Adaptive Importance Sampling meets Mirror Descent: a Bias-variance tradeoff

Anna Korba[1] François Portier[2]
[1]ENSAE, CREST, [2]ENSAI, CREST

JSTAR Rennes 2022

Outline

Problem : sample from a target distribution f over $\mathbb{R}^d$, whose density is typically known only up to a normalization constant, to compute quantities of the form $\int_{\mathbb{R}^d} gf$.

Problem : sample from a target distribution f over $\mathbb{R}^d$, whose density is typically known only up to a normalization constant, to compute quantities of the form $\int_{\mathbb{R}^d} gf$.

Main motivation = Bayesian inference:

Problem : sample from a target distribution f over $\mathbb{R}^d$, whose density is typically known only up to a normalization constant, to compute quantities of the form $\int_{\mathbb{R}^d} gf$.

Main motivation = Bayesian inference:

- Let $\mathcal{D} = (x_i, y_i)_{i=1}^m$ a labelled dataset of i.i.d. points.

Problem : sample from a target distribution f over $\mathbb{R}^d$, whose density is typically known only up to a normalization constant, to compute quantities of the form $\int_{\mathbb{R}^d} gf$.

Main motivation = Bayesian inference:

- Let $\mathcal{D} = (x_i, y_i)_{i=1}^m$ a labelled dataset of i.i.d. points.
- Assume an underlying model parametrized by $z \in \mathbb{R}^d$, e.g.

$$y = f(x, z) + \epsilon$$

($p(y|x, z)$ gaussian)

Problem : sample from a target distribution f over $\mathbb{R}^d$, whose density is typically known only up to a normalization constant, to compute quantities of the form $\int_{\mathbb{R}^d} gf$.

Main motivation = Bayesian inference:

- Let $\mathcal{D} = (x_i, y_i)_{i=1}^m$ a labelled dataset of i.i.d. points.
- Assume an underlying model parametrized by $z \in \mathbb{R}^d$, e.g.

$$y = f(x, z) + \epsilon$$

($p(y|x, z)$ gaussian)

- Compute the likelihood:

$$p(\mathcal{D}|z) = \prod_{i=1}^{m} p(y_i|x_i, z)$$

Problem : sample from a target distribution f over $\mathbb{R}^d$, whose density is typically known only up to a normalization constant, to compute quantities of the form $\int_{\mathbb{R}^d} gf$.

Main motivation = Bayesian inference:

- Let $\mathcal{D} = (x_i, y_i)_{i=1}^m$ a labelled dataset of i.i.d. points.
- Assume an underlying model parametrized by $z \in \mathbb{R}^d$, e.g.

$$y = f(x, z) + \epsilon$$

($p(y|x, z)$ gaussian)

- Compute the likelihood:

$$p(\mathcal{D}|z) = \prod_{i=1}^{m} p(y_i|x_i, z)$$

- Assume a prior distribution on the parameter $z \sim p$.

Bayes' rule : $f(z) := p(z|\mathcal{D}) = \frac{p(\mathcal{D}|z)p(z)}{C}$, $C = \int_{\mathbb{R}^d} p(\mathcal{D}|z)p(z)dz$.

f **is known up to a constant** since C is intractable.

How to sample from f then? e.g. to compute the "Bayesian model average":

$$p(y|x, \mathcal{D}) = \int_{\mathbb{R}^d} p(y|x, z)df(z)$$

Bayes' rule : $f(z) := p(z|\mathcal{D}) = \dfrac{p(\mathcal{D}|z)p(z)}{C}$, $C = \displaystyle\int_{\mathbb{R}^d} p(\mathcal{D}|z)p(z)dz$.

f **is known up to a constant** since C is intractable.

How to sample from f then? e.g. to compute the "Bayesian model average":

$$p(y|x, \mathcal{D}) = \int_{\mathbb{R}^d} p(y|x, z)df(z)$$

1. **MCMC methods** (Markov Chain Monte Carlo): generate a markov chain $(X_t)_{t\geq 0}$ whose law q_t converges to f as $t \to \infty$
2. **Variational Inference**

   $$\tilde{f} = \operatorname*{argmin}_{q\in\mathcal{Q}} \mathrm{KL}(q|f)$$

   where $\mathcal{Q}$ is a parametric family of probability distributions
3. **Importance sampling** : sample from a simple proposal distribution q dominating f and reweight

Naive Importance Sampling

Let X a random variable with distribution q dominating f. The basic idea of IS is to re-weight $g(X)$ by **the importance weight** $W(X) = f(X)/q(X)$.

Naive Importance Sampling

Let X a random variable with distribution q dominating f. The basic idea of IS is to re-weight $g(X)$ by **the importance weight** $W(X) = f(X)/q(X)$.

Since $\mathbb{E}[W(X)g(X)] = \int gf$ and using i.i.d. samples $X_1, \ldots, X_n \sim q$, one can build an (unbiased) IS estimator of $\int gf$ as

$$\int gf \approx \frac{1}{n}\sum_{k=1}^{n} \frac{f(X_k)}{q(X_k)} g(X_k) = \frac{1}{n}\sum_{k=1}^{n} W(X_k)g(X_k).$$

Naive Importance Sampling

Let X a random variable with distribution q dominating f. The basic idea of IS is to re-weight $g(X)$ by **the importance weight** $W(X) = f(X)/q(X)$.

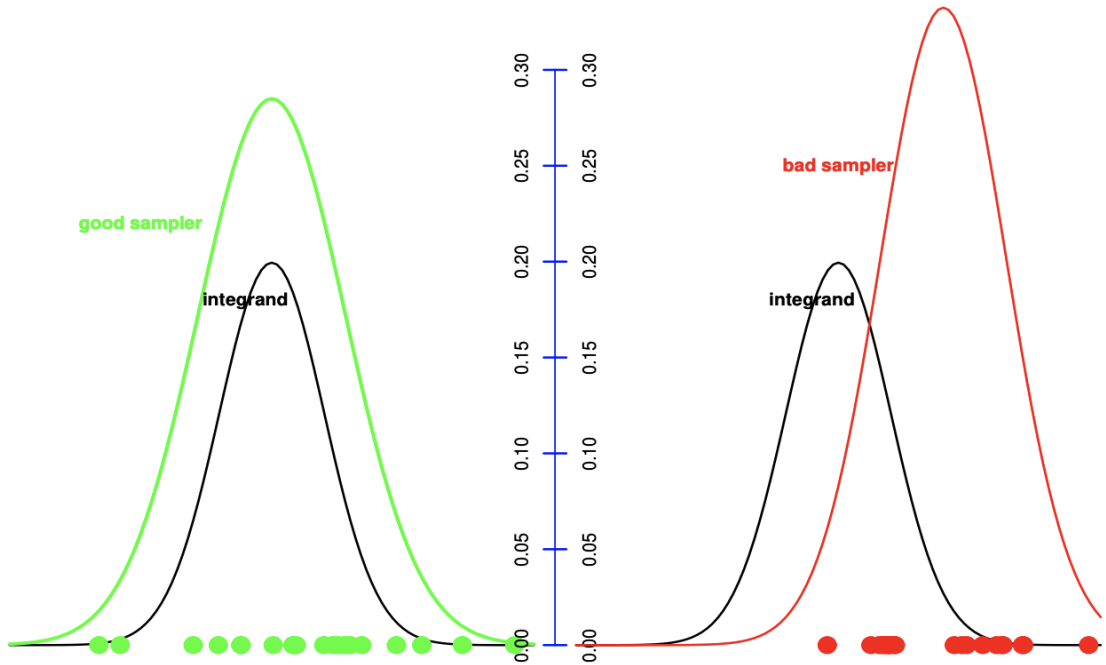
Since $\mathbb{E}[W(X)g(X)] = \int gf$ and using i.i.d. samples $X_1, \ldots, X_n \sim q$, one can build an (unbiased) IS estimator of $\int gf$ as

$$\int gf \approx \frac{1}{n}\sum_{k=1}^{n} \frac{f(X_k)}{q(X_k)} g(X_k) = \frac{1}{n}\sum_{k=1}^{n} W(X_k)g(X_k).$$

Remark: if f is known up to a normalization constant, use normalized weights $\sum_{k=1}^{n} W(X_k)g(X_k) / \sum_{k=1}^{n} W(X_k)$.

Naive importance sampling

Problem: if q is far from the target f, the importance weights may have a large variance (hence the IS estimator as well) !



The accuracy heavily depends on the choice of q

Contributions of the paper

Approach: Adaptive Importance Sampling (AIS)

Idea: use a sequence of proposals $(q_k)_{k\geq 0}$, learning from the past actions and data, to approximate f.

Contributions of the paper

Approach: Adaptive Importance Sampling (AIS)

Idea: use a sequence of proposals $(q_k)_{k\geq 0}$, learning from the past actions and data, to approximate f.

We propose a new non parametric AIS method, that

- (i) introduces a new regularization strategy which raises adaptively the importance sampling weights to a certain power ranging from 0 to 1
- (ii) uses a mixture between a kernel density estimate of the target and a safe reference density as proposal.

Outline

Regularized Importance Sampling

Idea: use regularized weights of the form $W(X)^{\eta}$, $\eta \in (0, 1)$.

Regularized Importance Sampling

Idea: use regularized weights of the form $W(X)^\eta$, $\eta \in (0, 1)$.

Lemma: Suppose that q dominates f and define $W(X) = f(X)/q(X)$ with X having density q. For all $\eta \in (0, 1]$:

$$\mathbb{E}[W(X)^\eta] \leq 1 \quad \text{and} \quad \mathrm{Var}[W(X)^\eta] \leq \mathrm{Var}[W(X)].$$

Regularized Importance Sampling

Idea: use regularized weights of the form $W(X)^\eta$, $\eta \in (0,1)$.

Lemma: Suppose that q dominates f and define $W(X) = f(X)/q(X)$ with X having density q. For all $\eta \in (0,1]$:

$$\mathbb{E}[W(X)^\eta] \leq 1 \quad \text{and} \quad \mathrm{Var}[W(X)^\eta] \leq \mathrm{Var}[W(X)].$$

Proof:

- Jensen's inequality: $\mathbb{E}[W(X)^\eta] \leq \mathbb{E}[W(X)]^\eta = 1$ since $f \ll q$.
- we have, since $|w^\eta - 1| \leq |w - 1|$ for all $w \geq 0$:

$$\begin{aligned}\mathrm{Var}[W(X)^\eta] &\leq \mathrm{Var}[W(X)^\eta] + (\mathbb{E}[W(X)^\eta] - 1)^2 \\ &= \mathbb{E}[(W(X)^\eta - 1)^2] \leq \mathbb{E}[(W(X) - 1)^2].\end{aligned}$$

Regularized Importance Sampling

Remarks:

- choosing η enables to balance bias and variance !
- $\mathbb{E}[W(X)^\eta g(X)] = \int f^\eta q^{1-\eta} g$

Hence, regularized IS "moves" from the initial density q ($\eta = 0$) to the target density $f^\eta q^{1-\eta}$ (=f if $\eta = 1$).

Regularized Importance Sampling

Remarks:

- choosing η enables to balance bias and variance !
- $\mathbb{E}[W(X)^\eta g(X)] = \int f^\eta q^{1-\eta} g$

Hence, regularized IS "moves" from the initial density q ($\eta = 0$) to the target density $f^\eta q^{1-\eta}$ (=f if $\eta = 1$).

Additional Remarks:

- different from simulated annealing (sequence of tempered posteriors) : $f \to f^\eta$
  $\Longrightarrow$ would yield IS weights of the form f^η / q instead of $(f/q)^\eta$

- can be seen as (entropic) mirror descent with step-size η_k:

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

Entropic mirror descent

Consider the sampling/variational inference objective:

$$q^\star = \operatorname*{arginf}_{q \in \mathcal{Q}} \ \mathrm{KL}(q|f), \quad \mathrm{KL}(q|f) = \int_{\mathbb{R}^d} \log\Big(\frac{q}{f}(x)\Big) dq(x)$$

Entropic mirror descent

Consider the sampling/variational inference objective:

$$q^\star = \operatorname*{arginf}_{q \in \mathcal{Q}} \ \mathrm{KL}(q|f), \quad \mathrm{KL}(q|f) = \int_{\mathbb{R}^d} \log\left(\frac{q}{f}(x)\right) dq(x)$$

Entropic mirror descent applied to this objective can be written at each time $k \geq 0$:

$$q^*_{k+1} = \operatorname*{argmin}_{q \in \mathcal{Q}} \ \eta_k \underbrace{\int_{\mathbb{R}^d} \log\left(\frac{q^*_k}{f}(x)\right) d(q - q^*_k)(x)}_{\langle \nabla \mathrm{KL}(q^*_k|f), q - q^*_k \rangle} + \mathrm{KL}(q|q^*_k). \quad (1)$$

Entropic mirror descent

Consider the sampling/variational inference objective:

$$q^\star = \operatorname*{arginf}_{q \in \mathcal{Q}} \ \mathrm{KL}(q|f), \quad \mathrm{KL}(q|f) = \int_{\mathbb{R}^d} \log\left(\frac{q}{f}(x)\right) dq(x)$$

Entropic mirror descent applied to this objective can be written at each time $k \geq 0$:

$$q_{k+1}^* = \operatorname*{argmin}_{q \in \mathcal{Q}} \ \eta_k \underbrace{\int_{\mathbb{R}^d} \log\left(\frac{q_k^*}{f}(x)\right) d(q - q_k^*)(x)}_{\langle \nabla \mathrm{KL}(q_k^*|f), q - q_k^* \rangle} + \mathrm{KL}(q|q_k^*). \quad (1)$$

Differentiate (1) w.r.t. q yields:

$$\eta_k \log\left(\frac{q_k^*}{f}\right) + \log\left(\frac{q_{k+1}^*}{q_k^*}\right) = 0$$

Entropic mirror descent

Consider the sampling/variational inference objective:

$$q^\star = \operatorname*{arginf}_{q \in \mathcal{Q}} \ \mathrm{KL}(q|f), \quad \mathrm{KL}(q|f) = \int_{\mathbb{R}^d} \log\left(\frac{q}{f}(x)\right) dq(x)$$

Entropic mirror descent applied to this objective can be written at each time $k \geq 0$:

$$q_{k+1}^* = \operatorname*{argmin}_{q \in \mathcal{Q}} \ \eta_k \underbrace{\int_{\mathbb{R}^d} \log\left(\frac{q_k^*}{f}(x)\right) d(q - q_k^*)(x)}_{\langle \nabla \mathrm{KL}(q_k^*|f), q - q_k^* \rangle} + \mathrm{KL}(q|q_k^*). \quad (1)$$

Differentiate (1) w.r.t. q yields:

$$\eta_k \log\left(\frac{q_k^*}{f}\right) + \log\left(\frac{q_{k+1}^*}{q_k^*}\right) = 0$$

Assuming $\mathcal{Q} = \mathcal{P}(\mathbb{R}^d)$, (1) is minimized for:

$$q_{k+1}^*(x) = \frac{f(x)^{\eta_k} q_k^*(x)^{1-\eta_k}}{\int_{\mathbb{R}^d} f(x')^{\eta_k} q_k^*(x')^{1-\eta_k} dx'} \propto f^{\eta_k}(x) q_k^{*(1-\eta_k)}(x).$$

Fast convergence of Entropic Mirror Descent (EMD)

The previous scheme is attractive thanks to its fast convergence.

Lemma

Let $(\eta_k)_{k\geq 1}$ *valued in* $(0,1]$ *and* $(q_k^*)_{k\geq 1}$ *be defined by EMD starting from an initial probability density function* q_1*. Then, for all* $n \in \mathbb{N}^*$*,*

$$\int_{\mathbb{R}^d} |f(x) - q_{n+1}^*(x)| dx \leq \sqrt{2\,\mathrm{KL}(f|q_1)} \prod_{k=1}^{n} (1-\eta_k)^{1/2},$$

- $\eta_k = c/k$ with $0 < c < 1$ yields the rate $O(n^{-c/2})$
- $\eta_k = c/k^\beta$ with $0 < c < 1$ and $\beta \in [0,1)$ yields the rate $O(\exp\left(-Cn^{(1-\beta)}\right))$ for some $C > 0$
- (Rk, ongoing work): linear rates in KL objective for constant step-size.

EMD in practice

Recall that EMD can be written at each time as:

$$q^*_{k+1} \propto f^{\eta_k} q_k^{*(1-\eta_k)}. \tag{2}$$

Unfortunately, running iteration (2) in practice is not feasible:

- each iteration depends on the whole density f (not some evaluations of f), maybe unknown in many applications
- (ii) even if f is known, computing $q^*_{k+1}(x)$ for $x \in \mathbb{R}^d$ would still be difficult due to the normalization following (2) which ensures q^*_{k+1} is a probability density function.

EMD in practice

Recall that EMD can be written at each time as:

$$q^*_{k+1} \propto f^{\eta_k} q_k^{*(1-\eta_k)}. \tag{2}$$

Unfortunately, running iteration (2) in practice is not feasible:

- each iteration depends on the whole density f (not some evaluations of f), maybe unknown in many applications
- (ii) even if f is known, computing $q^*_{k+1}(x)$ for $x \in \mathbb{R}^d$ would still be difficult due to the normalization following (2) which ensures q^*_{k+1} is a probability density function.

$\Longrightarrow$ we propose an implementable scheme approaching (2).

Outline

Safe and Regularized Adaptive Importance Sampling

We propose an *Adaptive Importance Sampling* (AIS) method which uses a sequence of proposals $(q_k)_{k\geq 0}$.

Safe and Regularized Adaptive Importance Sampling

We propose an *Adaptive Importance Sampling* (AIS) method which uses a sequence of proposals $(q_k)_{k\geq 0}$.

More specifically, as in [Delyon and Portier, 2021] we choose:

$$q_k = (1-\lambda_k) f_k + \lambda_k q_0, \qquad \forall k \geq 1$$

i.e. a mixture between

- a **safe density** q_0 (with heavy tails compared to f) , preventing too small values of q_k and high variance of IS weights,
- a **KDE estimate** f_k of the target f, accelerating the convergence to f

Safe and Regularized Adaptive Importance Sampling

We propose an *Adaptive Importance Sampling* (AIS) method which uses a sequence of proposals $(q_k)_{k\geq 0}$.

More specifically, as in [Delyon and Portier, 2021] we choose:

$$q_k = (1-\lambda_k) f_k + \lambda_k q_0, \qquad \forall k \geq 1$$

i.e. a mixture between

- a **safe density** q_0 (with heavy tails compared to f) , preventing too small values of q_k and high variance of IS weights,
- a **KDE estimate** f_k of the target f, accelerating the convergence to f

$$f_k(x) = \sum_{j=1}^{k} W_{k,j}^{(\eta_j)} K_{h_k}(x - X_j), \qquad \forall x \in \mathbb{R}^d,$$

where for all $j = 1, \ldots, k$:

$$W_{k,j}^{(\eta_j)} \quad \propto \quad W_j^{\eta_j} \quad = \quad \left(\frac{f(X_j)}{q_{j-1}(X_j)} \right)^{\eta_j}, \qquad \sum_{j=1}^{k} W_{k,j}^{(\eta_j)} \quad = \quad 1.$$

Safe and Regularized Adaptive Importance sampling (SRAIS) algorithm

---

Inputs: The safe density q_0, the sequences of bandwidths $(h_k)_{k=1,\dots,n}$, mixture weights $(\lambda_k)_{k=1,\dots,n}$, regularization parameters $(\eta_k)_{k=1,\dots,n}$.

For $k = 0, 1, \dots, n-1$:

1. Generate $X_{k+1} \sim q_k$.
2. Compute (a) $W_{k+1} = f(X_{k+1})/q_k(X_{k+1})$

   (b) $(W_{k+1,j}^{(\eta_j)})_{1 \le j \le k+1}$.
3. Return $q_{k+1} = (1-\lambda_{k+1}) f_{k+1} + \lambda_{k+1} q_0$ where

   $f_{k+1} = \sum_{j=1}^{k+1} W_{k+1,j}^{(\eta_j)} K_{h_{k+1}}(\cdot - X_j)$.

Remark: this algorithm can be used with a batch of m_k particles at each k.

SRAIS as stochastic approximation of mirror descent

Notice that

$$f_k(x) = \sum_{j=1}^{k} W_{k,j}^{(\eta_j)} K_{h_k}(x - X_j)$$

is a stochastic approximation of the mirror descent iteration $q_{k+1}^* \propto (q_k^*)^{1-\eta_k} f^{\eta_k}$. Indeed,

$$\mathbb{E}_{X_j \sim q_{j-1}}[W_j^{\eta_j} K_{h_k}(x - X_j)] = (f^{\eta_j} q_{j-1}^{1-\eta_j} \star K_{h_k})(x),$$

which approximates $f^{\eta_j} q_{j-1}^{1-\eta_j}$ when the bandwidth h_k is small.

Outline

Convergence of the scheme?

Does the kernel density estimate (KDE) f_k converge to the target distribution f?

Recall the hyperparameters of the algorithm :

- the safe density q_0 in the proposal (mixture between q_0 and f_k)
- mixture weights $(\lambda_k)_{k\geq 0}$
- KDE bandwidths $(h_k)_{k\geq 0}$
- regularization parameters $(\eta_k)_{k\geq 0}$

Assumptions

(A1)

1. The sequence $(\lambda_k)_{k\geq 1}$ is valued in $(0,1]$, nonincreasing, and $\lim_{k\to\infty} \lambda_k = 0$ and $\lim_{k\to\infty} \log(k)/(k\lambda_k) = 0$.

Assumptions

(A1)

1. The sequence $(\lambda_k)_{k\geq 1}$ is valued in $(0,1]$, nonincreasing, and $\lim_{k\to\infty} \lambda_k = 0$ and $\lim_{k\to\infty} \log(k)/(k\lambda_k) = 0$.
2. The sequence $(h_k)_{k\geq 1}$ is valued in $\mathbb{R}^+$, nonincreasing, and $\lim_{k\to\infty} h_k = 0$ and $\lim_{k\to\infty} \log(k)/(kh_k^d\lambda_k) = 0$.

Assumptions

(A1)

1. The sequence $(\lambda_k)_{k\geq 1}$ is valued in $(0,1]$, nonincreasing, and $\lim_{k\to\infty}\lambda_k = 0$ and $\lim_{k\to\infty}\log(k)/(k\lambda_k) = 0$.
2. The sequence $(h_k)_{k\geq 1}$ is valued in $\mathbb{R}^+$, nonincreasing, and $\lim_{k\to\infty}h_k = 0$ and $\lim_{k\to\infty}\log(k)/(kh_k^d\lambda_k) = 0$.
3. The sequence $(\eta_k)_{k\geq 1}$ is valued in $(0,1]$, and $\lim_{k\to\infty}\eta_k = 1$, $\lim_{k\to\infty}(1-\eta_k)\log(h_k) = 0$ and $\lim_{k\to\infty}(1-\eta_k)\log(\lambda_{k-1}) = 0$.

Assumptions

(A1)

1. The sequence $(\lambda_k)_{k\geq 1}$ is valued in $(0,1]$, nonincreasing, and $\lim_{k\to\infty} \lambda_k = 0$ and $\lim_{k\to\infty} \log(k)/(k\lambda_k) = 0$.
2. The sequence $(h_k)_{k\geq 1}$ is valued in $\mathbb{R}^+$, nonincreasing, and $\lim_{k\to\infty} h_k = 0$ and $\lim_{k\to\infty} \log(k)/(kh_k^d\lambda_k) = 0$.
3. The sequence $(\eta_k)_{k\geq 1}$ is valued in $(0,1]$, and $\lim_{k\to\infty} \eta_k = 1$, $\lim_{k\to\infty}(1-\eta_k)\log(h_k) = 0$ and $\lim_{k\to\infty}(1-\eta_k)\log(\lambda_{k-1}) = 0$.

(A2) The density q_0 is bounded and there exists $c > 0$ such that for all $x \in \mathbb{R}^d$, $q_0(x) \geq cf(x)$.

Assumptions

(A1)

1. The sequence $(\lambda_k)_{k\geq 1}$ is valued in $(0,1]$, nonincreasing, and $\lim_{k\to\infty}\lambda_k = 0$ and $\lim_{k\to\infty}\log(k)/(k\lambda_k) = 0$.
2. The sequence $(h_k)_{k\geq 1}$ is valued in $\mathbb{R}^+$, nonincreasing, and $\lim_{k\to\infty} h_k = 0$ and $\lim_{k\to\infty}\log(k)/(kh_k^d\lambda_k) = 0$.
3. The sequence $(\eta_k)_{k\geq 1}$ is valued in $(0,1]$, and $\lim_{k\to\infty}\eta_k = 1$, $\lim_{k\to\infty}(1-\eta_k)\log(h_k) = 0$ and $\lim_{k\to\infty}(1-\eta_k)\log(\lambda_{k-1}) = 0$.

(A2) The density q_0 is bounded and there exists $c > 0$ such that for all $x \in \mathbb{R}^d$, $q_0(x) \geq cf(x)$.

(A3) The function f is nonnegative, bounded by $B_f \geq 0$ and is l_f-Lipschitz.

Assumptions

(A1)

1. The sequence $(\lambda_k)_{k\geq 1}$ is valued in $(0,1]$, nonincreasing, and $\lim_{k\to\infty}\lambda_k = 0$ and $\lim_{k\to\infty}\log(k)/(k\lambda_k) = 0$.
2. The sequence $(h_k)_{k\geq 1}$ is valued in $\mathbb{R}^+$, nonincreasing, and $\lim_{k\to\infty} h_k = 0$ and $\lim_{k\to\infty}\log(k)/(kh_k^d\lambda_k) = 0$.
3. The sequence $(\eta_k)_{k\geq 1}$ is valued in $(0,1]$, and $\lim_{k\to\infty}\eta_k = 1$, $\lim_{k\to\infty}(1-\eta_k)\log(h_k) = 0$ and $\lim_{k\to\infty}(1-\eta_k)\log(\lambda_{k-1}) = 0$.

(A2) The density q_0 is bounded and there exists $c > 0$ such that for all $x \in \mathbb{R}^d$, $q_0(x) \geq cf(x)$.

(A3) The function f is nonnegative, bounded by $B_f \geq 0$ and is l_f-Lipschitz.

(A4) The kernel K is bounded by $K_\infty \geq 0$ and is L_K-Lipschitz with $L_K > 0$. Moreover, $\int K(u)du = 1$, $\int \|u\|K(u)du < \infty$, $\int K^{1/2}(u)d(u) < \infty$ and $\int \|u\|K(u)^{1/2}du < \infty$.

Uniform convergence of the scheme

Proposition: Assume **A1-A4**. Then, for any $r > 0$:

$$\sup_{\|x\| \le k^r} |f_k(x) - f(x)| \to 0 \quad \text{as } k \to \infty \text{ a.s.}$$

Proof: $f_k = N_k / D_k$ where

$$N_k(x) = \frac{1}{k} \sum_{j=1}^{k} W_j^{\eta_j} K_{h_k}(x - X_j), \quad D_k = \frac{1}{k} \sum_{j=1}^{k} W_j^{\eta_j}.$$

Uniform convergence of the scheme

Proposition: Assume **A1-A4**. Then, for any $r > 0$:

$$\sup_{\|x\| \le k^r} |f_k(x) - f(x)| \to 0 \quad \text{as } k \to \infty \text{ a.s.}$$

Proof: $f_k = N_k / D_k$ where

$$N_k(x) = \frac{1}{k} \sum_{j=1}^{k} W_j^{\eta_j} K_{h_k}(x - X_j), \quad D_k = \frac{1}{k} \sum_{j=1}^{k} W_j^{\eta_j}.$$

$$N_k - f = \underbrace{\frac{1}{k} \sum_{j=1}^{k} \left\{ W_j^{\eta_j} K_{h_k}(x - X_j) - \{ f^{\eta_j} q_{j-1}^{1-\eta_j} \} \star K_{h_k}(x) \right\}}_{(1)}$$

$$+ \underbrace{\left\{ \left(\frac{1}{k} \sum_{j=1}^{k} (f^{\eta_j} q_{j-1}^{1-\eta_j} - f) \right) \star K_{h_n} \right\}}_{(2)} + \underbrace{\{ f \star K_{h_k} - f \}}_{(3)}.$$

(1): avg of martingale increments, (2): regularization bias ($\to$0 as $\eta_k \to 1$), (3) KDE bias ($\to 0$ as $h_k \to 0$).

Outline

Adaptive Choice of Regularization (RAR)

Our conditions for uniform convergence require that the sequence $(\eta_k)_{k\geq 1}$ converges to 1. We propose an adaptive way to construct it.

Adaptive Choice of Regularization (RAR)

Our conditions for uniform convergence require that the sequence $(\eta_k)_{k\geq 1}$ converges to 1. We propose an adaptive way to construct it.

Idea: Draw m_k i.i.d samples $X_{k,1}, \ldots, X_{k,m_k}$ from q_{k-1}.

Let $\mathbb{P} = \sum_{l=1}^{m_k} W_{k,l}\delta_{X_{k,l}}$ and $\mathbb{Q} = \sum_{l=1}^{m_k} \frac{1}{m_k}\delta_{X_{k,l}}$

the reweighted and uniform distribution on the particles.

$\Longrightarrow$ If $q_{k-1} = f$, IS weights = 1 and $\mathbb{P} = \mathbb{Q}$.

$\Longrightarrow$ **penalize the divergence between $\mathbb{P}$ and $\mathbb{Q}$!**

Adaptive Choice of Regularization (RAR)

Our conditions for uniform convergence require that the sequence $(\eta_k)_{k\geq 1}$ converges to 1. We propose an adaptive way to construct it.

Idea: Draw m_k i.i.d samples $X_{k,1}, \ldots, X_{k,m_k}$ from q_{k-1}.

Let $\mathbb{P} = \sum_{l=1}^{m_k} W_{k,l}\delta_{X_{k,l}}$ and $\mathbb{Q} = \sum_{l=1}^{m_k} \frac{1}{m_k}\delta_{X_{k,l}}$

the reweighted and uniform distribution on the particles.

$\Longrightarrow$ If $q_{k-1} = f$, IS weights = 1 and $\mathbb{P} = \mathbb{Q}$.

$\Longrightarrow$ **penalize the divergence between $\mathbb{P}$ and $\mathbb{Q}$!**

We propose to use Renyi's α-divergences and set:

$$\eta_{k,\alpha} = 1 - \frac{D_\alpha(\mathbb{P}||\mathbb{Q})}{\log(m_k)}, \; D_\alpha(\mathbb{P}||\mathbb{Q}) = \frac{1}{\alpha - 1} \log\left(\sum_{\ell=1}^{m_k} W_{k,\ell}^\alpha m_k^{\alpha-1}\right).$$

RAR

$$\eta_{k,\alpha} = 1 - \frac{D_\alpha(\mathbb{P}||\mathbb{Q})}{\log(m_k)}, \; D_\alpha(\mathbb{P}||\mathbb{Q}) = \frac{1}{\alpha - 1} \log \left(\sum_{\ell=1}^{m_k} W_{k,\ell}^\alpha m_k^{\alpha-1} \right). \tag{3}$$

Proposition: Let $\alpha \in [0, 1]$ and let $(\eta_{k,\alpha})_{k \geq 1}$ be the sequence defined by (3) for all $k \geq 1$. Then, we have:

1. The sequence $(\eta_{k,\alpha})_{k \geq 1}$ is valued in $[0, 1]$, with $\eta_{k,\alpha} = 1$ iff $\mathbb{P} = \mathbb{Q}$;
2. $0 \leq \eta_{k,1} \leq \eta_{k,\alpha} \leq 1$;
3. Further assume that $(q_k)_{k \geq 1}$ is a sequence of probability density functions s.t. $\lim_{k \to \infty} |q_k(x) - f(x)| = 0$ almost everywhere and that $m_k = m$ for all $k \geq 1$ (fixed batch size). Then, $\lim_{k \to \infty} \eta_{k,\alpha} = 1$ in L_1.

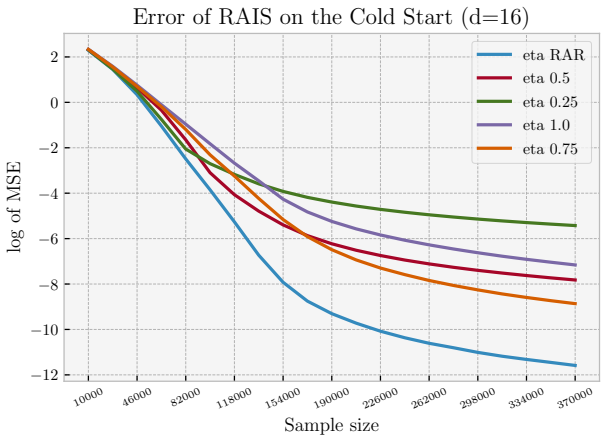
Renyi's α-divergences

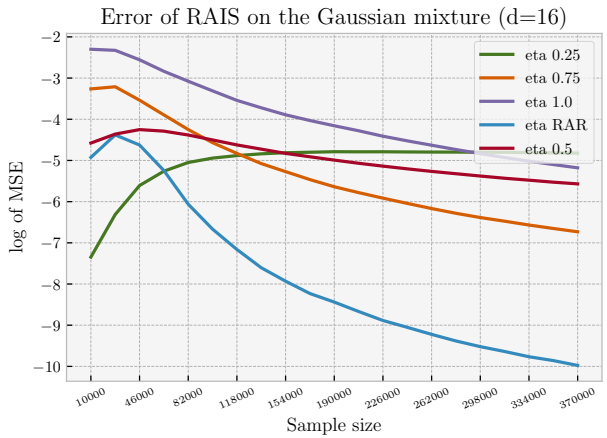
Table 1: Special cases in the Rényi divergence family.

α	Definition	Notes
$\alpha \to 1$	$\int p(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}$	*Kullback-Leibler (KL) divergence*, used in VI ($\mathrm{KL}[q\|\|p]$) and EP ($\mathrm{KL}[p\|\|q]$)
$\alpha = 0.5$	$-2 \log(1 - \mathrm{Hel}^2[p\|\|q])$	function of the square *Hellinger distance*
$\alpha \to 0$	$-\log \int_{p(\boldsymbol{\theta})>0} q(\boldsymbol{\theta}) d\boldsymbol{\theta}$	zero when $\mathrm{supp}(q) \subseteq \mathrm{supp}(p)$ (not a divergence)
$\alpha = 2$	$-\log(1 - \chi^2[p\|\|q])$	proportional to the χ^2-divergence
$\alpha \to +\infty$	$\log \max_{\boldsymbol{\theta} \in \Theta} \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}$	*worst-case regret* in *minimum description length principle* [24]

Outline

Toy Experiments





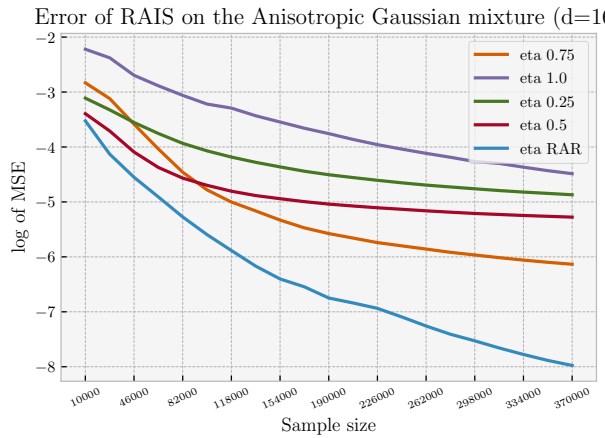


Figure: Logarithm of the average squared error for SRAIS for constant values of η or Adaptive η, over 50 replicates. 4×10^4 particles sampled from initial density, then $m_k = 18 \times 10^3$ particles from q_k at each $k \geq 1$.

Different target densities ($\phi_\Sigma = \mathcal{N}(0_d, \Sigma)$), initial densities have different means/variance than the target:

- "Cold Start" $f_1(x) = \phi_\Sigma(x - 5\mathbf{1}_d/\sqrt{d})$, $\Sigma = (0.16/d)\boldsymbol{I}_d$
- "Gaussian Mixture"
  $f_2(x) = 0.5\phi_\Sigma(x - \mathbf{1}_d/(2\sqrt{d})) + 0.5\phi_\Sigma(x + \mathbf{1}_d/(2\sqrt{d}))$
- "Anisotropic Gaussian Mixture"
  $f_3(x) = 0.25\phi_V(x - \mathbf{1}_d/(2\sqrt{d})) + 0.75\phi_V(x + \mathbf{1}_d/(2\sqrt{d}))$,
  $V = (.4/\sqrt{d})^2 \mathrm{diag}(10, 1, \ldots, 1)$

Evolution of Adaptive Regularization

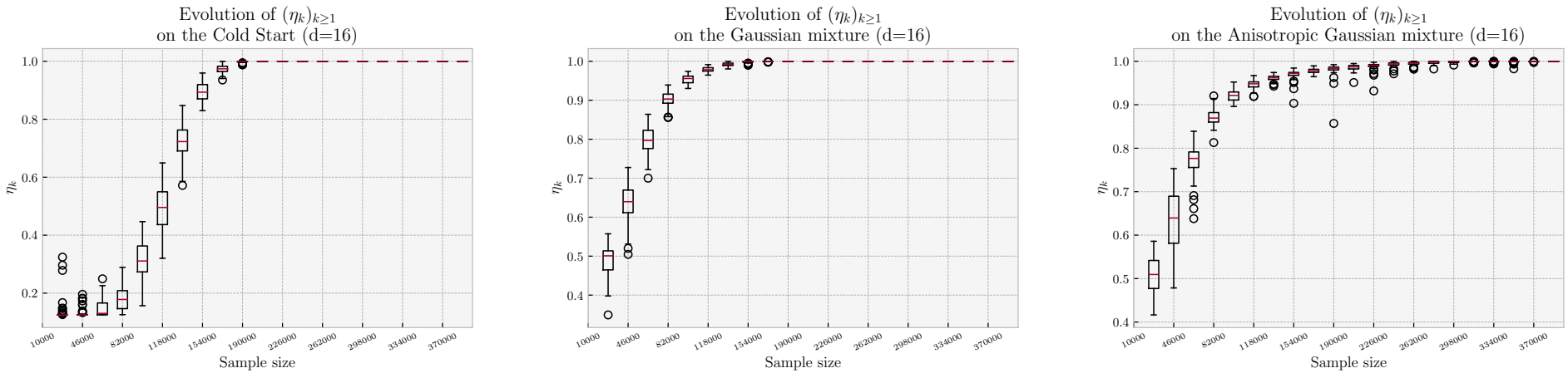


Figure: Boxplot of the values of $(\eta_{k,\alpha})_{k\geq 1}$ obtained from RAR (Adaptive η), with $\alpha = 0.5$.

- at the beginning of the algorithm when the policy is poor, the value of η_k is automatically set to a small value (leading to a uniformization of the weights)
- when the policy becomes better the value of $\eta_{k,\alpha}$ converges to 1.

Bayesian Logistic Regression (Waveform dataset, 5000 datapoints in $d = 22$)

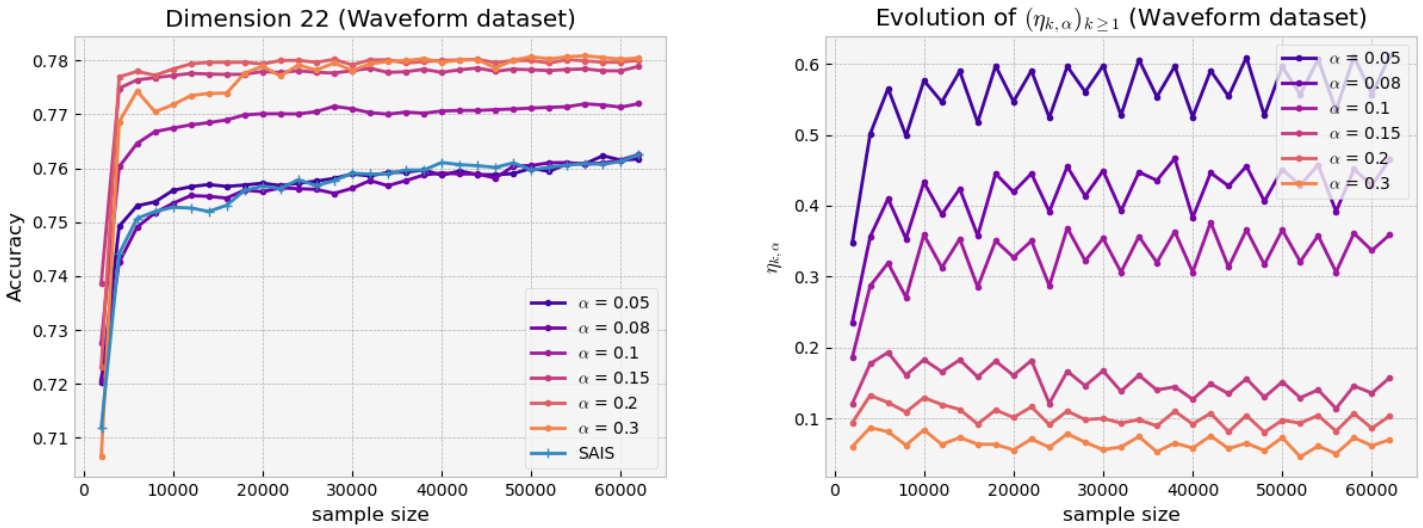


Figure: Left plot: Average accuracy over 100 trials of different learning policies $(\eta_{k,\alpha})_{k\geq 1}$ for Bayesian Logistic Regression on the Waveform dataset. Right plot: Averaged values of the learning policy $(\eta_{k,\alpha})_{k\geq 1}$ associated to each choice of α.

- a proper tuning of the parameter α allows us to outperform $(\eta_k)_{k\geq 1}$ constant and equal to 1
- the case $\alpha = 0.2$ yielding the best results here overall in terms of speed and accuracy

Conclusion

Contributions:

- We proposed a new algorithm for Adaptive Importance Sampling, that regularizes the importance weights by raising them to a certain power
- This algorithm is related to mirror descent on the space of probability distributions
- It enjoys a uniform convergence guarantee under mild assumptions on the target, safe density, and hyperparameters
- It outperforms numerically constant values of η

Conclusion

Future work:

- Non-asymptotic analysis of the scheme

Conclusion

Future work:

- Non-asymptotic analysis of the scheme
- Adaptive schedules for other hyperparameters

Conclusion

Future work:

- Non-asymptotic analysis of the scheme
- Adaptive schedules for other hyperparameters
- Replace KDE by a more scalable approximate ? e.g. normalizing flows [Papamakarios et al., 2021]:
  - generative models that transform a simple $q \in \mathcal{P}(\mathbb{R}^d)$ into $\tilde{f} \in \mathcal{P}(\mathbb{R}^d)$ ($\tilde{f} \approx f$)
  - through a sequence of invertible transformations:

$$\tilde{f} = G_{\#}q = (g_1 \circ \dots g_n)_{\#}q, \quad g_i : \mathbb{R}^d \to \mathbb{R}^d$$

  - $\implies$ sampling and evaluating the density are efficient

$$\tilde{f}(y) = q(z_0) \prod_{i=1}^{n} \left| det \left(\frac{\partial g_i}{\partial z_{i-1}} \right) \right|^{-1}, \quad y, z_0 \in \mathbb{R}^d$$

where $\frac{\partial g}{\partial z_{i-1}}$ is the Jacobian matrix of g_i.

Conclusion

Future work:

- Non-asymptotic analysis of the scheme
- Adaptive schedules for other hyperparameters
- Replace KDE by a more scalable approximate ? e.g. normalizing flows [Papamakarios et al., 2021]:
  - generative models that transform a simple $q \in \mathcal{P}(\mathbb{R}^d)$ into $\tilde{f} \in \mathcal{P}(\mathbb{R}^d)$ ($\tilde{f} \approx f$)
  - through a sequence of invertible transformations:

$$\tilde{f} = G_{\#}q = (g_1 \circ \dots g_n)_{\#}q, \quad g_i : \mathbb{R}^d \to \mathbb{R}^d$$

  - $\implies$ sampling and evaluating the density are efficient

$$\tilde{f}(y) = q(z_0) \prod_{i=1}^{n} \left| det \left(\frac{\partial g_i}{\partial z_{i-1}} \right) \right|^{-1}, \quad y, z_0 \in \mathbb{R}^d$$

where $\frac{\partial g}{\partial z_{i-1}}$ is the Jacobian matrix of g_i.

Thank you !

References I


Delyon, B. and Portier, F. (2021).
Safe adaptive importance sampling: A mixture approach.
The Annals of Statistics, 49(2):885–917.


Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021).
Normalizing flows for probabilistic modeling and inference.
Journal of Machine Learning Research, 22(57):1–64.