13}$$

The scaling matrix thus incorporates second order information on the target, whenever this yields to a definite positive matrix. Otherwise, it inherits the covariance of the proposal of the previous iteration, where $\boldsymbol{\Sigma}_n^{(0)}$ is set to a predefined default value (typically, a scalar times the identity matrix).

3.3. Design of the repulsion scheme

Our repulsion term is parameterized by a common time-dependent constant G_t and a proposal-dependent constant, m_n, for each $n = 1, \ldots, N$. By construction, Eq. (10) implies that the repulsion term vanishes whenever the proposals get further away (in Euclidean distance).

The interpretation of the functional form in Eq. (10) is discussed in the next section. The simpler choice is to keep $m_n = 1$, for all $n = 1, \ldots, N$, and to fix the common term G_t to be constant over the iterations, i.e., $G_t = G$. In this case, the repulsion never vanishes with the consequence of leading to a potential equilibrium positioning of the proposals in such a way that the interpreted mixture proposal, $\widetilde{q}^{(t)}(\mathbf{x}) \triangleq \frac{1}{N}\sum_{n=1}^{N} q_n^{(t)}(\mathbf{x}; \boldsymbol{\mu}_n^{(t)}, \boldsymbol{\Sigma}_n^{(t)}, \boldsymbol{\nu}_n)$ would overweight the tails of the target distribution. In this case, it is not guaranteed that the proposal adaptation converges in finite t. An alternative is to reduce the repulsion term in such a way that $r_{n,j}^{(t)} \to 0$ when $t \to \infty$. A natural choice is a decaying term in the form of

$$G_t = \exp(-\beta t), \qquad \beta > 0. \tag{14}$$

In such case, if the Newton scheme converges to a local maximum for each proposal, the whole mixture approximation would converge to a mixture of local Laplace approximations. The choice of β can be easily set depending on the repulsion strength desired in the last iteration, e.g., a 1% of attenuation in the last iteration leads to $\beta = \frac{-\log(0.01)}{T-1}$. It is also possible to set the repulsion term to zero in the last iteration, so a final set of samples can be simulated.

3.4. Repulsion term interpretation

We can interpret the repulsion term of Eq. (10) in general physical terms in $\mathbb{R}^{d_x}$. The following discussion uses the particle interpretation of the repulsion term mentioned in [62], formalizing it using the notion of Poisson fields [63]. For ease of presentation, we consider a simplified version of the update defined in (9). In particular, we consider a fixed, scalar step-size $\theta_n^{(t-1)} = \gamma$ for all $n = 1, \ldots, N$ and $t = 1, \ldots, T$ and $\boldsymbol{\Sigma}_n^{(t-1)} = I_{d_x}$, where I_{d_x} is an identity matrix. For the repulsion term, we also set $G_t = \gamma/N$ for all $t = 1, \ldots, T$ and $m_n = 1$ for all $n = 1, \ldots, N$. By using the scaled step-size γ/N as the repulsion term G_t, we recover a unified gradient descent formulation which clarifies the role of the repulsion term:

$$\boldsymbol{\mu}_n^{(t)} = \boldsymbol{\mu}_n^{(t-1)} + \gamma \nabla \log\left(\pi(\boldsymbol{\mu}_n^{(t-1)})\right) + \frac{\gamma}{N} \sum_{j=1, j\neq n}^{N} \mathbf{r}_{n,j}^{(t-1)}, \tag{15}$$

where $\gamma > 0$ is a scalar step-size and the repulsion term is $G_t = \gamma/N$ as mentioned above, and

$$\mathbf{r}_{n,j}^{(t-1)} = \frac{1}{S_{d_x-1}(1)\|\mathbf{d}_{n,j}^{(t-1)}\|^{d_x}}\mathbf{d}_{n,j}^{(t-1)}, \tag{16}$$

with $S_{d_x-1}(1)$ being a constant equals to surface area of unit (d_x-1) sphere. Thus,

$$S_{d_x-1}(1) = \frac{2\pi^{d_x/2}}{\Gamma(d_x/2)},$$

where Γ is the Gamma function and $\mathbf{d}_{n,j}^{(t-1)} = \boldsymbol{\mu}_n^{(t-1)} - \boldsymbol{\mu}_j^{(t-1)}$. Our first aim is to understand the last term in (15) as an empirical estimate of an integral. For this, consider the empirical measure constructed by the sequence $\{\boldsymbol{\mu}_n^{(t-1)}\}_{n=1}^N$ given by

$$\rho_{t-1,n}^N(\mathrm{d}\mathbf{u}) = \frac{1}{N}\sum_{j=1,j\neq n}^{N} \delta_{\boldsymbol{\mu}_j^{(t-1)}}(\mathrm{d}\mathbf{u}),$$

which is defined for the update of $\boldsymbol{\mu}_n^{(t-1)}$. Using this empirical measure, we can interpret the repulsion term in (15) as a discretised version of a particular integral, i.e.,

$$\boldsymbol{\mu}_n^{(t)} = \boldsymbol{\mu}_n^{(t-1)} + \gamma\nabla\log\left(\pi(\boldsymbol{\mu}_n^{(t-1)})\right) + \gamma\int g(\boldsymbol{\mu}_n^{(t-1)},\mathbf{v})\rho_{t-1,n}^N(\mathrm{d}\mathbf{v}), \tag{17}$$

with

$$g(\mathbf{u},\mathbf{v}) = \frac{1}{S_{d_x-1}(1)\|\mathbf{u}-\mathbf{v}\|^{d_x}}(\mathbf{u}-\mathbf{v}). \tag{18}$$

This implies that, if we set $g(\mathbf{u},\mathbf{v}) = -\nabla_{\mathbf{u}}G(\mathbf{u},\mathbf{v})$, where

$$G(\mathbf{u},\mathbf{v}) = \frac{1}{(d_x-2)S_{d_x-1}(1)}\frac{1}{\|\mathbf{u}-\mathbf{v}\|^{d_x-2}},$$

then the last term in (17) can be interpreted as a gradient [63], i.e.,

$$-\nabla\varphi_{t-1}^N(\mathbf{u}) = -\int \nabla_u G(\mathbf{u},\mathbf{v})\rho_{t-1,n}^N(\mathrm{d}\mathbf{v}),$$

where $G(\mathbf{u},\mathbf{v})$ is the extension of Green's function in the d_x-dimensional space [63]. We can then rewrite (15) as

$$\boldsymbol{\mu}_n^{(t)} = \boldsymbol{\mu}_n^{(t-1)} + \gamma\nabla\log\left(\pi(\boldsymbol{\mu}_n^{(t-1)})\right) - \gamma\nabla\varphi_{t-1}^N(\boldsymbol{\mu}_n^{(t-1)}), \tag{19}$$

where

$$-\nabla\varphi^N(\boldsymbol{\mu}_n^{(t-1)}) = \frac{1}{N}\sum_{j=1,j\neq n}^{N}\frac{\Gamma(d_x/2)}{2\pi^{d_x/2}}\frac{\mathbf{d}_{n,j}^{(t-1)}}{\|\mathbf{d}_{n,j}^{(t-1)}\|^{d_x}}.$$

The term $-\nabla\varphi^N(\mathbf{u})$, named *the Poisson field* [63], pushes particles away from *sources*.

Eq. (19) describes the precise balance achieved by the repulsion term. The term $\nabla \log \pi(\cdot)$ in (19) pushes the means towards the maximum-a-posteriori (MAP) estimate, which, if implemented alone, would cause all means to converge to the MAP (e.g., in settings where target has a single maximum, i.e., MAP is uniquely defined). The repulsion term creates a potential for means to stay away from each other. More precisely, the last term in (19) pushes means away from *sources*. In our case, sources for the adaptation of the n-th location parameter are the other location parameters $\{\boldsymbol{\mu}_j^{(t-1)}\}_{j=1,j\neq n}^N$. In other words, the addition of the term $-\nabla\varphi^N(\boldsymbol{\mu}_n^{(t-1)})$ to the gradient flow above creates a repulsive effect, pushing the updated mean away from the location parameters, which effectively spreads out the components in the mixture proposal. This interpretation also holds when we introduce back m_n, $n = 1,\dots,N$ terms, effectively determining the strength of the repulsion for a particular mean. From this viewpoint, G_t can also be seen as the adaptive weight that determines whether the repulsion term should be more or less active. High values of G_t might be useful in the initial exploration phase.

4. Numerical experiments

4.1. Gaussian mixtures

Let us consider a generic mixture of bivariate Gaussian pdfs as

$$(\forall \mathbf{x} \in \mathbb{R}^2) \quad \widetilde{\pi}(\mathbf{x}) = \frac{1}{L}\sum_{\ell=1}^{L} \mathcal{N}(\mathbf{x}; \boldsymbol{\gamma}_\ell, \boldsymbol{\Sigma}_\ell). \tag{20}$$

We start considering a toy example with $L = 2$ components to better understand the behavior of GRAMIS. In this case, the means are $\boldsymbol{\gamma}_1 = [-5,-5]^\top$and $\boldsymbol{\gamma}_2 = [6,4]^\top$, and the covariance are $\boldsymbol{\Sigma}_1 = [0.25,\ 0; 0,\ 0.25]$ and $\boldsymbol{\Sigma}_2 = [0.52,\ 0.48; 0.48,\ 0.52]$. We run GRAMIS displaying the adaptive behavior of different ablated version of the algorithm. We set $N = 50$ Gaussian proposals, $T = 20$ iterations, and $K = 20$ samples per proposal and iteration. The location parameters of the proposals are initialized in randomly initialized in the square $[1,6]\times[1,6]$.

Figure 1 shows the final location parameters (black dots) and scale parameters (black ellipses) of the proposals at time $t = 20$ for four ablated versions of GRAMIS. Plot (a) shows the modified GRAMIS without preconditioning matrix in the gradient update (as in [62] with $\lambda = 0.1$) and $G_t = 0$ (no repulsion). The arbitrary (suboptimal) step-size delays the convergence of the location parameter to the mode. Plot (b) shows GRAMIS with $G_t = 0$ (no repulsion). The adaptation is effective in recovering one mode, but not in finding the second mode. Plot (c) shows GRAMIS with constant repulsion $G_t = 0.5$. The mixture of proposals has *discovered* both modes, but since the repulsion is not decreased, the proposals cannot concentrate around the mode. Plot (d) shows GRAMIS with exponentially decayed repulsion $G_1 = 0.5$ (see Section 3.3), with the mixture proposal successfully approximating the target density. We denote this mixture as $\widetilde{q}^{(T)}(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} q_n^{(T)}(\mathbf{x}; \boldsymbol{\mu}_n^{(T)}, \boldsymbol{\Sigma}_n^{(T)})$, i.e., the mixture density composed of all the proposals at the last iteration of the algorithm. In all cases, we show also the marginal plots of $\widetilde{q}^{(T)}(\mathbf{x})$ and $\widetilde{\pi}(\mathbf{x})$.

We now extend the setting to $L = 5$ components, with 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 covariance matrices $\boldsymbol{\Sigma}_1 = [5,\ 2; 2,\ 5]$, $\boldsymbol{\Sigma}_2 = [2,\ -1.3; -1.3,\ 2]$, $\boldsymbol{\Sigma}_3 = [2,\ 0.8; 0.8,\ 2]$, $\boldsymbol{\Sigma}_4 = [3,\ 1.2; 1.2,\ 0.5]$ and $\boldsymbol{\Sigma}_5 = [0.2,\ -0.1; -0.1,\ 0.2]$. This target is particularly challenging since it requires the algorithms to discover 5 modes. We aim at estimating the first and second moments, and the normalizing constant, which are available in a closed form. The proposals are now randomly initialized in the square $[-15, 15] \times [-15, 15]$.

Table 2 shows the RMSE in the estimation of Z and the first and second moments of the target in an ablation study of GRAMIS. In particular, we test four versions of the algorithm, with/without preconditioning matrix in the update of the location parameters and with/without repulsion, i.e., the last column is the GRAMIS algorithm Table 1. In the case without preconditioning matrix, we set $\gamma = 10^{-1}$. In the case with repulsion, we use $G_1 = 0.05$ with exponential decay, otherwise we simply set $G_1 = 0$ to annihilate the repulsion effect. The MSE results are obtained over 100 independent runs, with estimators using the weighted samples on the half last iterations. It can be seen that the worst results are obtained when no preconditioning and no repulsion are implemented, while the best results are obtained by the full GRAMIS algorithm.

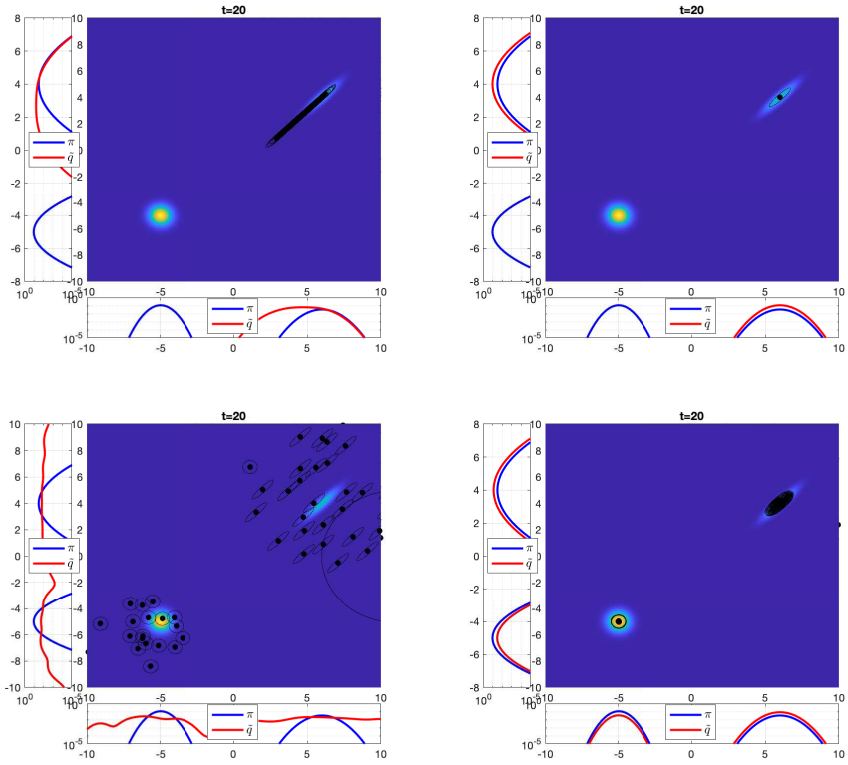

Figure 1: **Toy example.** Final location parameters (black dots) and scale parameters (black ellipses) of the proposals at time $t = 20$ for four ablated versions of GRAMIS. Upper left: modified GRAMIS without pre-conditioning matrix in the gradient update (as in [62] with $\lambda = 0.1$) and $G_t = 0$ (no repulsion). Upper right: GRAMIS with $G_t = 0$ (no repulsion). Bottom left: GRAMISwith constant repulsion $G_t = 0.5$. Bottom right: GRAMISwith exponentially decayed repulsion $G_1 = 0.5$ (see Section 3.3).

RMSE	No pre-cond./No repulsion			No pre-cond./Repulsion			Pre-cond./No repulsion			Pre-cond./Repulsion		
	$\sigma=1$	$\sigma=3$	$\sigma=5$	$\sigma=1$	$\sigma=3$	$\sigma=5$	$\sigma=1$	$\sigma=3$	$\sigma=5$	$\sigma=1$	$\sigma=3$	$\sigma=5$
Z	1.0108	0.0339	0.3152	0.0296	0.0163	0.0291	0.0224	0.0128	0.0192	**0.0096**	0.0168	0.0264
$\mathrm{E}_{\widetilde{\pi}}[\mathbf{X}]$	2.7567	1.6222	2.6798	0.7492	0.7253	0.5969	1.4756	0.9557	1.2745	**0.7694**	0.9097	1.5663
$\mathrm{E}_{\widetilde{\pi}}[\mathbf{X}^2]$	2.5058	1.7431	2.3161	0.6521	0.7439	0.3422	1.6427	0.8829	1.7884	**0.8137**	0.8895	1.6063

Table 2: **Gaussians-mixture target in Section 4.1.** RMSE of the IS estimators. We run an ablation study of GRAMIS with/without preconditioning matrix in the update of the location parameters and with/without repulsion. In the case without preconditioning matrix, we set $\gamma = 10^{-1}$. In the case with repulsion, $G_1 = 0.05$ with exponential decay. The MSE results are obtained over 100 independent runs, with estimators using the weighted samples on the half last iterations.

4.2. Banana-shaped distribution

We now consider a banana-shaped distribution [64, 65]. The shape of this target makes it particularly challenging for sampling methods. The target is the pdf of a r.v. resulting from a transformation of a d_x-dimensional multivariate Gaussian r.v. $\bar{\mathbf{X}} \sim \mathcal{N}(\mathbf{x}; \mathbf{0}_{d_x}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathrm{diag}(c^2, 1, ..., 1)$. The transformed r.v. is $(X_j)_{1 \leq j \leq d_x}$ such that $X_j = \bar{X}_j$ for $j \in \{1, ..., d_x\} \setminus 2$, and $X_2 = \bar{X}_2 - b(\bar{X}_1^2 - c^2)$, where we set $c = 1$ and $b = 3$ in our example.

First, we consider a toy example with $d_x = 2$ so we can obtain intuitive plots. We set $T = 100$, $N = 50$, and $K = 20$. Figure 2 shows the the target distribution, the final location parameters (black dots) and scale parameters (black ellipses) of the proposals at time $T = 100$, and the samples of the last iteration (red dots). We consider the GRAMIS scheme with constant repulsion, with $G_t \in \{0.02, 0.01, 0.005, 0.001, 0.0001, 0\}$. It can be seen that bigger values of G_t yield effectively a mixture with well separated proposals. In this example, the proposals remain in practice static after a few iterations. When G_t is smaller, the proposals tend to concentrate around the mode. In the extreme case with $G_t = 0$ (i.e., without) repulsion, all proposals are effectively the same, which in practice coincides with the Laplace approximation [66].

We now perform comparisons with competitive algorithms, namely PMC using either global (GR) or local (LR) resampling [20], AMIS [10], and O-PMC with LR [23], for various dimension d_x. In AMIS, we set $N = 1$, $K = 500$ and $T = 40$. The other algorithms set $N = 50$, $K = 20$, and $T = 20$, so all algorithms have the same

number of target evaluations. We measure the MSE of all algorithms in estimating $\mathrm{E}_{\widetilde{\pi}}[\mathbf{X}]$. The initialization of the location parameters is randomly done within the square $[-4,4]\times[-4,4]$. All algorithms are initialized with isotropic proposal covariances with $\sigma \in \{1,3,5\}$.

Table 3 shows the MSE of the proposed GRAMIS in the estimation of the target mean for $d_x \in \{5,20,50\}$. In Fig. 3, we compare GRAMIS, with LR-PMC, GR-PMC and O-PMC in a range of dimensions $d_x \in \{2,5,10,15,20,30,40,50\}$. The best performance is reached by GRAMIS in all dimensions, followed by the LR version of the O-PMC. We implement in this example a simple version of GRAMIS without repulsion, which simplifies the parameter tuning for different dimensions. In our GRAMIS and in O-PMC, the MSE tends to decrease when the dimensions grows, which can be explained by the strong structure of the banana-shaped target in high dimensions.

Finally, Fig. (4) shows the MSE in the estimation of $\mathrm{E}_{\widetilde{\pi}}[\mathbf{X}]$ versus the number of total iterations $T \in [10,200]$ for the proposed GRAMIS method (with $\sigma = 1$) using setting values of $G_{\text{rep}} \in \{0,\ 10^{-2}\}$. The estimators are built in all cases by using the last half of the total iterations. In the standard GRAMIS with $G_{\text{rep}} = 10^{-2}$, increasing the number of iterations improves the adaptation and thus the performance. However, if $G_{\text{rep}} = 0$ (i.e., no repulsion), running the algorithm for more iterations worsens the performance. This can be understood by seeing the last plot in Fig. (2). In this case, all proposals tend to be the same, and thus the mixture does not represent well the whole target.

	GR-PMC	LR-PMC	AMIS	LR-O-PMC	GAPIS	GRAMIS
$d_x = 5$	0.2515	0.3418	0.1758	0.0308	0.3007	**0.0029**
$d_x = 20$	0.3818	0.5340	0.1901	0.0098	1.5299	**0.0013**
$d_x = 50$	1.3134	2.3963	0.6074	0.0051	2.5524	**0.0009**

Table 3: **Banana-shaped target in Section 4.2.** 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 all methods, we set the initial proposal variance to $\sigma = 1$. In all PMC-based methods, $(N,K,T) = (50,20,20)$ while $(N,K,T) = (1,500,40)$ for AMIS.

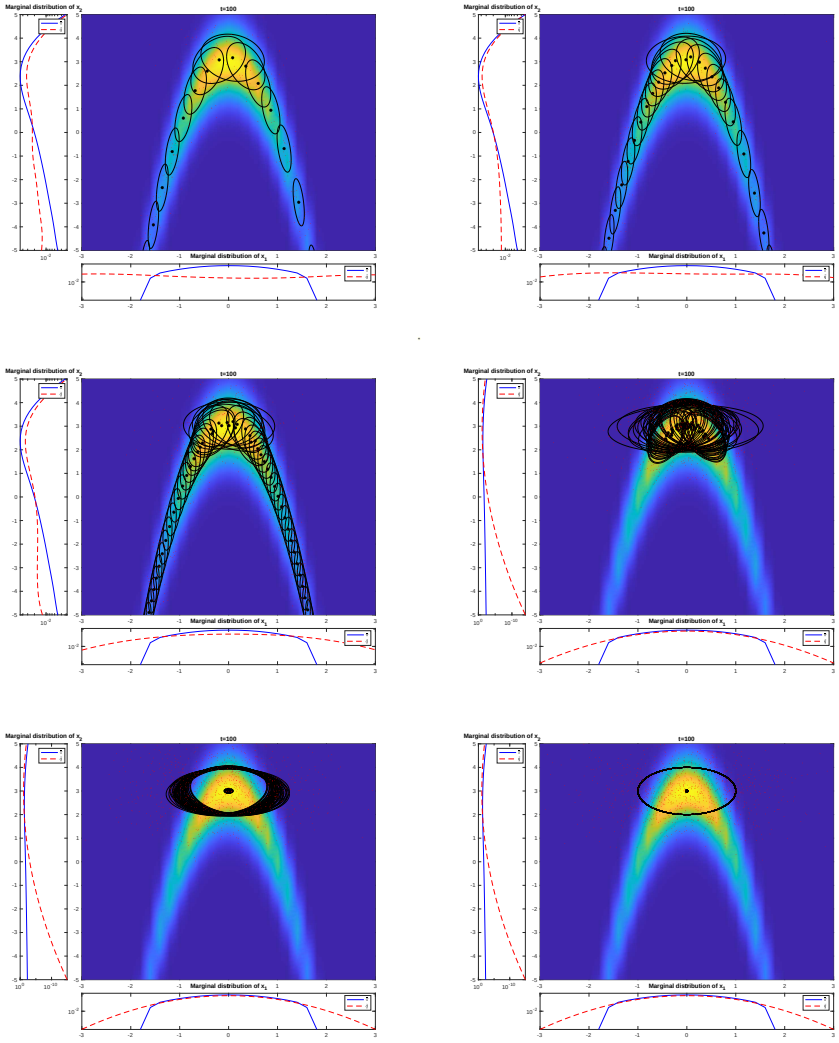

Figure 2: **Banana-shaped target in Section 4.2.** Final location parameters (black dots) and scale parameters (black ellipses) of the proposals at time $T = 100$ for six ablated versions GRAMIS with constant repulsion. In order: $G_t \in \{0.02, 0.01, 0.005, 0.001, 0.0001, 0\}$.

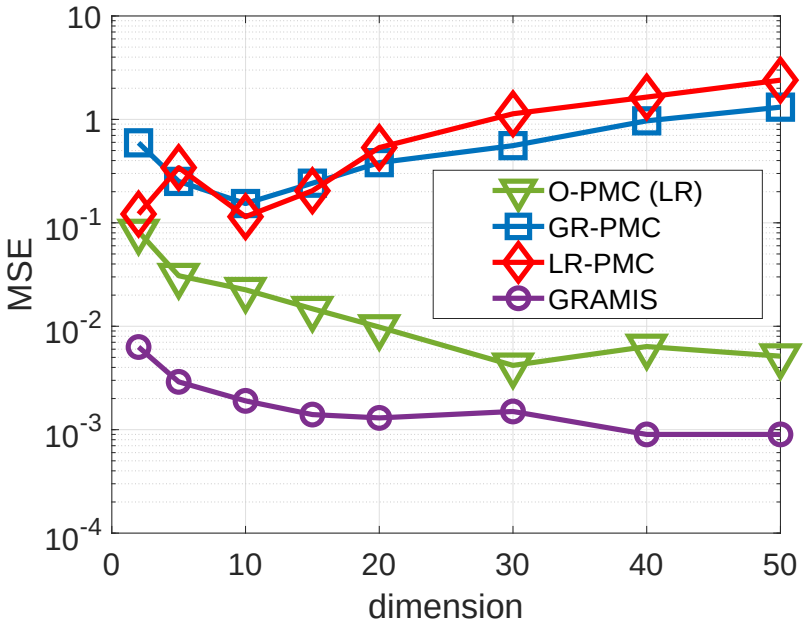

Figure 3: **Banana-shaped target in Section 4.2.** MSE in the estimation of $\mathrm{E}_{\tilde{\pi}}[\mathbf{X}]$ versus the dimension d_x, with GR-PMC, LR-PMC, O-PMC (using LR), and the proposed GRAMIS method (with $\sigma = 1$).

5. Conclusion

In this paper, we have proposed a new algorithm, called GRAMIS, that iteratively adapts a set of proposals with the goal of improving the performance of the importance sampling estimators. The geometric information of the target is exploited by using the first-order and second-order information to improve the location and the scale parameters. A cooperation in the adaptation is allowed by introducing a repulsion term, which can be justified through the lens of Poisson fields. This repulsion becomes essential in multi-modal scenarios and also to represent target densities that, even if uni-modal, cannot be well approximated by standard uni-modal proposals. The GRAMIS algorithm exhibits good exploratory capabilities and a powerful representation of complicated target densities, leading in most cases to lower-variance estimators compared to other AIS algorithms.

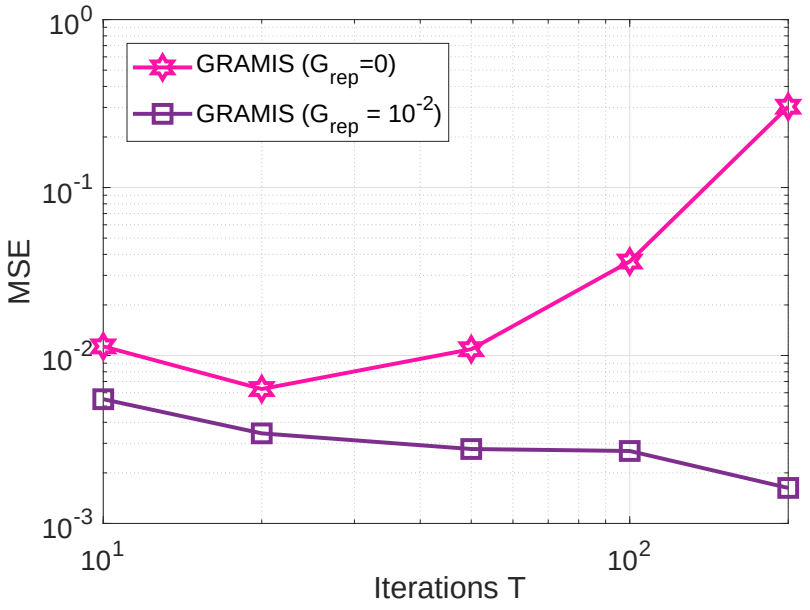

Figure 4: **Banana-shaped target in Section 4.2.** MSE in the estimation of $\mathrm{E}_{\tilde{\pi}}[\mathbf{X}]$ versus the number of iterations T for the proposed GRAMIS method (with $\sigma = 1$) setting $G_{\text{rep}} \in \{0,\ 10^{-2}\}$.

References

[1] C. P. Robert, 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, L. Martino, Advances in importance sampling, Wiley StatsRef: Statistics Reference Online (2021) 1–22.

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

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

[7] R. Douc, A. Guillin, J. M. Marin, C. P. Robert, Convergence of adaptive mixtures of importance sampling schemes, Annals of Statistics 35 (2007) 420–448.

[8] O. D. Akyildiz, J. Míguez, Convergence rates for optimised adaptive importance samplers, Statistics and Computing 31 (2) (2021) 1–17.

[9] Ö. D. Akyildiz, Global convergence of optimized adaptive importance samplers, arXiv preprint arXiv:2201.00409 (2022).

[10] J. M. Cornuet, J. M. Marin, A. Mira, C. P. Robert, Adaptive multiple importance sampling, Scandinavian Journal of Statistics 39 (4) (2012) 798–812.

[11] J.-M. Marin, P. Pudlo, M. Sedki, Consistency of adaptive importance sampling and recycling schemes, Bernoulli 25 (3) (2019) 1977–1998.

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

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

[14] T. Paananen, J. Piironen, P.-C. Bürkner, A. Vehtari, Implicitly adaptive importance sampling, Statistics and Computing 31 (2) (2021) 1–19.

[15] R. Douc, O. Cappé, 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, 2005, pp. 64–69.

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

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

[18] O. Cappé, R. Douc, A. Guillin, J. M. Marin, C. P. Robert, Adaptive importance sampling in general mixture classes, Statistical Computing 18 (2008) 447–459.

[19] E. Koblents, 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, 2013, pp. 1–5.

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

[21] V. Elvira, L. Martino, D. Luengo, M. F. Bugallo, Population monte carlo schemes with reduced path degeneracy, in: 2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), IEEE, 2017, pp. 1–5.

[22] C. Miller, J. N. Corcoran, M. D. Schneider, Rare events via cross-entropy population monte carlo, IEEE Signal Processing Letters 29 (2021) 439–443.

[23] V. Elvira, E. Chouzenoux, Optimized population monte carlo, IEEE Transactions on Signal Processing 70 (2022) 2489–2501.

[24] L. Martino, V. Elvira, D. Luengo, J. Corander, Layered adaptive importance sampling, Statistics and Computing 27 (3) (2015) 599–623.

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

[26] D. Rudolf, B. Sprungk, On a Metropolis–Hastings importance sampling estimator, Electronic Journal of Statistics 14 (1) (2020) 857–889.

[27] A. Mousavi, R. Monsefi, V. Elvira, Hamiltonian adaptive importance sampling, IEEE Signal Processing Letters 28 (2021) 713–717.

[28] M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. Hero, S. McLaughlin, A survey of stochastic simulation and optimization methods in signal processing, IEEE Journal on Selected Topics in Signal Processing 10 (2) (2016) 224–241.

[29] G. O. Roberts, O. Stramer, Langevin diffusions and Metropolis-Hastings algorithms, Methodology and Computing in Applied Probability 4 (4) (2002) 337–357.

[30] A. Durmus, E. Moulines, M. Pereyra, Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau, SIAM Journal on Imaging Sciences 11 (1) (2018) 473–506.

[31] A. Schreck, G. Fort, S. L. Corff, E. Moulines, A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection, IEEE Journal on Selected Topics in Signal Processing 10 (2) (2016) 366–375. doi:10.1109/JSTSP.2015.2496546.

[32] M. Girolami, B. Calderhead, Riemann manifold Langevin and Hamiltonian Monte Carlo methods, Journal of the Royal Statistical Society Series B (Statistical Methodology) 73 (91) (2011) 123–214.

[33] J. Martin, C. L. Wilcox, C. Burstedde, O. Ghattas, A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion, SIAM Journal on Scientific Computing 34 (3) (2012) 1460–1487.

[34] Y. Zhang, 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, 2011, pp. 2393–2401.

[35] Y. Qi, T. P. Minka, Hessian-based Markov Chain Monte-Carlo algorithms, Proceedings of the First Cape Cod Workshop on Monte Carlo

MethodsHttps://www.microsoft.com/en-us/research/publication/hessian-based-markov-chain-monte-carlo-algorithms/ (2002).

[36] Y. Marnissi, E. Chouzenoux, A. Benazza-Benyahia, J.-C. Pesquet, Majorize-Minimize adapted Metropolis-Hastings algorithm, IEEE Transactions on Signal Processing (68) (2020) 2356–2369.

[37] Y. Marnissi, A. Benazza-Benyahia, E. Chouzenoux, 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, 2014, pp. 1332–1336.

[38] Y. Marnissi, E. Chouzenoux, A. Benazza-Benyahia, J.-C. Pesquet, An auxiliary variable method for MCMC algorithms in high dimension, Entropy 20 (110) (2018).

[39] U. Simsekli, R. Badeau, A. T. Cemgil, 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, 2016, p. 642–651.

[40] I. Schuster, Gradient importance sampling, Tech. rep., https://arxiv.org/abs/1507.05781 (2015).

[41] M. Fasiolo, F. E. de Melo, S. Maskell, Langevin incremental mixture importance sampling, Statistical Computing 28 (3) (2018) 549–561.

[42] G. O. Roberts, L. R. Tweedie, Exponential convergence of Langevin distributions and their discrete approximations, Bernoulli 2 (4) (1996) 341–363.

[43] Y. El-Laham, V. Elvira, M. F. Bugallo, Robust covariance adaptation in adaptive importance sampling, IEEE Signal Processing Letters (2018).

[44] Y. El-Laham, V. Elvira, 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, 2019, pp. 624–628.

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

[46] E. Veach, 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), 1995, pp. 419–428.

[47] M. Welling, Y. W. Teh, Bayesian learning via stochastic gradient langevin dynamics, in: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), 2011.

[48] C. Rasmussen, A practical Monte Carlo implementation of bayesian learning, in: Proceedings of the Neural Information Processing Systems Conference (NIPS 1995), 1995.

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

[50] I. Kondapaneni, P. Vévoda, P. Grittmann, T. Skřivan, P. Slusallek, J. Křivánek, Optimal multiple importance sampling, ACM Transactions on Graphics (TOG) 38 (4) (2019) 1–14.

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

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

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

[54] V. Elvira, L. Martino, D. Luengo, M. F. Bugallo, Multiple importance sampling with overlapping sets of proposals, in: 2016 IEEE Statistical Signal Processing Workshop (SSP), IEEE, 2016, pp. 1–5.

[55] D. Talay, L. Tubaro, Expansion of the global error for numerical schemes solving stochastic differential equations, Stochastic Analysis and Applications 8 (4) (1991) 483–509.

[56] A. Durmus, E. Moulines, High-dimensional Bayesian inference via the unadjusted Langevin algorithm, Bernoulli (4A) (2019) 2854–2882.

[57] C. Vacar, J.-F. Giovannelli, 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, 2011, pp. 3964–3967.

[58] T. Xifara, C. Sherlock, S. Livingstone, S. Byrne, M. Girolami, Langevin diffusions and the Metropolis-adjusted Langevin algorithm, Statistics and Probability Letters 91 (2014) 14 – 19.

[59] S. Sabanis, Y. Zhang, Higher order Langevin Monte Carlo algorithm, Electronic Journal of Statistics 13 (2) (2019) 3805–3850.

[60] L. Martino, V. Elvira, D. Luengo, A. Artes-Rodriguez, J. Corander, Orthogonal mcmc algorithms, in: 2014 IEEE Workshop on Statistical Signal Processing (SSP), IEEE, 2014, pp. 364–367.

[61] L. Martino, V. Elvira, D. Luengo, A. Artés-Rodríguez, J. Corander, Smelly parallel mcmc chains, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 4070–4074.

[62] V. Elvira, L. Martino, L. Luengo, 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, 2015, pp. 4075–4079.

[63] Y. Xu, Z. Liu, M. Tegmark, T. Jaakkola, Poisson flow generative models, arXiv preprint arXiv:2209.11178 (2022).

[64] H. Haario, E. Saksman, J. Tamminen, Adaptive proposal distribution for random walk Metropolis algorithm, Computational Statistics 14 (3) (1999) 375–396.

[65] H. Haario, E. Saksman, J. Tamminen, An adaptive Metropolis algorithm, Bernoulli 7 (2) (2001) 223–242.

[66] Z. Shun, P. McCullagh, Laplace approximation of high dimensional integrals, Journal of the Royal Statistical Society: Series B (Methodological) 57 (4) (1995) 749–760.