Recursive Shrinkage Covariance Learning in Adaptive Importance Sampling

Yousef El-Laham⋆, Víctor Elvira‡, Mónica Bugallo⋆
⋆Department of Electrical & Computer Engineering, Stony Brook University, Stony Brook (USA)
‡IMT Lille Douai & CRIStAL laboratory, Villeneuve d'Ascq (France)

***Abstract*—The estimation of covariance matrices has been a central problem in a variety of disciplines, including quantitative finance, genomics, and signal processing. In Bayesian statistical inference, the efficiency of Monte Carlo methods, such as adaptive importance sampling (AIS), can be improved significantly if the distribution used to draw samples has a similar covariance structure to the posterior distribution of interest. Unfortunately, it is generally difficult to learn covariance matrices in high-dimensional settings due to the large number of samples needed for its appropriate estimation. This problem is intensified in the importance sampling context, where the usual weighted covariance estimators do not yield full rank estimates in most practical settings due to the weight degeneracy problem. In this work, we propose an AIS algorithm that robustly learns the covariance structure of the target distribution. The new method is based on applying shrinkage in a recursive manner, where the learned covariance matrix is constructed iteratively using a sequence of biased weighted covariance estimators. Simulation results indicate that the proposed method outperforms other state-of-the-art AIS methods, especially in the case where the number of samples drawn per iteration is relatively small.**

I. Introduction

Many problems of science and engineering require the computation of intractable integrals for estimating hidden parameters or inferring the probability density function (pdf) of those parameters. Monte Carlo (MC) methods use random samples for approximation of those intractable integrals and pdfs [1]. In the basic version, the MC technique simulates samples from the targeted pdf and provides an estimate by simply averaging the evaluation of those samples in a function of interest. However, in most practical cases, the simulation of samples from the target pdf is not possible, either because the pdf is not available in a closed form (e.g., due to an intractable normalizing constant) or because the distribution does not have a standard form where sampling is possible. In those scenarios, more advanced MC methods must be employed.

Importance sampling (IS) is a MC methodology that overcomes the above mentioned limitations by sampling from some *proposal* pdf [2]. Each sample receives an importance weight based on the mismatch between the target and the proposal pdfs. The key for a good performance in IS is in finding an adequate proposal pdf [3]. Many efforts have been devoted in the last two decades to developing adaptive IS (AIS) algorithms that can iteratively improve the proposal (see [4] for a survey). All AIS algorithms provide strategies for adapting the location parameter of the proposal, but only few of them adapt also the covariance matrix [5], [6], [7], [8]. The reason is the instability of the covariance estimator, especially in the early iterations of the algorithm where the mismatch between proposal and target pdfs is large and we encounter the weight degeneracy problem as a result of the *curse of dimensionality*.

In this work, we explore the use of shrinkage estimation [9], [10] to improve the performance of adaptive MC algorithms. Specifically, we incorporate a *recursive shrinkage* (RS) procedure to robustly adapt the proposal parameters. We prove that, under certain conditions, the proposed RS estimator is asymptotically unbiased. We also propose a gradual covariance learning approach that allows the method to achieve improved performance. When applied to AIS, the resulting mean and covariance estimators remain stable, even in the case that only a small number of samples are drawn per iteration. Numerical experiments reveal that the shrinkage-based AIS sampler outperforms other state-of-the-art methods, including our own previously proposed method [8], which adapts the covariance matrix using a nonlinear weight transformation to mitigate the effects of the weight degeneracy problem.

The remainder of the paper is organized as follows: in Section II we formulate the problem and in Section III, we provide a brief review of relevant prior work. We introduce the novel method in Section IV and validate the method in Section V through numerical simulations. Finally, we provide concluding remarks in Section VI.

II. Problem Formulation

We address the general problem of approximating the integral

$$I \triangleq \mathbb{E}\left[g(\boldsymbol{\theta})\right] = \int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{d_\theta}$ is a random vector with pdf $\pi(\boldsymbol{\theta})$, and $g(\boldsymbol{\theta})$ is a function that is integrable with respect to $\pi(\boldsymbol{\theta})$. In many applications, the difficulty is not only due to the intractability of (1) but also because $\pi(\boldsymbol{\theta})$ can be evaluated only up to an (unknown) normalizing constant Z. In this way, one only has access to $\tilde{\pi}(\boldsymbol{\theta}) = Z\pi(\boldsymbol{\theta})$, where $Z = \int \tilde{\pi}(\boldsymbol{\theta})d\boldsymbol{\theta}$ and $\tilde{\pi}(\boldsymbol{\theta})$ is the unnormalized non-negative function that can be evaluated. This is the case, for instance, in Bayesian inference, where $\tilde{\pi}(\boldsymbol{\theta})$ is the product between prior and likelihood, and Z is the marginal likelihood.

This work was carried out thanks to the support of the National Science Foundation (NSF) under Award CCF-1617986. The authors thank the Research Computing and Cyberinfrastructure and the Institute for Advanced Computational Science at Stony Brook University for access to the high-performance SeaWulf computing system sponsored by NSF (# 1531492).

III. Prior Work

A. Adaptive importance sampling (AIS)

A standard parametric AIS algorithm consists of the following iteratively applied steps: (1) *sampling*, (2) *weighting*, and (3) *adaptation*. In the sampling step, M samples are generated from a proposal distribution $q(\boldsymbol{\theta};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$, where i denotes the iteration index and $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ denote the proposal mean and covariance matrix, respectively. The generated samples $\{\boldsymbol{\theta}_i^{(m)}\}_{m=1}^{M} \overset{i.i.d.}{\sim} q(\boldsymbol{\theta};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ are then each assigned a corresponding importance weight

$$\tilde{w}_i^{(m)} = \frac{\tilde{\pi}(\boldsymbol{\theta}_i^{(m)})}{q(\boldsymbol{\theta}_i^{(m)};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}, \tag{2}$$

which can be normalized as $w_i^{(m)} = \frac{\tilde{w}_i^{(m)}}{\sum_{j=1}^{M}\tilde{w}_i^{(j)}}$ for $m = 1,\ldots,M$. The set of samples and normalized weights at each iteration form a discrete random measure $\mathscr{T}_i = \{\boldsymbol{\theta}_i^{(m)}, w_i^{(m)}\}_{m=1}^{M}$. Finally, at the end of each iteration, the proposal distribution is adapted using some adaptation rule. For instance, the mean of the proposal distribution can be adapted according to the weighted sample mean

$$\boldsymbol{\mu}_{i+1} = \sum_{m=1}^{M} w_i^{(m)}\boldsymbol{\theta}_i^{(m)}, \tag{3}$$

while the covariance matrix can be adapted according to the weighted empirical covariance

$$\boldsymbol{\Sigma}_{i+1} = \sum_{m=1}^{M} w_i^{(m)}(\boldsymbol{\theta}_i^{(m)} - \boldsymbol{\mu}_{i+1})(\boldsymbol{\theta}_i^{(m)} - \boldsymbol{\mu}_{i+1})^{\intercal}. \tag{4}$$

After I iterations, we obtain a set of samples and weights, which can be used to approximate the integral in (1) with

$$\hat{I}^{M,I}_{\text{AIS}} = \sum_{i=1}^{I}\sum_{m=1}^{M} \bar{w}_i^{(m)} g(\boldsymbol{\theta}_i^{(m)}), \tag{5}$$

and approximate the target distribution $\pi(\boldsymbol{\theta})$ with,

$$\hat{\pi}^{M,I}_{\text{AIS}}(\boldsymbol{\theta}) = \sum_{i=1}^{I}\sum_{m=1}^{M} \bar{w}_i^{(m)} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_i^{(m)}), \tag{6}$$

where $\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_i^{(m)})$ denotes a Dirac-delta function centered at $\boldsymbol{\theta}_i^{(m)}$ and $\bar{w}_i^{(m)} = \frac{\tilde{w}_i^{(m)}}{\sum_{i=1}^{I}\sum_{j=1}^{M}\tilde{w}_i^{(j)}}$ denotes the weight of the mth sample generated in the ith iteration normalized over all iterations. We summarize the approach in Algorithm 1.

Algorithm 1 Standard Parametric AIS

```
Initialization: Set the initial proposal parameters $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1$.
for $i = 1, \ldots, I$ do
    Draw $M$ samples from the proposal distribution,
        $\boldsymbol{\theta}_i^{(m)} \sim q(\boldsymbol{\theta};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i), \quad m = 1,\ldots,M.$
    Compute the importance weights of the samples,
        $\tilde{w}_i^{(m)} = \frac{\tilde{\pi}(\boldsymbol{\theta}_i^{(m)})}{q(\boldsymbol{\theta}_i^{(m)};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}, \quad m = 1,\ldots,M.$
    Compute $\boldsymbol{\mu}_{i+1}$ and $\boldsymbol{\Sigma}_{i+1}$ according to (3) and (4).
end for
```

B. Shrinkage in covariance estimation

Let $\widehat{\boldsymbol{\Sigma}}$ be a consistent estimator of the target covariance matrix, $\boldsymbol{\Sigma}$. Suppose that there also exists another estimator of the covariance $\widetilde{\boldsymbol{\Sigma}}$, which assumes a specific covariance structure. The estimator $\widetilde{\boldsymbol{\Sigma}}$ is biased, but is more "stable" than $\widehat{\boldsymbol{\Sigma}}$ for a smaller sample size. The shrinkage estimator of the target covariance matrix, $\widehat{\boldsymbol{\Sigma}}_{\beta}$, is given as

$$\widehat{\boldsymbol{\Sigma}}_{\beta} = (1-\beta)\widetilde{\boldsymbol{\Sigma}} + \beta\widehat{\boldsymbol{\Sigma}}, \tag{7}$$

where $0 \leq \beta \leq 1$. The biased estimator $\widetilde{\boldsymbol{\Sigma}}$ captures the simplified structure of the covariance matrix, while the estimator for the full covariance matrix $\widehat{\boldsymbol{\Sigma}}$ captures the correlations encoded by the samples. The shrinkage weight parameter β controls the bias-variance tradeoff, and is typically chosen to optimize some objective function $f(\beta)$, such as the mean squared error (MSE) of the shrinkage estimator.

IV. Proposed Method

Recall from Section III-A that the mean of the proposal distribution can be adapted according to the weighted sample mean as defined in (3) and the covariance matrix can be adapted according to (4). The weighted sample mean is an unbiased estimator for the mean, while the weighted empirical covariance is biased for a finite sample size. It is possible to form an unbiased estimator of the covariance matrix using weighted samples by applying Bessel's correction. This unbiased weighted covariance matrix is given as follows:

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{W_i}\sum_{m=1}^{M} w_i^{(m)}(\boldsymbol{\theta}_i^{(m)} - \boldsymbol{\mu}_{i+1})(\boldsymbol{\theta}_i^{(m)} - \boldsymbol{\mu}_{i+1})^{\intercal}, \tag{8}$$

where $W_i = 1 - \sum_{j=1}^{M}(w_i^{(m)})^2$. While the estimators in (3) and (8) are both unbiased, when d_θ is large, the weighted sample covariance matrix can be ill-conditioned due to the limited sample size M and the weight degeneracy in the early iterations of the algorithm.

One way to overcome the challenges posed by high dimensionality is to consider a parameter adaptation strategy based on applying the shrinkage principle recursively. In particular, we adapt the mean and covariance matrix as

$$\begin{aligned}\boldsymbol{\mu}_{i+1} &= (1-\alpha_i)\boldsymbol{\mu}_i + \alpha_i\hat{\boldsymbol{\mu}}_i \\ \boldsymbol{\Sigma}_{i+1} &= (1-\beta_i)\boldsymbol{\Sigma}_i + \beta_i\widehat{\boldsymbol{\Sigma}}_i\end{aligned} \tag{9}$$

where $0 < \alpha_i, \beta_i \leq 1$. We define $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ as the initial mean and covariance matrix, respectively. We note that $\boldsymbol{\Sigma}_1$ should be chosen to be well-conditioned and positive definite. The benefit of using the proposed adaptation scheme in (9) is that the resulting parameter adaptations remain stable. These estimators resemble adaptation schemes proposed in the adaptive MCMC literature in [11] and [12], where moment estimators like $\hat{\boldsymbol{\mu}}_i$

and $\hat{\mathbf{\Sigma}}_i$ are computed using samples from a Markov chain whose stationary distribution is the target. This is in contrast to our work, which uses unbiased estimators of the moments to adapt the parameters.

In the following, we show that, in the limit as $i \to \infty$, the estimators in (9) are unbiased. We show this result for the recursive covariance estimator, but it also follows for the mean. We remark that if the sequence $\beta_1, \beta_2, \dots, \beta_i$ are chosen according to the Robbins-Monro stochastic approximation conditions [13], the covariance of the proposal converges to the covariance of the target distribution as $i \to \infty$.

A. Asymptotic unbiasedness

Proposition 1. *Let $\beta_1, \beta_2, \dots, \beta_i$ be a sequence of constants that satisfy $0 < \beta_j \leq 1$, and let $\widehat{\mathbf{\Sigma}}_1, \widehat{\mathbf{\Sigma}}_2, \dots, \widehat{\mathbf{\Sigma}}_i$ be a sequence of unbiased estimators of the covariance matrix $\mathbf{\Sigma}$. For any positive definite matrix $\mathbf{\Sigma}_1$, the recursive shrinkage estimator*

$$\mathbf{\Sigma}_{i+1} = (1-\beta_i)\mathbf{\Sigma}_i + \beta_i \widehat{\mathbf{\Sigma}}_i$$

is asymptotically unbiased.

Proof. For $i = 1$, $\mathbf{\Sigma}_2 = (1-\beta_1)\mathbf{\Sigma}_1 + \beta_1 \widehat{\mathbf{\Sigma}}_1$. For $i = 2$,

$$\begin{aligned}\mathbf{\Sigma}_3 &= (1-\beta_2)\mathbf{\Sigma}_2 + \beta_2 \widehat{\mathbf{\Sigma}}_2 \\ &= (1-\beta_1)(1-\beta_2)\mathbf{\Sigma}_1 + \beta_1(1-\beta_2)\widehat{\mathbf{\Sigma}}_1 + \beta_2 \widehat{\mathbf{\Sigma}}_2\end{aligned}$$

For any i, $\mathbf{\Sigma}_{i+1}$ can be expressed as

$$\mathbf{\Sigma}_{i+1} = a_i \mathbf{\Sigma}_1 + \sum_{j=1}^{i} \gamma_{i,j} \widehat{\mathbf{\Sigma}}_j \tag{10}$$

where $a_i = \prod_{j=1}^{i}(1-\beta_j)$ and $\gamma_{i,j} = \beta_j \prod_{\tau=j+1}^{i}(1-\beta_\tau)$ for $j = 1, \dots, i-1$ and $\gamma_{i,i} = \beta_i$. We note that $a_i + \sum_{j=1}^{i} \gamma_{i,j} = 1$ always. Taking the expectation of both sides, we have that

$$\begin{aligned}\mathbb{E}\left[\mathbf{\Sigma}_{i+1}\right] &= \mathbb{E}\left[a_i \mathbf{\Sigma}_1 + \sum_{j=1}^{i} \gamma_{i,j} \widehat{\mathbf{\Sigma}}_j\right] \\ &= a_i \mathbf{\Sigma}_1 + \sum_{j=1}^{i} \gamma_{i,j} \mathbb{E}\left[\widehat{\mathbf{\Sigma}}_j\right] = a_i \mathbf{\Sigma}_1 + \left(\sum_{j=1}^{i} \gamma_{i,j}\right) \mathbf{\Sigma}\end{aligned}$$

We would like to show that $\lim_{i\to\infty} \mathbb{E}\left[\mathbf{\Sigma}_{i+1}\right] = \mathbf{\Sigma}$. First

$$\lim_{i\to\infty} a_i = \lim_{i\to\infty} \prod_{j=1}^{i}(1-\beta_j) = 0$$

since $0 \leq (1-\beta_j) < 1$ for all j. This implies that $\lim_{i\to\infty} \sum_{j=1}^{i} \gamma_{i,j} = 1$ and $\lim_{i\to\infty} \mathbb{E}\left[\mathbf{\Sigma}_{i+1}\right] = \mathbf{\Sigma}$. □

Intuitively, as the number of iterations gets large, the contribution of $\mathbf{\Sigma}_1$ in the recursive shrinkage estimator diminishes. Note that the result does not hold with consistent estimators that are biased for a finite sample size, e.g. the weighted empirical covariance matrix in (4); however, even in this situation, the contribution of the initial covariance matrix $\mathbf{\Sigma}_1$ goes to zero asymptotically with the number of iterations.

B. Gradual Covariance Learning

In the early iterations of an AIS algorithm, the weighted covariance estimates are poor due to the weight degeneracy problem. This can have a negative effect on the performance of the covariance update rule in (9). We overcome this issue by utilizing the following update for the covariance matrix:

$$\mathbf{\Sigma}_{i+1} = (1-\beta_i)\mathbf{\Sigma}_i + \beta_i(1-\eta_i)\widehat{\mathbf{\Sigma}}_i + \beta_i \eta_i \tilde{\mathbf{\Sigma}}_i, \tag{11}$$

where $\tilde{\mathbf{\Sigma}}_i$ is an estimator of an intermediate covariance target, and $\eta_1, \dots, \eta_i$ is a decreasing sequence of constants that satisfies: $\eta_1 = 1$, and $\lim_{i\to\infty} \eta_i = 0$. Intuitively, the intermediate covariance estimate is more stable than the target covariance estimate in the early iterations of AIS, and thus makes the update of the proposal covariance more robust in high-dimensional settings. In this work, we choose the intermediate covariance estimate to be the one proposed in our previous work [8], which estimates the covariance using a nonlinear transformation of the importance weights [14].

V. Example: Bayesian Linear Regression

We validate the proposed shrinkage-based adaptation technique with a simple example of Bayesian linear regression [15]. Suppose we observe a vector of N observations $\mathbf{y} = [y_1, \dots, y_N]^\intercal \in \mathbb{R}^N$, where each $y_n \in \mathbb{R}$ can be written as

$$y_n = \mathbf{x}_n^\intercal \boldsymbol{\theta} + u_n, \tag{12}$$

where $\mathbf{x}_n \in \mathbb{R}^{d_\theta}$ is a vector of features, $\boldsymbol{\theta} \in \mathbb{R}^{d_\theta}$ is a vector of weight coefficients, and $u_n \in \mathbb{R}$ is a zero-mean additive Gaussian noise with unity variance. Equivalently, we write $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{u}$, where $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^\intercal$ and $\mathbf{u} = [u_1, \dots, u_N]^\intercal$.

Under the Bayesian framework, our goal is to learn the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$, which can be evaluated up to a normalization constant accordingly:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto \tilde{\pi}(\boldsymbol{\theta}) \triangleq p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{13}$$

where $p(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}; \mathbf{X}\boldsymbol{\theta}, \mathbb{I}_{d_y})$ is the likelihood function of the parameters and $p(\boldsymbol{\theta})$ is the prior distribution. Under the choice $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}, \frac{1}{5}\mathbb{I}_{d_\theta})$, the posterior distribution can be determined analytically in closed form. For this reason, we evaluate our method on this example, since we can compare to the ground truth posterior. We simulate a synthetic data set with $N = 20$ and $d_\theta = 10$, where each feature vector $\mathbf{x}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_x)$ with $\mathbf{\Sigma}_x$ being a non-identity covariance matrix and the ground truth $\boldsymbol{\theta} = [\theta_1, \dots, \theta_{10}]^\intercal$ generated according to $\theta_i \sim \mathcal{U}(0,1)$ for $i = 1, \dots, 10$. For the proposed scheme, we fix $\alpha_i = 1$ and test four variants for adapting the covariance matrix according to (11): 1) Constant ($\beta_i = \beta_1$ and $\eta_i = 0$); 2) Decreasing ($\beta_i = (\beta_1)^{-0.5}$ and $\eta_i = 0$); 3) Constant/Gradual ($\beta_i = \beta_1$ and $\eta_i = i^{-1}$); and 4) Decreasing/Gradual ($\beta_i = (\beta_1)^{-0.5}$ and $\eta_i = i^{-1}$). For the initial shrinkage parameter, we test $\beta_1 \in \{0.1, 0.2, \dots, 0.9\}$.

We run two experiments to test the efficiency of the proposed scheme. In the first experiment, we test different sample sizes $M = \{100, 200, 500\}$ and run each algorithm for $I = 10^5/M$ iterations. The mean is initialized as $\boldsymbol{\mu}_1 \sim \mathcal{U}([-5, 5]^{10})$ and

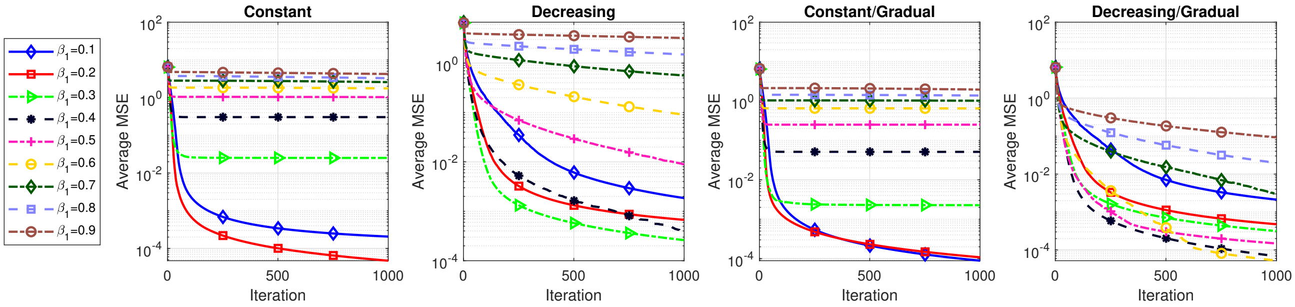

Fig. 1: Evolution of average MSE for proposed scheme with $M = 100$ and under different settings for β_1.

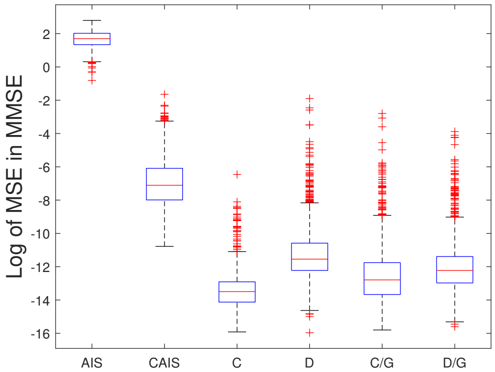

Fig. 2: MSE of the posterior mean for each method.

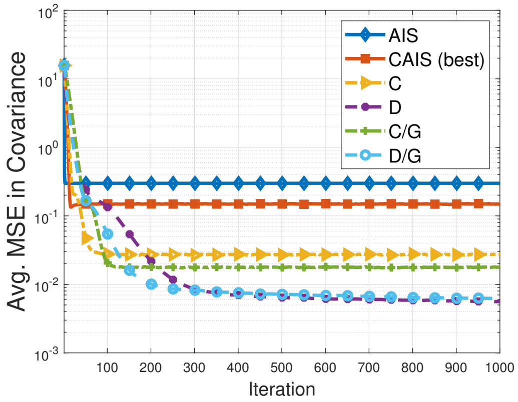

Fig. 3: Evolution of average MSE in covariance estimation.

the covariance matrix is initialized as $\mathbf{\Sigma}_1 = 5\mathbb{I}_{10}$. The results are averaged over 1000 runs. Figure 1 plots the evolution of the average MSE of the posterior mean estimate for different values of β_1. Based on the figure, it is clear that the gradual covariance learning approach is the most robust, as it achieves good performance regardless of the choice of β_1. This is verified in Table I, which shows performance averaged over the value of β_1 under different sample size settings. The results indicate that the gradual covariance learning approaches (both constant and decreasing β_i) achieve lower average MSE than the standard shrinkage approach.

	Constant	Decreasing	Const./Grad.	Decr./Grad.
$M = 100$	1.470 ± 1.591	0.598 ± 1.103	0.589 ± 0.717	$\mathbf{0.013 \pm 0.031}$
$M = 200$	1.094 ± 1.400	0.397 ± 0.785	0.239 ± 0.368	$\mathbf{0.002 \pm 0.003}$
$M = 500$	0.670 ± 1.100	0.281 ± 0.597	0.043 ± 0.091	$\mathbf{0.006 \pm 0.014}$

TABLE I: Average MSE of the posterior mean averaged over the tested values of β_1 (with standard errors).

In the second experiment, we compare the proposed method to the standard parametric AIS in Algorithm 1 and the covariance AIS (CAIS) algorithm in [8]. We fix $M = 100$ and $I = 1000$ and use the same proposal parameter initialization as the previous experiment. We fix β_1 as follows for each the proposed method variants: Constant ($\beta_1 = 0.2$), Decreasing ($\beta_1 = 0.3$), Constant/Gradual ($\beta_1 = 0.1$), and Decreasing/Gradual ($\beta_1 = 0.4$). The results of the experiment are averaged over 1000 runs and are shown in Figs. 2 and 3. Figure 2 shows a boxplot that compares the MSE in the estimates of the posterior mean. The results indicate that the novel method always outperforms the standard AIS and for the selected choices of β_1, it outperforms CAIS under its best configuration. This shows that our method is more robust than methods like CAIS in the case that the number of samples per iterate is small relative to the dimension of the target distribution. Figure 3 shows the evolution of the distance between the target covariance and proposal covariance for each of the AIS methods (measured with the Frobenius norm). The plot indicates that the adapted covariance in standard parametric AIS and CAIS are further from the target covariance than the proposed schemes. Furthermore, it is evident that, when a gradual covariance learning strategy is used, the proposal covariance approaches the target covariance at the fastest rate.

VI. Conclusions and Future Work

In this work, we introduced a proposal parameter adaptation scheme for general adaptive Monte Carlo methods. The scheme is based on applying shrinkage recursively, which allows for stable parameter updates. We also presented a variant of the method that gradually learns the target covariance by approximating the covariance of an intermediate target. The adaptation scheme was applied to standard parametric adaptive importance sampling (AIS). Numerical experiments showed that the shrinkage-based AIS methods outperformed other competing algorithms in practice.

References

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

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

[3] A. B. Owen, *Monte Carlo theory, methods and examples*, 2013.

[4] 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.

[5] 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.

[6] J. Marin, P. Pudlo, and M. Sedki, "Consistency of the adaptive multiple importance sampling," *arXiv preprint arXiv:1211.2548*, 2012.

[7] 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.

[8] Y. El-Laham, V. Elvira, and M. F. Bugallo, "Robust covariance adaptation in adaptive importance sampling," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 1049–1053, 2018.

[9] O. Ledoit and M. Wolf, "Honey, i shrunk the sample covariance matrix," *The Journal of Portfolio Management*, vol. 30, no. 4, pp. 110–119, 2004.

[10] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, 2005.

[11] 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.

[12] H. Haario, E. Saksman, J. Tamminen, et al., "An adaptive metropolis algorithm," *Bernoulli*, vol. 7, no. 2, pp. 223–242, 2001.

[13] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[14] E. Koblents and J. Míguez, "A population monte carlo scheme with transformed weights and its application to stochastic kinetic models," *Statistics and Computing*, vol. 25, no. 2, pp. 407–425, 2015.

[15] D. V. Lindley and A. F. M. Smith, "Bayes estimates for the linear model," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 1, pp. 1–18, 1972.