IMPROVED ADAPTIVE IMPORTANCE SAMPLING BASED ON VARIATIONAL INFERENCE

Matthew Dowling, Josue Nassar*, Petar M. Djurić, and Mónica F. Bugallo* †

Department of Electrical & Computer Engineering
Stony Brook University, Stony Brook, NY 11794-2350
{matthew.dowling,josue.nassar,petar.djuric,monica.bugallo}@stonybrook.edu

ABSTRACT

In Monte Carlo-based Bayesian inference, it is important to generate samples from a target distribution, which are then used, e.g., to compute expectations with respect to the target distribution. Quite often, the target distribution is the posterior of parameters of interest, and drawing samples from it can be exceedingly difficult. Monte Carlo-based methods, like adaptive importance sampling (AIS), is built on the *importance sampling principle* to approximate a target distribution using a set of samples and their corresponding weights. Variational inference (VI) attempts to approximate the posterior by minimizing the Kullback-Leibler divergence (KLD) between the posterior and a set of simpler parametric distributions. While AIS often performs well, it struggles to approximate multimodal distributions and suffers when applied to high dimensional problems. By contrast, VI is fast and scales well with the dimension, but typically underestimates the variance of the target distribution. In this paper, we combine both methods to overcome their individual drawbacks and create an efficient and robust novel technique for drawing better samples from a target distribution. Our contribution is two-fold. First, we show how to do a smart initialization of AIS using VI. Second, we propose a method for adapting the parameters of the proposal distributions of the AIS, where the adaptation depends on the performance of the VI step. Computer simulations reveal that the new method improves the performance of the individual methods and shows promise to be applied to challenging scenarios.

Index Terms— Adaptive importance sampling, Markov chain Monte Carlo, variational inference, Bayesian inference

1. INTRODUCTION

In a Bayesian setting, all relevant information about the parameters of interest is contained in the posterior distribution. In practice, the posterior distribution is used to find expectations of interest, which is usually difficult, if not impossible, to express in closed form. Thus, a possible alternative consists of approximating such expectations through numerical methods, and in particular, Monte Carlo-based methods.

One of the most popular Monte Carlo sampling methods are the Markov chain Monte Carlo (MCMC) methods, which construct a Markov chain whose stationary distribution is the posterior [1]. Once the Markov chain converges, the drawn samples by the chain are considered to come from the target distribution. This allows for sample-based approximation of the posterior. However, not all the samples are used towards the approximation, which may lead to a considerable computational inefficiency.

Recently, adaptive importance sampling (AIS) has emerged as a viable alternative to MCMC. It builds upon the principle of *importance sampling,* where a set of proposal functions are iteratively updated and are used to generate samples with assigned weights. In contrast to MCMC, all the samples and corresponding weights are used to form an approximation of the posterior. There is also no need for a burn-in period when implementing AIS. However, AIS suffers when the posterior is characterized as multimodal and/or the problem is of high dimensionality. There have been attempts to overcome some of these challenges through various AIS implementations, which rely on different weight calculations and proposal updates such as AMIS (Adaptive Multiple Importance Sampling) [2], DM-PMC (Deterministic Mixture Population Monte Carlo) [3], M-PMC (Mixture Population Monte Carlo) [4], and others [5].

Variational inference (VI) is another popular method, primarily used in the machine learning community, for approximating posteriors. Contrary to Monte Carlo methods, which use sampling, VI is based on optimization to approximate the posterior [6], [7]. VI aims to minimize the Kullback-Leibler divergence (KLD) between the posterior and a family of probability distributions [8]. However, it can often be hard to find closed form expressions for updating the equations involved in the algorithm [9]. In addition, since the final variational distributions come from a known family, they may not fully resemble the targeted posterior.

In this paper, we propose a novel approach to AIS, where VI is used to obtain a coarse estimate of the posterior, which is then used to initialize AIS. This allows AIS to take advantage

*These two authors contributed equally

†The authors thank the support of the National Science Foundation under Award CCF-1618999 (P. M. Djurić) and under Award CCF-1617986 (M. F. Bugallo).

ISBN 978-90-827970-1-5 © EURASIP 2018

of the estimate supplied by the VI algorithm and to provide a more accurate estimate of the posterior that will also more closely resemble it. We also propose a variance adaptation technique for the proposal distributions that is robust to the performance of VI. Using AIS alone we run into the aforementioned problem of many modes, while using VI alone, the variational approximation may not closely resemble the true posterior. By combining both methods, we take advantage of the positive attributes of both methodologies. Computer simulations show that the proposed method results in an approximation that more closely resembles the target distribution than the methods can achieve on their own.

The remainder of the paper is structured as follows. In Section 2, we review the theory of AIS and VI. In Section 3, we discuss both the proposed method and the variance adaptation. In the following section, we present results of a series of experiments that compare the performance of the proposed method with some state-of-the-art methods. In Section 5, we provide concluding remarks and discuss future work.

2. REVIEW

2.1. Preliminaries

We consider problems where the challenge is to estimate some target distribution $\pi(x)$, where $x \in \mathbb{R}^d$. In the Bayesian framework, this target distribution typically represents a posterior distribution of the parameters of interest. Both AIS and VI can give us an approximation of this target. The approximation via AIS is provided by a series of samples and their corresponding weights. On the other hand, the approximation given through VI is provided through a series of parameterized distributions from known families.

2.2. Adaptive Importance Sampling

AIS uses the principle of *importance sampling* to iteratively improve the proposal distribution, allowing for a better approximation of the target density, $\pi(x)$. At iteration 0, we draw M samples from a proposal distribution, $q_0(x)$, and form an approximation of the target

$$\hat{\pi}_0(x) = \sum_{m=1}^{M} w_0^{(m)} \delta\left(x_0^{(m)}\right), \tag{1}$$

where $w_0^{(m)} = \frac{\bar{w}_0^{(m)}}{\sum_{j=1}^{M} \bar{w}_0^{(j)}}$ represents the normalized weight of the mth generated sample, $x_0^{(m)}$, with $\bar{w}_0^{(m)} = \frac{\pi\left(x_0^{(m)}\right)}{q_0\left(x_0^{(m)}\right)}$. The random measure $\chi_0 = \left\{x_0^{(m)}, w_0^{(m)}\right\}_{m=1}^{M}$ can be used to construct a better proposal, $q_1(x)$ [10]. Then, M samples are drawn from $q_1(x)$ and their weights are properly calculated. The samples and weights from these first two iterations can be combined to obtain a better approximation of the target [11]. Moreover, this new approximation is used to construct an even better proposal distribution for the next iteration, $q_2(x)$. The procedure continues iteratively, and at iteration i, the proposal distribution $q_i(x)$ is adapted using the past random measures.

One popular implementation of the AIS methodology is the population Monte Carlo (PMC) algorithm [12]. In PMC, a set of proposals is used to generate samples, i.e., $x_i^{(m)} \sim q_i^{(m)}\left(x|\mu_i^{(m)}, \Sigma_i^{(m)}\right)$ for $m = 1, \cdots, M$ and $i = 1, \cdots, I$. The proposal distributions are updated by adapting their location parameters $\mu_i^{(m)}$, while keeping $\Sigma_i^{(m)}$ fixed, i.e., $\Sigma_i^{(m)} = \Sigma$, $\forall i, m$. The adaptation is performed using a resampling method [3]. Here, for simplicity, we use multinomial resampling. The PMC implementation is summarized in Algorithm 1.

Algorithm 1 PMC

Select the adaptive parameters $\mu_1^{(m)}, \quad m = 1, \cdots, M$
Select the static parameter Σ

1: **for** $i = 1, \cdots, I$ **do**

2: Generate samples:
draw $x_i^{(m)} \sim q(x_i^{(m)}|\mu_i^{(m)}, \Sigma) \quad m = 1, \cdots, M$

3: Update the weights:
$\bar{w}_i^{(m)} = \frac{\pi(x_i^{(m)})}{q(x_i^{(m)}|\mu_i^{(m)}, \Sigma)} \quad m = 1, \cdots, M$

4: Normalize the weights:
$w_i^{(m)} = \frac{\bar{w}^{(m)}}{\sum_{j=1}^{M} \bar{w}^{(j)}} \quad m = 1, \cdots, M$

5: Perform resampling using $\chi_i = \left\{x_i^{(m)}, w_i^{(m)}\right\}$
to select $\mu_{i+1}^{(m)}, \quad m = 1, \cdots, M$

6: **end for**

7: Recalculate the weights

8: $\rho_i^{(m)} = \frac{\bar{w}_i^{(m)}}{\sum_{i=1}^{I}\sum_{m=1}^{M} \bar{w}_i^{(m)}}, \; m = 1, \cdots, M, \; i = 1, \cdots, I$

9: $\hat{\pi}(\mathbf{x}) = \sum_{i=1}^{I}\sum_{m=1}^{M} \rho_i^{(m)} \delta(\mathbf{x}_i^{(m)})$

2.3. Variational Inference

Let us define the set of tractable distributions as $\mathcal{D}$. The goal of VI is to select an element, $q(x) \in \mathcal{D}$, which provides the best approximation to the target, $\pi(x)$. This is done by recasting the optimization problem that selects the distribution which minimizes its KLD with the target [13]:

$$q^*(x) = \arg\max_{q(x)\in\mathcal{D}} \left\{\text{KLD}\left(q(x)||\pi(x)\right)\right\}. \tag{2}$$

ISBN 978-90-827970-1-5

Often the mean-field approximation methodology is followed [9], which assumes that $q^*(x)$ can be factored as

$$q(x) = \prod_{k=1}^{K} q_k^*(x_k). \tag{3}$$

This allows for the optimization of K functions individually instead of just one, which eases computation. It can be shown that the optimal variational distribution takes the form

$$q_k^*(x_k) \propto \exp(E_{-q_k^*}(\log(\pi(x)))), \tag{4}$$

where the expectation is taken with respect to the variational distributions, not including $q_k^*(x)$. Solving for the variational distributions as above will result in a series of equations that allow for the updates of the parameters of the distributions, assuming the distributions come from a known family. The algorithm runs either for a fixed number of iterations or until the difference between consecutive updates drops below a threshold. The standard VI implementation is summarized in Algorithm 2.

Algorithm 2 Standard VI Algorithm

Randomly initialize the parameters of the distributions $q_k(\mathbf{x}_k) \quad k = 1, \cdots, K$

for $i = 1, \cdots, I$ **do**

  for $k = 1, \cdots, K$ **do**

    Optimize $q_k(x_k)$ fixing $q_{-k}(x_{-k})$

  end for

end for

$\hat{\pi}(x) = \prod_{k=1}^{K} q_k(x_k)$

3. PROPOSED METHOD

In our proposed method, we aim to combine the strengths of both methods. Taking advantage of the low computational complexity of VI, we can have a quick and efficient search of high probability state spaces. This serves to circumvent the problem of initialization of AIS algorithms. Poor initialization will affect proper exploration by the proposals and will lead to poor results. The initialization becomes increasingly important when many modes are present in order to avoid locking into few modes.

More precisely, we propose to run the VI for a fixed number of iterations. The modes of the variational distributions $\{\lambda_1, \cdots, \lambda_K\}$ are then used to initialize the location parameters for AIS. For instance, using M proposals after estimating N modes of the target, we can set the adaptive location paramters as $\mu_{1:\lfloor M/K \rfloor} = \lambda_1, \cdots, \mu_{M-\lfloor M/K \rfloor:M} = \lambda_K$, where $\lfloor \cdot \rfloor$ represents the floor function. In this manner an equal number of samples are generated around each of the discovered modes.

A key part of the proposed method is that the AIS step adapts depending on the performance of the VI step. For example, if the proposals are Gaussians and the static parameter Σ represents a fixed covariance matrix for the M proposals, we would like that if the VI step did well, we stay within those modes, and therefore we use small covariance matrices that do not encourage far reaching exploration. However, if the VI step did not perform well, we should like that the initial proposals do explore far outside the VI initialization and the selection of Σ encourages that exploration. We account for this by selecting a set of possible static parameters $\{\Sigma_1, \cdots, \Sigma_L\}$ that influence the exploration of the space. Running L parallel AIS algorithms for I iterations each with a different Σ_l and the parameters passed from the VI step, the Σ_l that produces the highest average likelihood is chosen to run for a full number of iterations. That is, we choose the static parameters as follows:

$$\Sigma = \underset{\Sigma_l,\, l \in \{1,\cdots,L\}}{\arg\max} \left\{ \frac{1}{IM} \sum_{i=1}^{I} \sum_{m=1}^{M} \pi(x_i^{(m)}) \right\}. \tag{5}$$

This adaptation step, using a sufficiently diverse set of values $\{\Sigma_1, \cdots, \Sigma_L\}$, will ensure proper exploration of the space. The newly proposed method is outlined as Algorithm 3.

Algorithm 3 Proposed Method

1: Select the possible static parameters $\{\Sigma_1, \cdots, \Sigma_L\}$ for testing

2: Run the VI algorithm to obtain $\lambda_1, \cdots, \lambda_K$

3: **for** $l = 1, \cdots, L$ **do**

4: Run AIS with Σ_l as the static parameter of the proposals, and adaptive parameters set as

$$\mu_{1:\lfloor M/K \rfloor} = \lambda_1, \cdots, \mu_{M-\lfloor M/K \rfloor:M} = \lambda_K$$

5: **end for**

6: Choose Σ, the most favorable static parameter, according to equation (5)

7: Continue with the AIS which used the optimal static parameter for the remaining iterations.

ISBN 978-90-827970-1-5 © EURASIP 2018

4. NUMERICAL RESULTS

4.1. The problem

Consider N samples drawn from a mixture of K Gaussians. Specifically let

$$x = \{\mu_1, \cdots, \mu_K\}, \tag{6}$$

$$y_n|x \sim \sum_{k=1}^{K} \frac{1}{K}\mathcal{N}(y_n|\mu_k, \Sigma_X), \quad n = 1, \cdots, N, \tag{7}$$

where $\mu_k, y_n \in \mathbb{R}^d$ for $\forall n, k$, y_n represents an observed sample, Σ_X is a known covariance matrix, and I_d is the $d \times d$ identity matrix. The goal is to estimate x, the set of means. The posterior can be written as

$$f(x|\boldsymbol{Y}) = \frac{\prod_{n=1}^{N}\sum_{k=1}^{K}\frac{1}{K}\mathcal{N}(y_n|\mu_k, \Sigma_X)\ p(x)}{f(\mathbf{Y})}, \tag{8}$$

where the prior placed on the means is $p(x) = \prod_{k=1}^{K}\mathcal{N}(\mu_0, \Sigma_0)$, and $\boldsymbol{Y} = \{y_1, \ldots, y_N\}$. Due to the complexity of the model, it is not trivial to estimate x. We solve the problem using standard AIS, standard VI, and the proposed method.

4.2. Variational Inference Updates

Standard AIS is straightforward to implement but more work is required to obtain the variational inference implementation. To ease the derivation, we add a latent variable, $z_n \in \mathbb{R}^K$, a 1-hot vector with a 1 in the k^{th} position if y_n comes from the mixand k. The joint distribution of all the variables is then

$$f(x, \mathbf{Z}, \mathbf{Y}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\frac{1}{K}\mathcal{N}(y_n|\mu_k, \Sigma_X)^{z_{n_k}}\ p(x)p(\mathbf{Z}), \tag{9}$$

where z_{n_k} is the kth element of z_n, $\mathbf{Z} = \{z_1, \cdots, z_N\}$, and $p(\boldsymbol{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K}(\frac{1}{K})^{z_{n_k}}$. Under the Mean Field approximation, the variational distributions can be factored as $q(x, \mathbf{Z}) = q(\mathbf{Z})\prod_{k=1}^{K} q(\mu_k)$. It can be shown that

$$q(\mu_k) = \mathcal{N}(\lambda_k, \Sigma_k), \tag{10}$$

$$\Sigma_k = \left(\Sigma_0^{-1} + \sum_{n=1}^{n} E[z_{n_k}]\,\Sigma_X^{-1}\right)^{-1}, \tag{11}$$

$$\lambda_k = \Sigma_k\left(\Sigma_X^{-1}\sum_{n=1}^{N} y_n E[z_{n_k}] + \Sigma_0^{-1}\mu_0\right), \tag{12}$$

and

$$q(\mathbf{Z}) = \prod\nolimits_{n=1}^{N}\prod\nolimits_{k=1}^{K}\Omega_{n_k}^{z_{n_k}}, \tag{13}$$

where for normalization purposes,

$$\Omega_{n_k} = \frac{\omega_{n_k}}{\sum_j \omega_{n_j}}, \tag{14}$$

and

$$\log\omega_{n_k} = -\frac{1}{2}\Big(y_n^T\Sigma_X^{-1}y_n - 2y_n^T\Sigma_X^{-1}E[\mu_k] \tag{15}$$

$$+ \operatorname{tr}\left(\Sigma_X^{-1}\operatorname{var}(\mu_k) + E[\mu_k]^T\Sigma_X^{-1}E[\mu_k]\right)\Big). \tag{16}$$

Next we present the results of three experiments.

4.3. Experiment 1

Let the mean components of the mixture be

$$x = [-40, -30, -20, -10, 0, 10, 20, 30, 40]^\top$$

and $N = 500$ data points. PMC and VI are both run for $I_{PMC} = I_{VI} = 250$ iterations while the proposed method, called VI-PMC in Table 1, uses 125 iterations for each step. For the standalone PMC, we generated $M = 250$ samples, and set the variance of the proposal densities to be 10. For the PMC part of the proposed method, we set $M = 250$ and the variance of the proposals comes from $\sigma^2 \in \{9/10, 9\}$. We used Gaussian proposals and the prior $\prod_{k=1}^{K}\mathcal{N}(0, 50)$. In Table 1, the MSE was obtained by averaging over 100 realizations for each method. The MSEs were obtained for various values of $\Sigma_X = \sigma_X^2$, the variance of the data. The results show that the proposed method had the best performance.

Table 1. Experiment 1 Results

Method	$\sigma_X^2 = 0.5$	$\sigma_X^2 = 1$	$\sigma_X^2 = 4$	$\sigma_X^2 = 9$
PMC	30.68	34.11	27.81	41.92
VI	149.6	101.38	61.32	0.6628
VI-PMC	6.6	4.26	5.0054	0.1426

4.4. Experiment 2

In this experiment, we considered a multivariate case, where the mean components of the mixture were $\mu_1 = [-40; -40]$, $\mu_2 = [-30; -30]$, $\mu_3 = [-20; -20]$, $\mu_4 = [-10; -10]$, $\mu_5 = [0; 0]$, $\mu_6 = [10; 10]$, $\mu_7 = [20; 20]$, $\mu_8 = [30; 30]$, $\mu_9 = [40; 40]$. All the simulation parameters were the same as before except that the variance of the proposals for the standalone PMC were $10I_2$. For the PMC part of the proposed method, the variance of the proposals were $\Sigma = \sigma^2 I$, where $\sigma^2 \in \{9/10, 9\}$. The results are presented in Table 2. Again, the proposed method had the best performance.

ISBN 978-90-827970-1-5 © EURASIP 2018

Table 2. Experiment 2 Results

Method	$\sigma_X^2 = 0.5$	$\sigma_X^2 = 1$	$\sigma_X^2 = 4$	$\sigma_X^2 = 9$
PMC	259.68	191.46	68.88	32.71
VI	316.17	295.26	126.92	56.75
VI-PMC	43.30	17.75	43.91	21.78

4.5. Experiment 3

In the third experiment, we considered a 20-dimensional multivariate Gaussian, where the mean components of the mixture were $\mu_1 = -40 \times \mathbf{1}_{20}$, $\mu_2 = 0 \times \mathbf{1}_{20}$, $\mu_3 = 40 \times \mathbf{1}_{20}$, where $\mathbf{1}_{20}$ represents the $20-$dimensional column vector of all ones. All the simulation parameters were the same as before except that the variance of the proposals for the standalone PMC were $10I_{20}$. For the PMC part of the proposed method, the variance of the proposals were $\Sigma = \sigma^2 I$ where $\sigma^2 \in \{9/10, 9\}$.

The results of this experiment are shown in Table 3. In this experiment, the proposed method outperformed the PMC and VI the most.

Table 3. Experiment 3 Results

Method	$\sigma_X^2 = 0.5$	$\sigma_X^2 = 1$	$\sigma_X^2 = 4$	$\sigma_X^2 = 9$
PMC	6020.9	6177.0	5590.7	4664.9
VI	4024.1	3759.7	3610.5	3479.1
VI-PMC	173.6	133.84	134.1	146.5

In summary, we see from the above that the combination of both methods always produces better results. We can attribute this to the VI-step, which hones in close to the proper areas of high probability where samples should be generated. The AIS-step is then able to produce a fine grained approximation of the target by properly exploring regions of high probability, in contrast to a random initialization where the AIS may never be in position to explore them.

5. CONCLUSION

In this paper we proposed a new strategy of combining variational inference (VI) and adaptive importance sampling (AIS). The method applies AIS which is initialized via VI. We also proposed an adaptation step that allows the method to be robust to the performance of the VI initialization step. Numerical results show that the proposed method outperforms the respective individual implementations of AIS via PMC and VI. The improvement in performance is more dramatic as the number of modes increases and the dimension of the target distribution becomes higher.

6. REFERENCES

[2] J.-M. Cornuet, J.-M. Marin, A. Mira, and C. P. Robert, "Adaptive Multiple Importance Sampling," no. 1987, pp. 1–20, 2009.

[1] G. Casella, S. Fienberg, and I. Olkin, *Springer Texts in Statistics*, vol. 102. 2006.

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

[4] O. Cappé, R. Douc, A. Guillin, J.-m. Marin, O. Cappé, R. Douc, A. Guillin, J.-m. Marin, and C. R. A. Im, "Adaptive Importance Sampling in General Mixture Classes To cite this version : HAL Id : inria-00181474 Adaptive Importance Sampling in General Mixture Classes," 2008.

[5] R. Douc, A. Guillin, J. M. Marin, and C. P. Robert, "Convergence of adaptive mixtures of importance sampling schemes," *Annals of Statistics*, vol. 35, no. 1, pp. 420–448, 2007.

[6] M. Jordan and J. Kleinberg, *Information Science and Statistics*, vol. 4. 2006.

[7] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in Variational Inference," *CoRR*, vol. abs/1711.05597, 2017.

[8] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*. New York, NY, USA: Cambridge University Press, 2002.

[9] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[10] M. F. Bugallo, L. Martino, and J. Corander, "Adaptive importance sampling in signal processing," *Digital Signal Processing: A Review Journal*, vol. 47, pp. 36–49, 2015.

[11] 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.

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

[13] D. M. Blei, A. Kucukelbir, and J. D. Mcauliffe, "Variational Inference: A Review for Statisticians," 2017.

ISBN 978-90-827970-1-5 © EURASIP 2018