A VARIATIONAL ADAPTIVE POPULATION IMPORTANCE SAMPLER

Yousef El-Laham, Petar M. Djurić, Mónica F. Bugallo

Department of Electrical and Computer Engineering
Stony Brook University, Stony Brook, NY 11794-2350
{yousef.ellaham, petar.djuric, monica.bugallo}@stonybrook.edu

ABSTRACT

Adaptive importance sampling (AIS) methods are a family of algorithms which can be used to approximate Bayesian posterior distributions. Many AIS algorithms exist in the literature, where the differences arise in the manner by which the proposal distribution is adapted at each iteration. The adaptive population importance sampler (APIS), for example, deterministically samples from a mixture distribution and uses the local information given by the samples and weights to adapt the location parameter of each proposal. The update rules by nature are heuristic, but effective, especially in the case that the target posterior is multimodal. In this work, we introduce a novel AIS scheme which incorporates modern techniques in stochastic optimization to improve the methodology for higher-dimensional posterior inference. More specifically, we derive update rules for the parameters of each proposal by means of deterministic mixture sampling and show that the method outperforms other state-of-the-art approaches in high-dimensional scenarios.

***Index Terms*—** Monte Carlo methods, adaptive importance sampling, mixture distributions, stochastic optimization

1. INTRODUCTION

Monte Carlo (MC) methods are computational schemes which use the concept of random sampling to obtain numerical approximations [1]. A well-known application of MC schemes is in Bayesian inference, where we wish to approximate the posterior distribution of a set of unknowns given noisy observations [2]. For decades, the preferred strategy for posterior inference has been Markov chain Monte Carlo (MCMC) sampling, an approach that constructs a Markov process in order to sample from the posterior distribution. MCMC methods are coupled with nice theoretical guarantees, in that the Markov process eventually converges to a stationary distribution that is the posterior [3]. However, a significant drawback of MCMC sampling is that the methodology may break down for complex systems (i.e., large number of unknowns or large sets of observed data) due to poor mixing of the constructed Markov chain. In this setting, a strong competitor to MCMC sampling is variational inference (VI). VI approaches the inference problem using optimization, where the goal is to minimize the Kullback-Leibler divergence (KLD) between the posterior distribution and a member of a family of probability distributions [4]. A strong advantage of using VI is in its scalability for complex models and big data applications, thanks to advances in stochastic optimization [5, 6]. Unfortunately, VI methods are not guaranteed to converge to the true posterior. Thus, it is of utmost interest in Bayesian signal processing to develop methodologies that possess similar theoretical guarantees to MCMC samplers while also obtaining faster results for complex probabilistic models in the way VI does. A natural area of study is in hybrid methodologies which utilize MC and optimization for scalable posterior inference.

Importance sampling (IS) is an MC methodology that allows for approximation of some target distribution using weighted samples generated from another proposal distribution [7]. The variance of the IS estimator depends on the fit of the proposal to the target. This presents a challenge in high-dimensional systems because the choice of the proposal distribution is not straightforward. Adaptive importance sampling (AIS) implements an iterative version of IS, whereby the proposal distribution is adapted with each iteration to better fit the target density [8]. Numerous algorithms have been proposed in the literature which have advanced AIS. Advances include the development of alternative weighting schemes and effective parameter updates for mixture proposal distributions [9, 10, 11, 12, 13]. Some hybrid methods have even been proposed which combine MCMC and AIS schemes [14]. Alternatively, an interesting approach is the incorporation of stochastic optimization techniques within the AIS framework. For example, in [15, 16] the parameters of a proposal distribution from the exponential family were adapted to minimize the per-sample variance of the IS estimator. Other techniques of this flavor have been proposed [17], but little work has been done in the context of efficiently optimizing a mixture proposal distribution.

In this work, we explore the optimization of the parameters of a mixture proposal distribution in the context of AIS. Our contribution is twofold. First, we show that the considered objective functions and their gradients can be reformulated as a combination of expectations w.r.t. the mixture components separately. This allows for simplicity in implementation through deterministic sampling methods, which guarantees representation of each mixand in the computation of the stochastic gradients. Second, we develop a novel AIS scheme which adapts a population of proposal distributions efficiently using stochastic optimization techniques. We show that the novel sampler attains superior performance compared to other AIS schemes.

The paper is organized as follows. First, in Section 2, we formulate the problem. Section 3 reviews preliminaries and prior work, while Section 4 describes the novel scheme. We provide simulation results in Section 5 and conclude the paper in Section 6.

2. PROBLEM FORMULATION

Suppose that we have a set of i.i.d. observed data $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N \sim p(\mathbf{y}|\mathbf{x})$, where $\mathbf{y}_i \in \mathbb{R}^{d_y}$ for $i = 1, ..., N$ and $\mathbf{x} \in \mathbb{R}^{d_x}$ is a vector of unknowns. We address the problem of fully Bayesian inference,

M.B. thanks the support of the NSF under Award CCF-1617986, and P. M. D. the support of the NSF under Award CCF-1618999. The authors would like to thank Stony Brook Research Computing and Cyberinfrastructure, and the Institute for Advanced Computational Science at Stony Brook University for access to the high-performance SeaWulf computing system, which was made possible by a $1.4M NSF grant (# 1531492).

in which the goal is to sample from the posterior distribution,

$$\pi(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \propto \ell(\mathbf{y}|\mathbf{x})p(\mathbf{x}), \tag{1}$$

where $\ell(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{N} p(\mathbf{y}_i|\mathbf{x})$ is the likelihood of the unknowns, $p(\mathbf{x})$ is the prior distribution of the unknowns, and $Z = p(\mathbf{y})$ is called the evidence, which depends only on the data. We define $\tilde{\pi}(\mathbf{x}) \equiv \ell(\mathbf{y}|\mathbf{x})p(\mathbf{x})$. We wish to estimate the evidence through IS from a proposal distribution $q(\mathbf{x};\boldsymbol{\theta})$. The IS estimate is,

$$\hat{Z}_{IS} = \frac{1}{M}\sum_{m=1}^{M} \frac{\tilde{\pi}(\mathbf{x}^{(m)})}{q(\mathbf{x}^{(m)};\boldsymbol{\theta})}, \tag{2}$$

where $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)} \sim q(\mathbf{x};\boldsymbol{\theta})$. For efficient approximation of the target posterior, our goal is to minimize the variance of (2) by optimizing over the proposal parameters $\boldsymbol{\theta}$.

3. PRELIMINARIES

In this section, we review IS and AIS for estimation of arbitrary target distributions. We also summarize some advances in the field of VI which allow for scalable approximate Bayesian inference through stochastic optimization. Furthermore, we review a methodology which combines stochastic optimization and AIS.

3.1. Adaptive Importance Sampling (AIS)

IS is an MC tool used to approximate properties of statistical distributions through weighted samples drawn from proposal distributions. For example, suppose M samples are generated as follows: $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)} \sim q(\mathbf{x};\boldsymbol{\theta})$. If we wish to estimate an expectation of a particular function $f(\mathbf{x})$ w.r.t. to the unnormalized distribution $\tilde{\pi}(\mathbf{x})$, we can apply the self-normalized IS estimator,

$$\hat{I}_{IS} = \frac{1}{\sum_{m=1}^{M} w^{(m)}} \sum_{m=1}^{M} w^{(m)} f(\mathbf{x}^{(m)}), \tag{3}$$

where $w^{(m)} = \tilde{\pi}(\mathbf{x}^{(m)})/q(\mathbf{x}^{(m)};\boldsymbol{\theta})$ and the term $\sum_{m=1}^{M} w^{(m)}$ is a normalization constant. For simplicity we denote the set of normalized importance weights as $\{\bar{w}^{(m)}\}_{m=1}^{M}$. Besides point estimation, there is a growing interest in applying IS in the context of fully Bayesian inference, whereby a set of samples and their corresponding normalized importance weights $\{\mathbf{x}^{(m)}, \bar{w}^{(m)}\}_{m=1}^{M}$ can approximate the normalized target distribution $\pi(\mathbf{x})$,

$$\hat{\pi}_{IS}(\mathbf{x}) = \sum_{m=1}^{M} \bar{w}^{(m)} \delta(\mathbf{x} - \mathbf{x}^{(m)}). \tag{4}$$

The accuracy of the estimate in (4) depends on the fit of the proposal to the target distribution, which is challenging in high dimensions. To tackle this challenge, we can employ AIS. AIS uses a learning algorithm to improve IS, where $q(\mathbf{x};\boldsymbol{\theta})$ is adapted over a set of iterations. Given a set of samples drawn from $q(\mathbf{x};\boldsymbol{\theta}_t)$ and their unnormalized weights $\{\mathbf{x}_t^{(m)}, w_t^{(m)}\}_{m=1}^{M}$ for $t = 1, ..., T$, with T being the number of iterations, the approximation of the target is given by,

$$\hat{\pi}_{AIS}(\mathbf{x}) = \frac{1}{\sum_{t=1}^{T}\sum_{m=1}^{M} w_t^{(m)}} \sum_{t=1}^{T}\sum_{m=1}^{M} w_t^{(m)} \delta(\mathbf{x} - \mathbf{x}_t^{(m)}). \tag{5}$$

Many algorithms exist which use different update rules to efficiently adapt the proposal distribution to the target. Our focus will be on utilizing techniques in stochastic optimization which will iteratively minimize some discrepancy measure between the proposal and target distributions.

Algorithm: Projected Stochastic Gradient Descent

1. Initialization: Select the initial parameter value $\boldsymbol{\theta}_1$.
2. For $t = 1, ..., T$
a. Calculate the stochastic gradient $\tilde{g}(\boldsymbol{\theta}_t)$.
b. Update the parameter value,
$$\boldsymbol{\theta}_{t+1} = \Pi_{\mathcal{C}}\big(\boldsymbol{\theta}_t - \eta_t \tilde{g}(\boldsymbol{\theta}_t)\big).$$
3. Return $\boldsymbol{\theta}_{T+1}$.

Table 1: The parameter η_t denotes a decreasing learning rate ($\eta_t \to 0$ as $t \to \infty$). $\Pi_{\mathcal{C}}(\cdot)$ denotes the projection onto the feasible set $\mathcal{C}$.

3.2. Black Box Variational Inference (BBVI)

VI approaches the approximate inference problem using optimization. Black box variational inference (BBVI) [6] uses stochastic optimization to minimize a discrepancy measure between the target distribution and the proposal distribution. A common discrepancy between probability distributions is the KLD, $\mathcal{D}_{KL}(q_{\boldsymbol{\theta}}||\pi)$. In VI, minimizing this discrepancy measure is equivalent to minimizing the negative evidence lower-bound (ELBO),

$$\mathcal{L}(\boldsymbol{\theta}) = -\int_{-\infty}^{\infty} q(\mathbf{x};\boldsymbol{\theta}) \log\left(\frac{\tilde{\pi}(\mathbf{x})}{q(\mathbf{x};\boldsymbol{\theta})}\right) d\mathbf{x}, \tag{6}$$

w.r.t. the proposal parameters $\boldsymbol{\theta}$. The gradient of this can be written as an expectation w.r.t. the distribution $q(\mathbf{x};\boldsymbol{\theta})$,

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = -\mathbb{E}_q\left[\log\left(\frac{\tilde{\pi}(\mathbf{x})}{q(\mathbf{x};\boldsymbol{\theta})}\right)\nabla_{\boldsymbol{\theta}}\Big(\log q(\mathbf{x};\boldsymbol{\theta})\Big)\right], \tag{7}$$

which can be approximated using an MC estimate with samples drawn from $q(\mathbf{x};\boldsymbol{\theta})$. In BBVI, these stochastic gradients (coupled with variance reduction tricks) [18] allow for scalable Bayesian inference using optimizers such as projected stochastic gradient descent (see Table 1). The optimized distribution, however, is not guaranteed to converge to the true posterior distribution. We also note that, for general cases, the minimization of the negative ELBO is a nonconvex optimization problem w.r.t. $\boldsymbol{\theta}$.

3.3. Convex AdaMC

There are also algorithms that couple AIS methods with stochastic optimization. For instance, an algorithm called Convex AdaMC (CAMC) [15, 16] minimizes the Rényi divergence,

$$\mathcal{D}_{\alpha}(\pi||q_{\boldsymbol{\theta}}) = \frac{1}{\alpha - 1}\log\left(\int_{-\infty}^{\infty} \pi(\mathbf{x})^{\alpha} q(\mathbf{x};\boldsymbol{\theta})^{1-\alpha} d\mathbf{x}\right). \tag{8}$$

We note that the limiting case $\alpha = 1$ corresponds to $\mathcal{D}_{KL}(\pi||q_{\boldsymbol{\theta}})$. It can be shown that minimizing (8) does not depend on the normalization constant and we can replace $\pi(\mathbf{x})$ with $\tilde{\pi}(\mathbf{x})$. CAMC implicitly minimizes the per-sample variance of the IS estimator by minimizing a monotonic transformation of (8) for $\alpha > 1$, i.e., $\exp((\alpha - 1)\mathcal{D}_{\alpha}(\pi||q_{\boldsymbol{\theta}}))$. For the estimator in (2), the per-sample variance can be written as

$$\begin{aligned} V(\boldsymbol{\theta}) &= \mathbb{E}_q\left[\left(\frac{\tilde{\pi}(\mathbf{x})}{q(\mathbf{x};\boldsymbol{\theta})}\right)^2\right] - \left(\mathbb{E}_q\left[\frac{\tilde{\pi}(\mathbf{x})}{q(\mathbf{x};\boldsymbol{\theta})}\right]\right)^2 \\ &= \int_{-\infty}^{\infty} \frac{\tilde{\pi}(\mathbf{x})^2}{q(\mathbf{x};\boldsymbol{\theta})} d\mathbf{x} - Z^2. \end{aligned} \tag{9}$$

Algorithm: VAPIS

1. **Initialization:** Set $\boldsymbol{\theta}_{1,k}$ for $k = 1, ..., K$.
2. **For** $t = 1, ..., T$
   a. Draw N samples from K proposal distributions,
   $$\mathbf{x}_{t,k}^{(n)} \sim q_k(\mathbf{x}; \boldsymbol{\theta}_{t,k}), \quad \begin{matrix} n = 1, \dots, N, \\ k = 1, \dots, K. \end{matrix}$$
   b. Compute the deterministic mixture weights,
   $$w_{t,k}^{(n)} = \frac{\tilde{\pi}(\mathbf{x}_{t,k}^{(n)})}{\frac{1}{K}\sum_{k=1}^{K} q_k(\mathbf{x}_{t,k}^{(n)}; \boldsymbol{\theta}_{t,k})}, \quad \begin{matrix} n = 1, \dots, N, \\ k = 1, \dots, K. \end{matrix}$$
   c. Approximate the target distribution,
   $$\hat{\pi}_t(\mathbf{x}) = \frac{\sum_{\tau=1}^{t}\sum_{k=1}^{K}\sum_{n=1}^{N} w_{\tau,k}^{(n)} \delta(\mathbf{x} - \mathbf{x}_{\tau,k}^{(n)})}{\sum_{\tau=1}^{t}\sum_{k=1}^{K}\sum_{n=1}^{N} w_{\tau,k}^{(n)}}.$$
   d. **For** $k = 1, ..., K$
      i. Compute the stochastic gradient $\tilde{g}(\boldsymbol{\theta}_{t,k})$.
      ii. Update the vector of proposal parameters $\boldsymbol{\theta}_{t,k}$,
      $$\boldsymbol{\theta}_{t+1,k} = \Pi_{\mathcal{C}}\big(\boldsymbol{\theta}_{t,k} - \eta_t \tilde{g}(\boldsymbol{\theta}_{t,k})\big)$$
3. Return the approximation $\hat{\pi}_T(\mathbf{x})$.

Table 2: Proposed methodology.

The optimization problem will only depend on the first term, which is proportional to $\exp(\mathcal{D}_2(\pi||q_{\boldsymbol{\theta}}))$. We can compute the gradient as,

$$\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) = -\mathbb{E}_q\left[\left(\frac{\tilde{\pi}(\mathbf{x})}{q(\mathbf{x};\boldsymbol{\theta})}\right)^2 \nabla_{\boldsymbol{\theta}}\Big(\log q(\mathbf{x};\boldsymbol{\theta})\Big)\right], \tag{10}$$

which can also be approximated using an MC estimate. Furthermore, when $q(\mathbf{x};\boldsymbol{\theta})$ is chosen from the exponential family (e.g., Gaussian), the minimization of (9) is a convex optimization problem. Coupled with off-the-shelf stochastic optimization algorithms, we can find the optimal vector of parameters $\boldsymbol{\theta}$, which minimizes the variance of the estimator in (2). Unfortunately, this convexity does not hold in general for mixture proposal distributions.

4. METHODOLOGY

In this section, we introduce a novel framework for AIS in high dimensions. Specifically, we show that the gradients of the considered objective functions can be reformulated when mixture proposal distributions are used to enable deterministic sampling from each mixand. We then introduce a novel algorithm called *variational adaptive population importance sampling* (VAPIS) which adapts the parameters of a population of proposal distributions.

4.1. Reformulation of Stochastic Gradients

In general, the optimization problem we would like to solve is,

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\in\mathcal{C}}{\arg\min}\ C(\boldsymbol{\theta}), \tag{11}$$

where $C(\boldsymbol{\theta})$ denotes the objective function and $\mathcal{C}$ denotes the feasible set. In order to solve the minimization problem, we need access to the stochastic gradients. We consider objective functions such that the gradients can generalize as follows:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} C(\boldsymbol{\theta}) &= -\mathbb{E}_q\big[\Phi(\mathbf{x},\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}(\log q(\mathbf{x};\boldsymbol{\theta}))\big] \\ &= -\int_{-\infty}^{\infty} q(\mathbf{x};\boldsymbol{\theta})\Phi(\mathbf{x},\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}(\log q(\mathbf{x};\boldsymbol{\theta}))d\mathbf{x},\end{aligned} \tag{12}$$

where the term $\Phi(\mathbf{x},\boldsymbol{\theta})$ depends on the chosen objective function. For the Rényi divergence, if $\alpha = 1$, we directly minimize $\mathcal{D}_1(\pi||q_{\boldsymbol{\theta}})$. For $\alpha > 1$, we minimize $\exp((\alpha-1)\mathcal{D}_\alpha(\pi||q_{\boldsymbol{\theta}}))$ as in [16]. This yields the following multiplier for minimizing the Rényi divergence:

$$\Phi(\mathbf{x},\boldsymbol{\theta}) = \left(\frac{\tilde{\pi}(\mathbf{x})}{q(\mathbf{x};\boldsymbol{\theta})}\right)^{\alpha}, \quad \alpha \geq 1. \tag{13}$$

Consider a mixture proposal distribution with K mixands, $q(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{K}\rho_k q_k(\mathbf{x};\boldsymbol{\theta}_k)$, where ρ_k and $\boldsymbol{\theta}_k$ denote the mixand weight and parameters, respectively. In the current form, the gradient w.r.t. the parameters of the kth mixand is given by,

$$\nabla_{\boldsymbol{\theta}_k} C(\boldsymbol{\theta}) = -\mathbb{E}_q\left[\frac{\nabla_{\boldsymbol{\theta}_k}(q(\mathbf{x};\boldsymbol{\theta}))}{q(\mathbf{x};\boldsymbol{\theta})}\Phi(\mathbf{x},\boldsymbol{\theta})\right]. \tag{14}$$

Suppose $\nabla_{\boldsymbol{\theta}_k}(q(\mathbf{x};\boldsymbol{\theta})) = \rho_k q_k(\mathbf{x};\boldsymbol{\theta}_k)\Psi(\mathbf{x},\boldsymbol{\theta}_k)$, where $\Psi(\mathbf{x},\boldsymbol{\theta}_k)$ is a function that depends on the choice of $q_k(\mathbf{x};\boldsymbol{\theta}_k)$. Then, we can alternatively write the gradient of the cost w.r.t. $\boldsymbol{\theta}_k$ as an expectation w.r.t. the individual mixand $q_k(\mathbf{x};\boldsymbol{\theta}_k)$,

$$\nabla_{\boldsymbol{\theta}_k} C(\boldsymbol{\theta}) = -\rho_k \mathbb{E}_{q_k}\big[\Psi(\mathbf{x},\boldsymbol{\theta}_k)\Phi(\mathbf{x},\boldsymbol{\theta})\big]. \tag{15}$$

This refined form allows for computation of the gradients by deterministically sampling from each mixand separately.

4.2. Mixture of Exponential Family Members

Consider that each mixand $q_k(\mathbf{x};\boldsymbol{\theta}_k)$ belongs to the exponential family of probability distributions, i.e.,

$$q_k(\mathbf{x};\boldsymbol{\theta}_k) = h(\mathbf{x})\exp(\boldsymbol{\beta}(\boldsymbol{\theta}_k)^{\mathsf{T}}\mathbf{T}(\mathbf{x}) - A(\boldsymbol{\theta}_k)). \tag{16}$$

The gradient of such distributions w.r.t. $\boldsymbol{\theta}_k$ is given by,

$$\nabla_{\boldsymbol{\theta}_k} q_k(\mathbf{x};\boldsymbol{\theta}_k) = q_k(\mathbf{x};\boldsymbol{\theta}_k)\nabla_{\boldsymbol{\theta}_k}(\boldsymbol{\beta}(\boldsymbol{\theta}_k)^{\mathsf{T}}\mathbf{T}(\mathbf{x}) - A(\boldsymbol{\theta}_k)). \tag{17}$$

Thus, we have that if $q_k(\mathbf{x};\boldsymbol{\theta}_k)$ is in the exponential family, then $\Psi(\mathbf{x},\boldsymbol{\theta}_k) = \nabla_{\boldsymbol{\theta}_k}(\boldsymbol{\beta}(\boldsymbol{\theta}_k)^{\mathsf{T}}\mathbf{T}(\mathbf{x}) - A(\boldsymbol{\theta}_k))$ and the decomposition made to obtain (15) is possible.

4.3. Comments on Gradient Computation

Suppose that we would like to compute the stochastic gradient of $C(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}_k$. If we were to compute using (14), this would require us to sample from the full mixture $q(\mathbf{x};\boldsymbol{\theta})$. To reduce computation, a small mini-batch of samples from the mixture could be drawn; however, this would also limit the representation of $q_k(\mathbf{x};\boldsymbol{\theta}_k)$ in the computation of the gradient. If we use (15) to compute the gradient, we only need to sample from $q_k(\mathbf{x};\boldsymbol{\theta}_k)$, guaranteeing representation of the mixand regardless of the size of the mini-batch of samples.

4.4. Algorithm Summary

The newly proposed algorithm, called *variational adaptive population importance sampling* (VAPIS) is summarized in Table 2. The algorithm deterministically samples from an equally weighted mixture distribution. We approximate the target using the so-called *deterministic mixture weights* [19, 20] allowing for a reduced-variance solution at a higher computational cost. Parameter updates are governed by minimizing the objective function through stochastic optimization. We note that the stochastic gradients $\tilde{g}(\boldsymbol{\theta}_{t,k})$ are computed using an MC estimate of (15). Furthermore, we emphasize that step d. of the algorithm is a trivially parallelizable task in practice. The general algorithm incorporates projected stochastic gradient descent

σ^2	2	4	6	8	10
M-PMC	27.90	27.91	27.65	27.77	27.67
DM-PMC	7.06	9.05	13.90	16.99	19.90
APIS	1.58	3.52	9.20	14.13	18.76
VAPIS	**0.05**	**0.04**	**0.18**	**0.23**	**0.84**

Table 3: MSE in the estimation of $\mathbb{E}_\pi[\mathbf{x}]$.

σ^2	2	4	6	8	10
M-PMC	496.2	480.3	500.6	419.3	415.6
DM-PMC	483.5	457.9	659.2	552.4	631.7
APIS	134.4	195.9	472.2	563.9	563.3
VAPIS	**21.1**	**21.8**	**55.9**	**42.8**	**84.3**

Table 4: MAE in the estimation of Z.

to solve the optimization task. Alternatively, we can employ a more sophisticated optimization algorithm such as RMSprop (see [21] for a review), which allows for adaptive learning rates for each component of $\boldsymbol{\theta}_k$.

4.4.1. Location Parameters for a Mixture of Gaussians

Consider the optimization of the means of an equally weighted mixture of Gaussians w.r.t. the objective $C(\boldsymbol{\theta}) = \exp(\mathcal{D}_2(\pi||q_{\boldsymbol{\theta}}))$. The proposal is given by $q(\mathbf{x}; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}) = \frac{1}{K}\sum_{k=1}^{K} q_k(\mathbf{x}; \boldsymbol{\mu}_{t,k}, \boldsymbol{\Sigma}_k)$, where $q_k(\mathbf{x}; \boldsymbol{\mu}_{t,k}, \boldsymbol{\Sigma}_k) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{t,k}, \boldsymbol{\Sigma}_k)$. The stochastic gradient w.r.t. $\boldsymbol{\mu}_{t,k}$ is given by,

$$\tilde{g}(\boldsymbol{\mu}_{t,k}) = -\frac{\boldsymbol{\Sigma}_k^{-1}}{KN}\sum_{n=1}^{N}\left(\frac{\tilde{\pi}(\mathbf{x}_{t,k}^{(n)})}{q(\mathbf{x}_{t,k}^{(n)}; \boldsymbol{\mu}_t, \boldsymbol{\Sigma})}\right)^2(\mathbf{x}_{t,k}^{(n)} - \boldsymbol{\mu}_{t,k}), \quad (18)$$

where $\mathbf{x}_{t,k}^{(n)} \sim \mathcal{N}(\boldsymbol{\mu}_{t,k}, \boldsymbol{\Sigma}_k)$ for $n = 1, \ldots, N$. If we assume that $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbb{I}_{d_x}$, where $\mathbb{I}_{d_x}$ denotes the $d_x \times d_x$ identity matrix, then the computation of $\boldsymbol{\Sigma}_k^{-1}$ is trivial. When we consider the algorithm in Table 2 with only adapting the location parameters, we obtain a method that resembles the APIS technique [12]. APIS uses locally weighted estimates of the target mean in order to adapt the location of each proposal. Our algorithm instead uses stochastic optimization, which implies two key advantages. First, VAPIS is explicitly optimizing a desired objective function, i.e., the per-sample variance of the normalizing constant. Second, the parameter adaptations in VAPIS can be resolved with any off-the-shelf stochastic optimization algorithm, allowing for scalability to higher-dimensional probabilistic models.

5. SIMULATIONS

In this section, we present numerical experiments to demonstrate the performance of the proposed methodology. We simulated the proposed VAPIS algorithm according to the update rules in Section 4.4.1 and utilized the RMSprop algorithm to run the optimization.

We considered the toy example of approximating a Gaussian mixture in $\mathbb{R}^{20}$ of the following form:

$$\pi(\mathbf{x}) \propto \tilde{\pi}(\mathbf{x}) = \sum_{j=1}^{5} \tilde{\rho}_j \mathcal{N}(\mathbf{x}; \boldsymbol{m}_j, \boldsymbol{\Lambda}_j), \quad (19)$$

where $\boldsymbol{\Lambda}_j = \tilde{\boldsymbol{\Lambda}}_j + 2 \times \mathbb{I}_{20}$, such that $\tilde{\boldsymbol{\Lambda}}_j \sim \mathcal{IW}(\mathbb{I}_{20}, 20)$ for $j = 1, \ldots 5$, i.e., an inverse Wishart distribution with scale matrix $\mathbb{I}_{20}$ and 20 degrees of freedom. We assumed that we could compute $\tilde{\pi}(\mathbf{x})$ but the parameters of the mixture were unknown. We generated the target means as $\boldsymbol{m}_j \sim \mathcal{U}([-10, 10]^{20})$ and the unnormalized mixture weights $\tilde{\rho}_j \sim \mathcal{G}(10, 10)$ for $j = 1, .., 5$, i.e., they were drawn from a Gamma distribution with shape and scale parameters equal to 10. The objective was to estimate the normalization constant $Z = \sum_{j=1}^{5} \tilde{\rho}_j = 511.3$ and the target mean $\mathbb{E}_\pi[\mathbf{x}] = \frac{1}{Z}\sum_{j=1}^{5} \tilde{\rho}_j \boldsymbol{m}_j$.

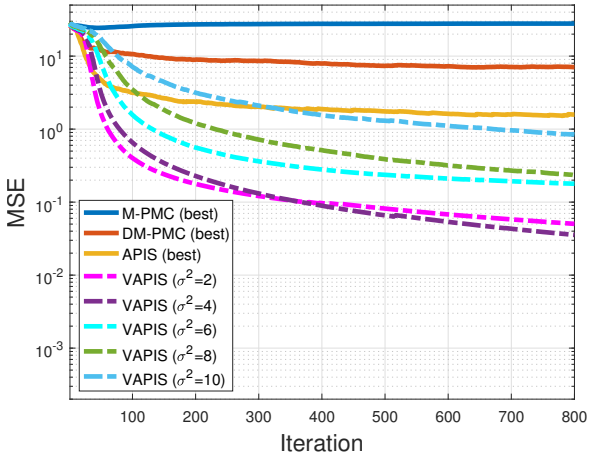

Fig. 1: Evolution of the MSE for each of the compared methods.

We tested and compared the following algorithms: the mixture population Monte Carlo (M-PMC) [10], the population Monte Carlo with deterministic mixture weights (DM-PMC) [20], APIS [12] and the novel VAPIS method with different configurations. The considered error metric was the mean absolute error (MAE) for the normalizing constant and the mean square error (MSE) for the target mean. For each algorithm, we used $K = 100$ mixands for the proposal. We generated $M = 1000$ total samples per iteration (10 samples per mixand) over $I = 800$ iterations and we averaged the results over 500 MC simulations. The prior means and covariances were generated according to $\boldsymbol{\mu}_{1,k} \sim \mathcal{U}([-10, 10]^{20})$ and $\boldsymbol{\Sigma}_k = \sigma^2 \mathbb{I}_{20}$ for $k = 1, ..., 100$, where $\sigma^2 \in \{2, 4, 6, 8, 10\}$.

Tables 3 and 4 show the results of the numerical experiment. We can see that for each value of σ^2, the proposed method outperforms the state-of-the-art approaches. The closest competitors to VAPIS are DM-PMC and APIS, but due to the small number of samples generated per proposal, the update rules for DM-PMC and APIS (resampling and locally weighted mean estimates) are not robust to the high dimension of the target. VAPIS, on the other hand, performs very well despite the small number of samples generated per proposal. These results are confirmed in Fig. 1, where VAPIS clearly outperforms the rest of the methods (under their best setting of σ^2) in terms of MSE for all choices of σ^2.

6. CONCLUSIONS

In this work, we proposed a technique that embeds stochastic optimization within the AIS framework. We showed that when a mixture proposal distribution is utilized, the gradient of interesting objective functions, such as the variance of the IS estimate of the normalizing constant of a target posterior, can be decomposed in a way that allows for calculation of components through deterministic mixture sampling. This led to a novel AIS algorithm which efficiently adapts the parameters of a mixture distribution. We tested the new methodology on a high-dimensional multimodal target distribution and showed that it outperforms other state-of-the-art AIS methods.

7. REFERENCES

[1] J. S. Liu, *Monte Carlo strategies in scientific computing*, Springer Science & Business Media, 2008.

[2] J. V. Candy, *Bayesian signal processing: classical, modern, and particle filtering methods*, vol. 54, John Wiley & Sons, 2016.

[3] A. Gelman, H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*, Chapman and Hall/CRC, 2013.

[4] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[5] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[6] R. Ranganath, S. Gerrish, and D. M. Blei, "Black box variational inference," *arXiv preprint arXiv:1401.0118*, 2013.

[7] C. Robert and G. Casella, *Monte Carlo statistical methods*, Springer Science & Business Media, 2013.

[8] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, 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.

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

[10] O. Cappé, R. Douc, A. Guillin, J. Marin, and C. P. Robert, "Adaptive importance sampling in general mixture classes," *Statistics and Computing*, vol. 18, no. 4, pp. 447–459, 2008.

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

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

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

[14] L. Martino, V. Elvira, D. Luengo, and J. Corander, "Layered adaptive importance sampling," *Statistics and Computing*, vol. 27, no. 3, pp. 599–623, 2017.

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

[16] E. K. Ryu, *Convex optimization for Monte Carlo: Stochastic optimization for importance sampling*, Ph.D. thesis, Stanford University, 2016.

[17] J. Han and Q. Liu, "Stein variational adaptive importance sampling," *arXiv preprint arXiv:1704.05201*, 2017.

[18] C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt, "Advances in variational inference," *arXiv preprint arXiv:1711.05597*, 2017.

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

[20] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Improving population monte carlo: Alternative weighting and resampling schemes," *Signal Processing*, vol. 131, pp. 77–91, 2017.

[21] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.