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

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

AISTATS 2022

Contributions of the paper

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 $\int_{\mathbb{R}^d} gf$.

Contributions of the paper

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 $\int_{\mathbb{R}^d} gf$.

Adaptive Importance Sampling (AIS) is one increasingly popular way to tackle this problem, whose idea is to sample from an alternative, simpler proposal probability density q_k at time k of the algorithm to approximate f.

In this paper, 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.

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 $\sum_{k=1}^{n} W(X_k)g(X_k) / \sum_{k=1}^{n} W(X_k)$.

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

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 $\eta \in (0,1]$:

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

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 to the target density $f^\eta q^{1-\eta}$ (=f if $\eta = 1$).

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 $\eta \in (0, 1]$:

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

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 to the target density $f^\eta q^{1-\eta}$ (=f if $\eta = 1$).

Additional Remarks:

- different from simulated annealing (f^η/q instead of $(f/q)^\eta$)
- it corresponds to mirror descent with step-size η_k:

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

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 $j = 1, \ldots, k$:

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

SRAIS algorithm

Algorithm 1 *Safe and Regularized Adaptive Importance sampling (SRAIS)*

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

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

(i) Generate $X_{k+1} \sim q_k$.

(ii) Compute (a) $W_{k+1} = f(X_{k+1})/q_k(X_{k+1})$ and (b) $(W_{k+1,j}^{(\eta_j)})_{1 \le j \le k+1}$.

(iii) 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,

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

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

Uniform convergence of the scheme

$(\mathbf{A}_1)$(i) The sequence $(\lambda_k)_{k\geq 1}$ is valued in $(0,1]$, nonincreasing, and $\lim_{k\to\infty}\lambda_k = 0$ and $\lim_{k\to\infty}\log(k)/(k\lambda_k) = 0$.

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

(iii) 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 $\lim_{k\to\infty}(1-\eta_k)\log(\lambda_{k-1}) = 0$.

$(\mathbf{A}_2)$ The density q_0 is bounded and there exists $c > 0$ such that for all $x \in \mathbb{R}^d$, $q_0(x) \geq cf(x)$.

$(\mathbf{A}_3)$ The function f is nonnegative, L-Lipschitz and bounded by $U \in \mathbb{R}^+$.

$(\mathbf{A}_4)$ $\int K = 1$, $\int \|u\|K(u)\mathrm{d}u < \infty$, $\int K^{1/2} < \infty$ and $\int \|u\|K(u)^{1/2}\mathrm{d}u < \infty$. The kernel K is bounded by $K_\infty \geq 0$ and is L_K-Lipschitz with $L_K > 0$, i.e. :

$$|K(x+u) - K(x)| \leq L_K\|u\| \quad \text{for all } x, u \in \mathbb{R}^d.$$

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

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

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:

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

Prop: $\lim_{k\to\infty} \eta_{k,\alpha} \to 1$ (in L^1) if $\lim_{k\to\infty} |q_k(x) - f(x)| = 0$ a.e.

Toy Experiments

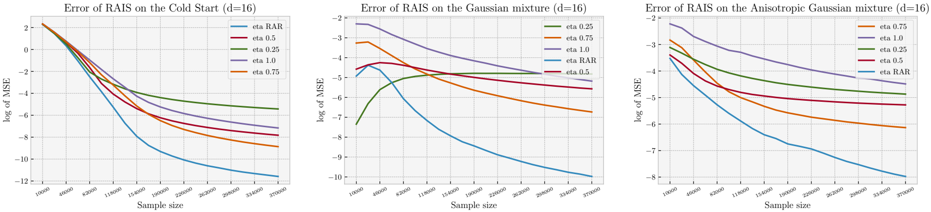

Figure: Logarithm of the average squared error for SRAIS for constant values of η or Adaptive η, over 50 replicates. 4×10^4 particles sampled from initial density, then $m_k = 18 \times 10^3$ particles from q_k at each $k \geq 1$.

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

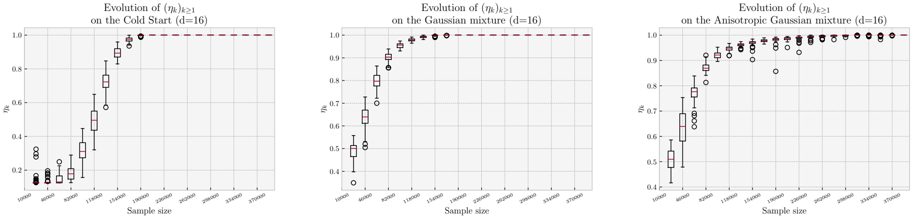

Figure: Boxplot of the values of $(\eta_{k,\alpha})_{k\geq1}$ obtained from RAR (Adaptive η), with $\alpha = 0.5$.

- 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 $d = 22$)

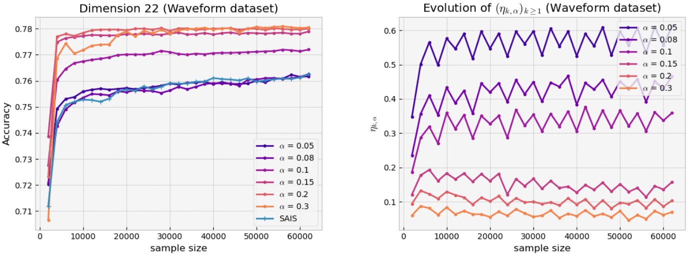

Figure: Left plot: Average accuracy over 100 trials of different learning policies $(\eta_{k,\alpha})_{k\geq 1}$ for Bayesian Logistic Regression on the Waveform dataset. Right plot: Averaged values of the learning policy $(\eta_{k,\alpha})_{k\geq 1}$ associated to each choice of α.

- 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

Future work:

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

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.