A Variational Adaptive Population Importance Sampler

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

Department of Electrical & Computer Engineering

Stony Brook University, Stony Brook, NY, USA

1. Introduction

- Monte Carlo (MC) methods and variational inference (VI) are the two main approaches used to approximate Bayesian posterior distributions.
- Each approach has its own challenges:
  - MC – scalability to complex systems.
  - VI – accuracy of the variational approximation.
- **Goal**: To apply robust techniques in stochastic optimization to scale adaptive importance sampling (AIS) methods for inference in high-dimensional probabilistic models.

2. Problem Formulation

- Given a set of i.i.d. observations $\mathbf{y}_1, \ldots, \mathbf{y}_N \sim p(\mathbf{y}|\mathbf{x})$, where each $\mathbf{y}_i \in \mathbb{R}^{d_y}$, we would like find the posterior probability of $\mathbf{x}$ given the observations:

$$\pi(\mathbf{x}) \equiv p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \propto \tilde{\pi}(\mathbf{x}) \equiv p(\mathbf{y}|\mathbf{x})p(\mathbf{x}).$$

- The normalizing constant $p(\mathbf{y})$ is unknown and can be estimated using importance sampling:

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

- We want to learn the best proposal, $q(\mathbf{x};\boldsymbol{\theta})$, by minimizing the variance of $\hat{Z}_{IS}$ with respect to the parameters $\boldsymbol{\theta}$.

3. Algorithm Summary

Sampling

Draw N samples from K proposal distributions,

$$\mathbf{x}_{t,k}^{(n)} \sim q_k(\mathbf{x};\boldsymbol{\theta}_{t,k}), \qquad n = 1,\ldots,N, \quad k = 1,\ldots,K.$$

Weighting

Compute the deterministic mixture weights,

$$w_{t,k}^{(n)} = \frac{\tilde{\pi}(\mathbf{x}_{t,k}^{(n)})}{\frac{1}{K}\sum_{k=1}^{K} q_k(\mathbf{x}_{t,k}^{(n)};\boldsymbol{\theta}_{t,k})}, \qquad n = 1,\ldots,N, \quad k = 1,\ldots,K.$$

Adaptation

For $k = 1, \ldots, K$

i. Compute the stochastic gradient $\tilde{g}(\boldsymbol{\theta}_{t,k})$.

ii. Update the vector of proposal parameters $\boldsymbol{\theta}_{t,k}$,

$$\boldsymbol{\theta}_{t+1,k} = \Pi_{\mathcal{C}}\left(\boldsymbol{\theta}_{t,k} - \eta_t \tilde{g}(\boldsymbol{\theta}_{t,k})\right)$$

$t := t+1$

4. Proposed Methodology

- The optimization problem we would like to solve is:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} C(\boldsymbol{\theta})$$

- For example, $C(\boldsymbol{\theta})$ could be chosen as to minimize a monotonic transformation of the Rényi divergence:

$$C(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \pi(\mathbf{x})^{\alpha} q(\mathbf{x};\boldsymbol{\theta})^{1-\alpha} d\mathbf{x}, \quad \alpha > 1$$

- The gradient of $C(\boldsymbol{\theta})$ is given as:

$$\nabla_{\boldsymbol{\theta}} C(\boldsymbol{\theta}) = -\mathbb{E}_q\left[\left(\frac{\tilde{\pi}(\mathbf{x})}{q(\mathbf{x};\boldsymbol{\theta})}\right)^{\alpha} \nabla_{\boldsymbol{\theta}}\left(\log q(\mathbf{x};\boldsymbol{\theta})\right)\right]$$

- Consider that $q(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \rho_k q_k(\mathbf{x};\boldsymbol{\theta}_k)$, where ρ_k and $\boldsymbol{\theta}_k$ denote the weight and parameters of the kth mixand. Then, the gradient of $C(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}_k$ is given by:

$$\nabla_{\boldsymbol{\theta}_k} C(\boldsymbol{\theta}) = -\mathbb{E}_q\left[\left(\frac{\tilde{\pi}(\mathbf{x})}{q(\mathbf{x};\boldsymbol{\theta})}\right)^{\alpha} \frac{\nabla_{\boldsymbol{\theta}_k}(q(\mathbf{x};\boldsymbol{\theta}))}{q(\mathbf{x};\boldsymbol{\theta})}\right]$$

- If there exists a function $\Psi(\mathbf{x},\boldsymbol{\theta}_k)$ such that

$$\nabla_{\boldsymbol{\theta}_k}(q(\mathbf{x};\boldsymbol{\theta})) = \rho_k q_k(\mathbf{x};\boldsymbol{\theta}_k)\Psi(\mathbf{x},\boldsymbol{\theta}_k),$$

  then the gradient $\nabla_{\boldsymbol{\theta}_k} C(\boldsymbol{\theta})$ can alternatively be written as:

$$\nabla_{\boldsymbol{\theta}_k} C(\boldsymbol{\theta}) = -\rho_k \mathbb{E}_{q_k}\left[\left(\frac{\tilde{\pi}(\mathbf{x})}{q(\mathbf{x};\boldsymbol{\theta})}\right)^{\alpha} \Psi(\mathbf{x},\boldsymbol{\theta}_k)\right] \quad (1)$$

Proposition: *Let $q_k(\mathbf{x};\boldsymbol{\theta}_k)$ be a member of the exponential family of probability distributions. Then, $\Psi(\mathbf{x},\boldsymbol{\theta}_k)$ exists and is given by*

$$\Psi(\mathbf{x},\boldsymbol{\theta}_k) = \nabla_{\boldsymbol{\theta}_k}(\boldsymbol{\beta}(\boldsymbol{\theta}_k)^{\top}\mathbf{T}(\mathbf{x}) - A(\boldsymbol{\theta}_k)),$$

where $\boldsymbol{\beta}(\boldsymbol{\theta}_k)$, $\mathbf{T}(\mathbf{x})$ and $A(\boldsymbol{\theta}_k)$ are known functions. Then, the gradient $\nabla_{\boldsymbol{\theta}_k} C(\boldsymbol{\theta})$ can be expressed according to (1).

Example: Location parameters of a Gaussian mixture for AIS

- Let $q(\mathbf{x};\boldsymbol{\theta}_t) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_{t,k},\boldsymbol{\Sigma}_k)$. We derive an expression for the stochastic gradient $g(\boldsymbol{\mu}_{t,k}) \approx \nabla_{\boldsymbol{\mu}_{t,k}} C(\boldsymbol{\theta}_t)$ as follows:

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

  where $\mathbf{x}_{t,k}^{(n)} \sim \mathcal{N}(\boldsymbol{\mu}_{t,k},\boldsymbol{\Sigma}_k)$ for $n = 1,\ldots,N$.
- Choosing $\alpha = 2$ minimizes the variance of $\hat{Z}_{IS}$.

5. Simulations

- Our goal is to approximate the following target in $\mathbb{R}^{20}$:

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

- **Goal**: Estimate the normalizing constant $Z = \sum_{j=1}^{5}\tilde{\rho}_j$ and the target mean $\mathbb{E}_{\pi}[\mathbf{x}] = \frac{1}{Z}\sum_{j=1}^{5}\tilde{\rho}_j \boldsymbol{m}_j$.
- We used our method to adapt the location parameters of a mixture of Gaussians as in (2). We set $\boldsymbol{\Sigma}_k = \sigma^2\mathbb{I}_{20}$.

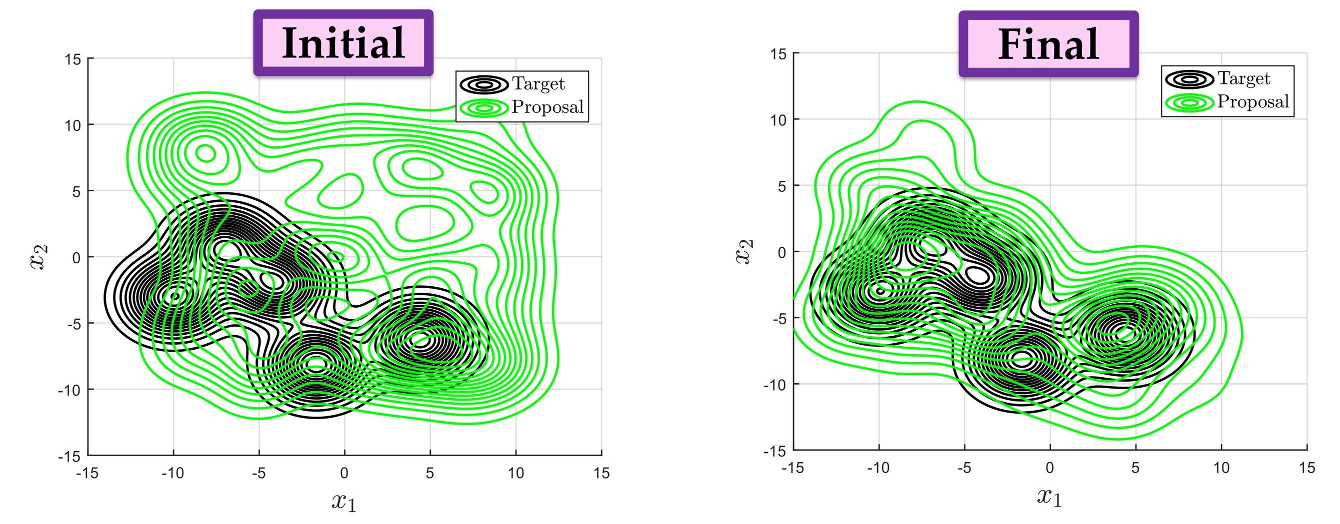
σ^2	2	4	6	8	10
M-PMC	27.90	27.91	27.65	27.77	27.67
DM-PMC	7.06	9.05	13.90	16.99	19.90
APIS	1.58	3.52	9.20	14.13	18.76
VAPIS	**0.05**	**0.04**	**0.18**	**0.23**	**0.84**

Table 1: MSE in the estimation of $\mathbb{E}_{\pi}[\mathbf{x}]$.

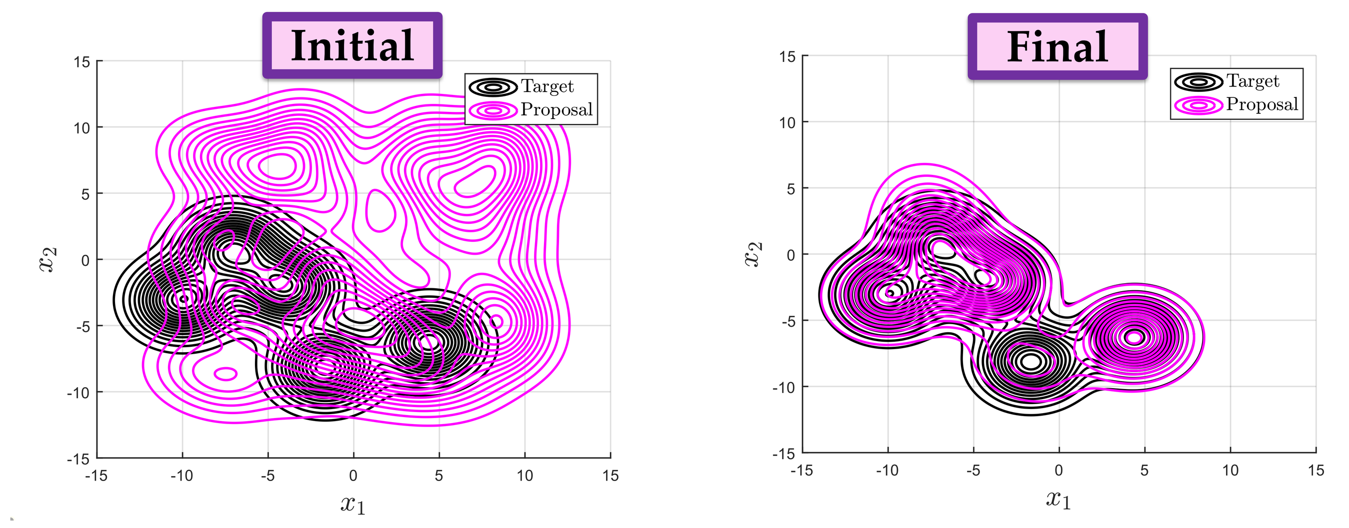
σ^2	2	4	6	8	10
M-PMC	496.2	480.3	500.6	419.3	415.6
DM-PMC	483.5	457.9	659.2	552.4	631.7
APIS	134.4	195.9	472.2	563.9	563.3
VAPIS	**21.1**	**21.8**	**55.9**	**42.8**	**84.3**

Table 2: MSE in the estimation of Z.

- Initial and final proposal of the APIS method:



- Initial and final proposal of the proposed method:



6. Conclusions

- We proposed a novel adaptation scheme for AIS samplers that explicitly optimizes a mixture's parameters by means of deterministic mixture sampling.
- The results of the numerical experiment showed that the proposed method outperforms other AIS samplers when dealing with high-dimensional target distributions.