# Stochastic mirror descent for nonparametric adaptive importance sampling

P. Bianchi[1] B. Delyon [2] F. Portier[3] **V. Priser**[1]

[1]LTCI, Télécom Paris

[2]IRMAR, Université de Rennes

[3]CREST, ENSAI

https://arxiv.org/abs/2409.13272

# Objective

Target: A probability measure admitting a density $f : \mathbb{R}^d \to \mathbb{R}$.

- Unnormalized density: $f_u$ such that $f_u(x)/\int f_u = f(x)$.
- $f_u(x)$ is known for $x \in \mathbb{R}^d$.
- $\int f_u$ is unknown and challenging to compute.
- No access to the gradient $\nabla f_u$.

**Question**

How can we evaluate the quantity

$$\mathbb{E}_{X\sim f}(\phi(X)) := \int \phi(x) f(x) dx,$$

for some function $\phi$?

In the sequel: $f = f_u$.
Applications: Variational Inference, Bayesian Inference, Reinforcement Learning, Stochastic Optimization, ...

# Application: Bayesian Inference

## Model

- Parameter: $\theta \sim \underbrace{p(\cdot)}_{\text{prior}}$
- Data: $x|\theta \sim \underbrace{p(\cdot|\theta)}_{\text{likelihood}}$

Objective: Find the posterior distribution $p(\cdot|x)$ of $\theta$, given the data $x$.

$$p(\cdot|x) := \frac{p(x|\cdot)p(\cdot)}{\int p(x|\theta)p(\theta)d\theta}$$

Target: $f_u(\cdot) = p(x|\cdot)p(\cdot)$.

# Context

Several types of algorithms:

- **Markov Chain Monte Carlo (MCMC)**: Construction of a Markov chain with an invariant distribution equal to the target distribution.
  - Metropolis-Hastings Metropolis et al. 1953
  - Hamiltonian Monte Carlo (HMC) Duane et al. 1987
  - No-U-Turn Sampler (NUTS) Hoffman and Gelman 2011
  - Langevin algorithm Meyn and Tweedie 1993
- **Importance Sampling Methods**: Repeated random sampling with weights assigned to each sample to approximate the target distribution.
  - Importance Sampling (IS) Kloek and Van Dijk 1978
  - Adaptive Importance Sampling Oh and Berger 1992
  - Annealed Importance Sampling Neal 2001
- **Variational Inference**: Optimization of a functional defined in the measure space (e.g., Kullback-Leibler divergence), where the minimizer corresponds to the target distribution.
  - Stein Variational Gradient Descent (SVGD) Liu and Wang 2016
  - Mollified Interaction Energy Descent (MIED) Li et al. 2023
  - Normalizing Flows Tabak and Vanden-Eijnden 2010

# Table of Contents

We propose an Importance Sampling method whose dynamics are inspired by Variational Inference (Mirror Descent).

# Important Sampling Methods

- Generate $n$ particles $(X_1, \ldots, X_n)$.
- Assign a weight $w_i$ to each particle $X_i$.

**Importance Sampling Estimator**

$$\hat{\phi}_n := \frac{1}{n} \sum_{i=1}^{n} w_i \phi(X_i)$$

Good estimator if:

- $\mathbb{E}(\hat{\phi}_n) = \int \phi(x) f(x) dx$.
- $\mathbb{V}(\hat{\phi}_n) \leq \frac{C}{n}$.

# Vanilla Importance Sampling algorithm

$q_0$: known density

## Vanilla Importance Sampling

- $(X_1, \ldots, X_n) \sim_{i.i.d.} q_0$
- $w_i := \frac{f}{q_0}(X_i)$

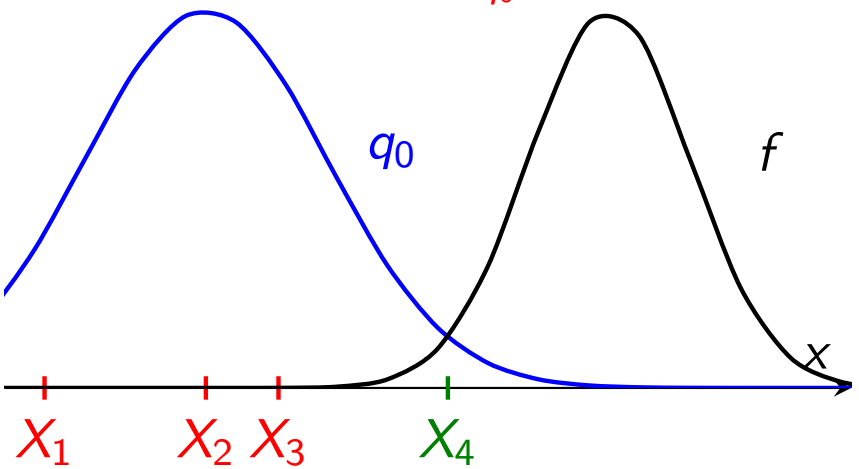
- $\mathbb{E}\left(\hat{\phi}_n\right) = \int \frac{f(x)}{q_0(x)} q_0(x)\phi(x)dx = \int f(x)\phi(x)dx\,.$
- $\mathbb{V}\left(\hat{\phi}_n\right) = \frac{1}{n} \underbrace{\left( \int \frac{f(x)^2}{q_0(x)}\phi(x)^2 dx - \left( \int f(x)\phi(x)dx \right)^2 \right)}_{\leq C}$

Assumption: $q_0 \geq cf$

# Issue with Importance Sampling

$$\mathbb{V}\left(\hat{\phi}_n\right) = \frac{1}{n}\left(\int \frac{f(x)^2}{q_0(x)}\phi(x)^2 dx - \left(\int f(x)\phi(x)dx\right)^2\right)$$

behaves badly when $\frac{f}{q_0}(x) \gg 1$ for some $x$.




$$\sum_{i=1}^{3} \frac{f(X_i)}{q_0(X_i)} \ll \frac{f(X_4)}{q_0(X_4)}$$

$$\hat{\phi}_4 \simeq \frac{f(X_4)}{q_0(X_4)}\phi(X_4)$$

## Optimal Importance Sampling

$$q_0 = \underset{q\,:\,density}{\arg\min}\,\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)}{q(X_i)}\phi(X_i)\right) = f$$

# Adaptive strategy

## Important Sampling (IS) algorithm

For $k = 0, \ldots, n$,

1. $X_{k+1} \sim q_0$
2. $w_{k+1} = \frac{f(X_{k+1})}{q_0(X_{k+1})}$

- To minimize the variance, we want: $q_0 = f$, but $f$ is unknown.
- $\tilde{q}_k := \frac{\sum_{i=1}^{k} w_i \delta_{X_i}}{\sum_{i=1}^{k} w_i}$ converges (in the weak-* topology) to $f$.
- $\tilde{q}_k = q_0$ should perfom better but $\tilde{q}_k$ has no density.

## Conclusion

At step $k$, we replace $q_0$ with a probability measure $q_k$ in the IS algorithm. Where $q_k$:

- has a density.
- $q_k \to f$.

# Kernel Density Estimation

$K$: a kernel and $K_h := \frac{1}{h^d} K(\frac{x}{h})$; $(X_1, \ldots, X_n) \sim_{i.i.d.} q_0$.

## Kernel density estimator Parzen 1962

$$g_n := \frac{1}{n} \sum_{i=1}^{n} \left( \frac{f}{q_0}(X_i) \right) K_h(\cdot - X_i)$$

- $\mathbb{E}(g_n(x)) = f * K_h(x) \to_{h \to 0} f(x)$
- $\mathbb{V}(g_n(x)) = \frac{1}{n} \left( \underbrace{\int \frac{f(y)^2}{q_0(y)} K_h(x-y)^2 dy}_{\leq C h^{-d}} - (f * K_h(x))^2 \right)$

## Conditions

$h \to 0$ and $\frac{1}{nh^d} \to 0$

# Adaptive Importance sampling

$g_0 = 0$

## Adaptive Importance Sampling algorithm

For $k = 0, \ldots, n$:

1. $X_{k+1} \sim q_k$
2. $w_{k+1} = \frac{f(X_{k+1})}{q_k(X_{k+1})}$
3. $g_{k+1} = (1 - \frac{1}{k+1})q_k + \frac{1}{k+1} w_{k+1} K_{h_k}(\cdot - X_{k+1})$
4. $q_{k+1} = \frac{g_{k+1}}{\int g_{k+1}}$

$g_k$ is the Kernel density estimator of $f$:

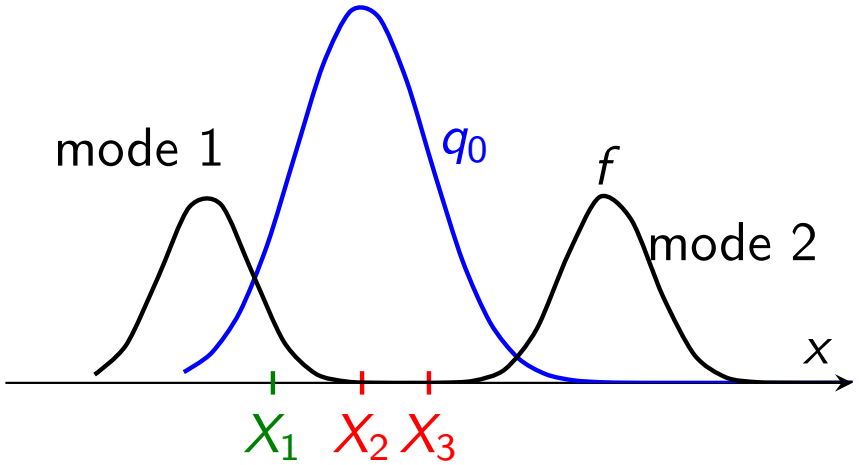
$$g_k(x) = \frac{1}{k} \sum_{i=1}^{k} w_i K_{h_i}(\cdot - X_i).$$

## Mini-Batch Variant

At each iteration, instead of simulating a single particle, we simulate $m$ particles according to the distribution $q_k$.

# Challenges in Adaptive Importance Sampling

Step $k = 0$; batch-size $m = 3$, $(X_1, \ldots, X_3) \sim_{i.i.d.} q_0$.




$$g_1 \simeq \frac{f}{q_0}(X_1) K_{h_0}(\cdot - X_1) \implies q_1 \propto K_{h_0}(\cdot - X_1)$$

New particles are sampled around $X_1 \longrightarrow$ mode 2 cannot be recovered.

## Solution

$$q_k = (1 - \lambda_k) \frac{g_k}{\int g_k} + \lambda_k q_0$$

$\lambda_k$ adds a safe component to $q_k$.

# Safe Adaptative Importance Sampling (SAIS)

$(\lambda_k)$: mixture weights satisfying $\lambda_k \to 0$.

## Safe Adaptative Importance Sampling

For $k = 0, \ldots, n$:

1. $X_{k+1} \sim q_k$
2. $w_{k+1} = \frac{f(X_{k+1})}{q_k(X_{k+1})}$
3. $g_{k+1} = q_k(1 - \frac{1}{k+1}) + \frac{1}{k+1} w_{k+1} K_{h_k}(x - X_{k+1})$
4. $q_{k+1} = (1 - \lambda_{k+1}) \frac{g_{k+1}}{\int g_{k+1}} + \lambda_{k+1} q_0$

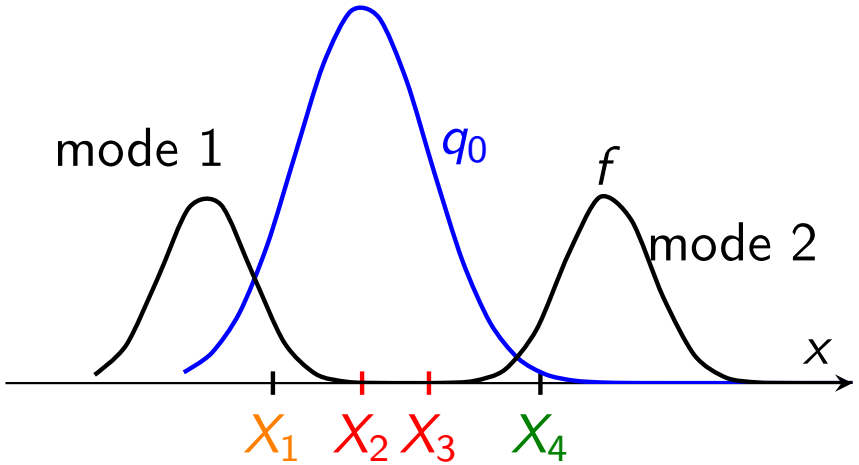
## Delyon and Portier 2021

$$\lim_{n \to \infty} \sup_{x \in \mathbb{R}^d} |q_n(x) - f(x)| = 0 \quad a.s.$$

An important assumption is $\sup_{x \in \mathbb{R}^d} \frac{f}{q_0}(x) < \infty$.

# Problem with SAIS

Step $k = 0$; batch-size $m = 4$, $(X_1, \ldots, X_4) \sim q_0$.




$$\sum_{i=2}^{3} \frac{f}{q_0}(X_i) \ll \frac{f}{q_0}(X_1) \ll \frac{f}{q_0}(X_4)$$

$$g_1 \simeq w_1 K_{h_0}(x - X_1) + w_4 K_{h_0}(x - X_1)$$

- We sample around mode 1 with probability $\frac{w_1}{w_1+w_4} \ll 1$.
- We sample around mode 2 with probability $\frac{w_4}{w_1+w_4} \simeq 1$.

**Conlcusion**

Very low probability of sampling around mode 1, which can be forgotten by the algorithm.

# Variance Problem

## Reformulation of the problem

The weights $(w_1, \dots, w_n)$ have high variance.

The probability to sample around a particule $i$:

$$\frac{w_i}{\sum_{k=1}^{n} w_{k+1}},$$

is small for most of the particles.

## Idea from Korba and Portier 2022

Let $\eta \in (0, 1)$:

$$\mathbb{V}_{X \sim q}\left(\left(\frac{f}{q}(X)\right)^{\eta}\right) \leq \mathbb{V}_{X \sim q}\left(\frac{f}{q}(X)\right)$$

At step $k$,

$$g_{k+1} = (1 - \frac{1}{k+1})g_k + \frac{1}{k+1} w_{k+1}^{\eta} K_{h_k}(x - X_{k+1}).$$

# The Mirror Descent algorithm in $\mathbb{R}^d$

$$\min_{x \in \mathbb{R}^d} h(x)$$

Gradient Descent with learning rate $\eta \in (0, 1]$.

$$x_{k+1} = x_k - \eta \nabla h(x_k) = \arg\min\{x \mapsto \langle \nabla h(x_k), x - x_k \rangle + \frac{1}{\eta} \frac{\|x - x_k\|^2}{2}\}$$

**Bregman divergence**

Let $\psi$ be a convex function. For $x, y \in \mathbb{R}^d$:

$$D_\psi(y|x) := \psi(y) - \psi(x) - \langle \nabla \psi(x), y - x \rangle .$$

Mirror Descent for some function $\psi$:

$$x_{k+1} = \arg\min\{x \mapsto \langle \nabla h(x_k), x - x_k \rangle + \frac{1}{\eta} D_\psi(x|x_k)\}$$

Depending of $\psi$, there are many interesting dynamics.

# Mirror Descent in $\mathcal{P}_2(\mathbb{R}^d)$

$$\min_{q \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(q) = KL(q|f) := \int \log\left(\frac{q}{f}(x)\right) q(x)dx$$

For $\psi = \mathcal{F}$, the Mirror Descent algorithm on $\mathcal{F}$ with learning rate $\eta$:

$$q_{k+1} \propto q_k^{1-\eta} f^{\eta}$$

**Chopin, Crucinio, and Korba 2024**

$$KL(q_n|f) \leq \frac{(1-\eta)^n}{\eta} KL(q_0|f)$$

**Problem:**

$\int q^{1-\eta} f^{\eta}$ is unknown. The algorithm cannot be implemented.

# Link between the variance reduction and Mirror Descent

Variance reduction hinted:

$$g_{k+1} = (1 - \frac{1}{k+1})g_k + \frac{1}{k+1} w_{k+1}^{\eta} K_{h_k}(x - X_{k+1}).$$

Let $\mathcal{F}_k := \sigma(X_i : i \leq k)$.

$$\mathbb{E}[g_{k+1}|\mathcal{F}_k] = (1 - \frac{1}{k+1})g_k + \frac{1}{k+1} \underbrace{f^{\eta} q_k^{1-\eta}}_{\text{Miror Descent}} * K_{h_k}.$$

**Conclusion**

The variance reduction approach yields the same dynamics as Mirror Descent.

# MIrror Descent for Adaptive Sampling (MIDAS)

$\eta \in (0, 1]$: learning rate. Let $g_0 = 0$.

## MIDAS

For $k = 0, \ldots, n$:

1. $X_{k+1} \sim q_k$
2. $w_{k+1} = \frac{f(X_{k+1})}{q_k(X_{k+1})}$
3. $g_{k+1} = (1 - \frac{1}{k+1})g_k + \frac{1}{k+1} w_{k+1}^{\eta} K_{h_k}(x - X_{k+1})$
4. $q_{k+1} = \frac{g_{k+1}}{\int g_{k+1}}(1 - \lambda_{k+1}) + \lambda_{k+1} q_0$

## Bianchi et al. 2024

For every compact sets $A \subset \mathbb{R}^d$:

$$\lim_{n \to \infty} \sup_{x \in A} |q_n(x) - f(x)| = 0 \quad a.s.$$

Corrolary: $\sqrt{n}(\hat{\phi}_n - \mathbb{E}_{X \sim f}(\phi(X))) \to \mathcal{N}(0, \underbrace{\mathbb{V}_{X \sim f}(\phi(X))}_{\text{optimal variance}})$.

# A spurious stationnary distribution

We recall

$$\mathbb{E}[g_{k+1}|\mathcal{F}_k] = (1 - \frac{1}{k+1})g_k + \frac{1}{k+1} f^\eta q_k^{1-\eta} * K_{h_k}.$$

## Stationnary Distributions

Stationnary distributions $(q_*, g_*)$ solve the equation (since $K_{h_k} \to \delta_0$):

$$g_*(x) = f^\eta q_*^{1-\eta}(x)$$

for every $x$ such that $g_*(x), q_*(x) < \infty$.

When $\eta < 1$, we have two stationnary distributions:

- $q_*(x) = g*(x) = f$, for every $x$.
- $g_*(x) = q_*(x) = 0$ almost everywhere $\implies$ $q_*$ is degenerated.

# Avoidance of the trap $q_* = g_* = 0$

Define (for some $v > 0$):

$$u_{k+1} = (1 - \frac{1}{k+1})u_k + \frac{1}{k+1} u_k^{1-\eta} v^\eta \, .$$

When $u_k \leq v$, $u_{k+1} \geq u_k$. Therefore, when $u_0 > 0$, we can't have $u_k \to 0$. Then, by classical arguments, $u_k \to v$.

$$g_{k+1}(x) = (1 - \frac{1}{k+1})g_k(x) + \frac{1}{k+1} f^\eta q_k^{1-\eta} * K_{h_k}(x) + \frac{1}{k+1} \underbrace{\xi_{k+1}(x)}_{\mathbb{E}[\xi_{k+1}(x)|\mathcal{F}_k]=0} \, .$$

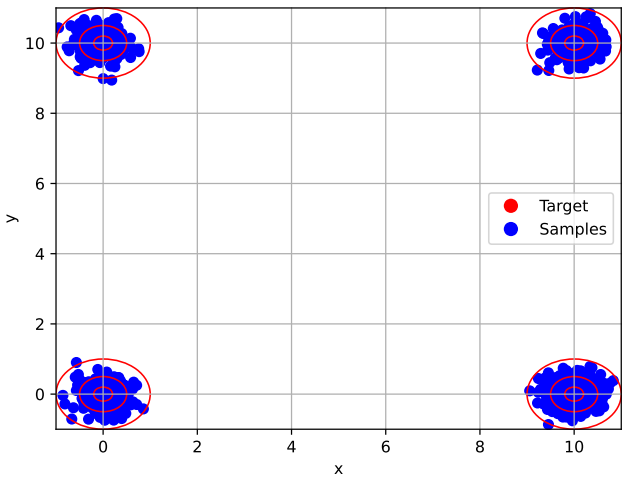
Iterating:

$$g_n(x) = \underbrace{A_n(x)}_{\text{deterministic algorithm}} + \underbrace{B_n q_0(x)}_{\text{safe component}} + \underbrace{M_n(x)}_{\text{martingale increment}} \, .$$

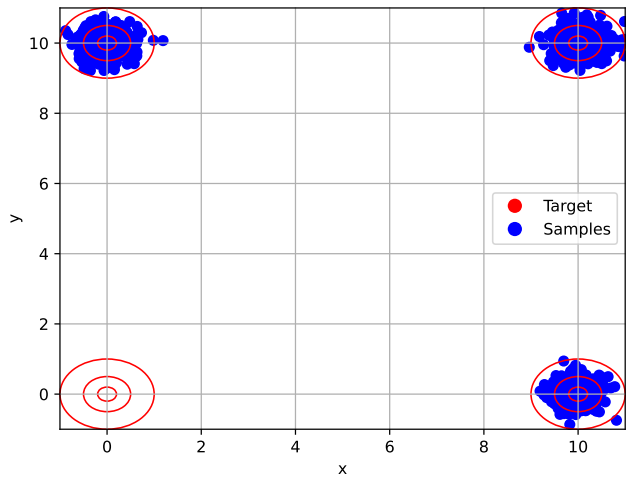
By a Freedman-type inequality, we show that for every $x$ in a compact space $A \subset \mathbb{R}^d$ and for $n \geq n_0$, $B_n q_0(x) + M_n(x) \geq 0$. The noise cannot lead the algorithm to the unstable equilibrium.

# Simulation

Target $f$: mixture of four gaussians centered in $(10i, 10j)$ for $i, j = 0, 1$.




(a) All the modes are found




(b) One mode is missing

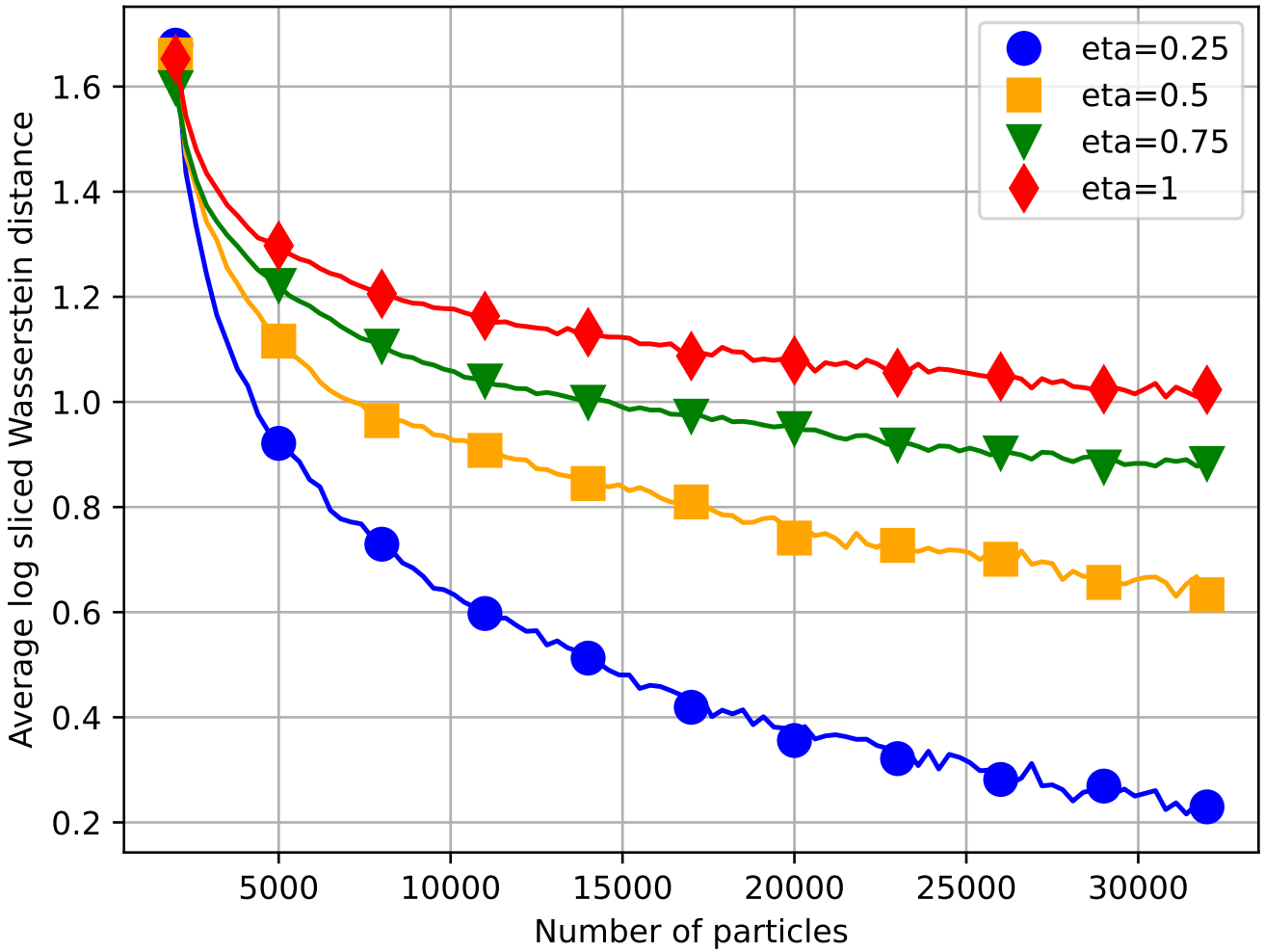



Figure: Average over 50 independent runs of the logarithm of sliced Wasserstein distance for different values of $\eta$

# Conclusion

Contribution:

- Demonstrated the convergence of a novel algorithm that introduces new dynamics at the intersection of Importance Sampling methods and Variational Inference.
- Achieved improved performance compared to the Safe Adaptive Importance Sampling algorithm, particularly in identifying modes of the target distribution.

Extension/Application of this result:

- Analysis of convergence speed.
- Guidelines for selecting the parameter $\eta$.
- Established connections with the annealed importance sampling algorithm.