Adaptive Importance Sampling meets Mirror Descent: a Bias-Variance Tradeoff

Anna Korba
CREST, ENSAE, Institut Polytechnique de Paris

François Portier
CREST, ENSAI

Abstract

Adaptive importance sampling is a widely spread Monte Carlo technique that uses a reweighting strategy to iteratively estimate the so-called target distribution. A major drawback of adaptive importance sampling is the large variance of the weights which is known to badly impact the accuracy of the estimates. This paper investigates a regularization strategy whose basic principle is to raise the importance weights at a certain power. This regularization parameter, that might evolve between zero and one during the algorithm, is shown (i) to balance between the bias and the variance and (ii) to be connected to the mirror descent framework. Using a kernel density estimate to build the sampling policy, the uniform convergence is established under mild conditions. Finally, several practical ways to choose the regularization parameter are discussed and the benefits of the proposed approach are illustrated empirically.

1 INTRODUCTION

Many machine learning problems rely on our ability to compute the expected value of a function g of interest according to a target distribution f. Unfortunately, these calculations may be hindered by the fact that we are unable to sample from f, which is often intractable. Adaptive Importance Sampling (AIS) is one increasingly popular way to tackle this problem in the machine learning literature. It is notably used in stochastic optimization to achieve variance reduction and speed-up of stochastic gradient descent (El Hanchi and Stephens, 2020; Needell et al., 2014; Zhao and Zhang, 2015; Bouchard et al., 2015; Stich et al., 2017; Katharopoulos and Fleuret, 2018; Johnson and Guestrin, 2018; Csiba and Richtárik, 2018). In addition, it is commonly used in reinforcement learning (Huang and Jiang, 2020; Liu et al., 2020) and has also been integrated recently to the tuning of Bayesian neural networks (Ariafar et al., 2021).

The idea of AIS is to sample from an alternative, simpler *proposal* probability density q_k at time k of the algorithm to approximate f. It updates the proposal distribution at each iteration based on samples from the previous iteration and their associated importance sampling weights, in an online manner. AIS methods in the literature differ on the type or number of proposals, on the weighting scheme, and on the update of the proposal. Most methods use parametric proposals, e.g. Gaussians (Martino et al., 2015) or generic parametric proposals (Cornuet et al., 2012), and update the parameters of these proposals at each iteration based on weighted moment estimation of the target distribution. Alternatively, nonparametric methods were proposed using kernel density estimates of the target (West, 1993; Givens and Raftery, 1996; Zhang, 1996; Neddermeyer, 2009).

It is well-known that a bad proposal can lead to a high variance in high dimension (Bengtsson et al., 2008) and part of the art of using AIS then consists in selecting a *policy* $(q_k)_{k\geq 1}$, i.e. a sequence of proposals, that is practical and provides good efficiency gains, especially as the dimension increases (Oh and Berger, 1992; Owen and Zhou, 2000; Douc et al., 2007; Hartmann and Richter, 2021; Portier and Delyon, 2018), see Bugallo et al. (2017); Elvira and Martino (2021) for detailed reviews.

In this paper, we propose a new non parametric Adaptive Importance Sampling 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. In particular, (i) will be connected to the entropic mirror descent algorithm on the space

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

of probability distributions as in Dai et al. (2016) , while (ii) will be closely related to the safe adaptive importance sampling approach of Delyon and Portier (2021). We prove that this algorithm enjoys a uniform convergence result and we exhibit a way to carry out the regularisation procedure adaptively. On practical AIS applications, this procedure allows us to make significant speed-ups and performance gains empirically compared to the (classical) AIS framework.

The paper is organized as follows. In Section 2, we motivate our regularisation strategy by showing how it balances between the bias and the variance in an importance sampling setting. Furthermore, we underline its connections with an entropic mirror descent algorithm over the space of probability distributions. Section 3 is devoted to presenting our Safe and Regularised Adaptive Importance Sampling (SRAIS) approach. Notably, we state our theoretical guarantees and we describe our regularization schedule at each iteration. Lastly, our numerical results are to be found in Section 5.

In the whole paper, all densities are with respect to the Lebesgue measure and $\int g$ denotes the integral of g with respect to the Lebesgue measure.

2 MOTIVATION

2.1 Naive Importance Sampling

Given a target density function f defined on $\mathbb{R}^d$, our goal is to compute $\int gf$ for some (possibly many) integrable functions g. In many situations, it is impossible or too complicated to get samples from f and use a classical Monte Carlo estimator; in which case one can resort to Importance Sampling (IS) methods.

Letting q be another density function defined on $\mathbb{R}^d$ and assuming that X is a random variable with distribution q, the basic idea of IS is to re-weight $g(X)$ by $W(X) = f(X)/q(X)$, the so-called *importance weight*. We further say that q dominates f if for almost every $x \in \mathbb{R}^d$, $q(x) > 0$ implies that $f(x) > 0$. Whenever q dominates f, since $\mathbb{E}[W(X)g(X)] = \int gf$ and using i.i.d. samples $X_1, \dots, X_n \sim q$, one can build an (unbiased) IS estimator of $\int gf$ as

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

The previous estimate is referred to as *naive importance sampling* since it uses a single proposal q, in contrast to adaptive importance sampling (AIS). The weakness of naive importance sampling is that if the proposal q is not sufficiently close to f, the importance weights $(W(X_k))_{k=1}^n$ may have a large variance. However, a large variability of the importance weights naturally results in a high variance of the IS estimators Bengtsson et al. (2008). To tackle this problem, we introduce the concept of *regularized Importance Sampling*, which we describe in the next section.

Remark 1. In many applications, e.g. in Bayesian inference, the density f is often known only up to a normalization constant. In this case, one can use the normalized counterpart of (1), namely the *self-normalized importance sampling* estimator defined as $\sum_{k=1}^n W(X_k)g(X_k)/\sum_{k=1}^n W(X_k)$, where $X_1, \dots, X_n \sim q$; and which suffers from a small bias of order $\mathcal{O}(n)$ (Owen, 2013).

2.2 Regularized Importance Sampling

Regularized importance sampling consists in using *regularized* weights of the form $W(X)^\eta$, $\eta \in (0, 1)$ instead of the classical importance weights $W(X)$. In this way, and as written in Lemma 1 below, we will be able to diminish the variance at the cost of increasing the bias, which illustrates the typical bias-variance tradeoff and justifies the use of the term *regularized importance sampling*.

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

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

If $q \neq f$, then both inequalities become equalities if and only if $\eta = 1$.

The proof of Lemma 1 is deferred to Appendix A.1 and we now make a few comments. A first comment is that via a proper choice of the regularization parameter η, one will be able to balance between the bias and the variance in IS. A second comment is that if we consider the estimate $W(X)^\eta g(X)$ of $\int gf$ for some regularization parameter $\eta \in (0, 1)$, it has the following expectation

$$\mathbb{E}[W(X)^\eta g(X)] = \int f^\eta q^{1-\eta} g$$

when X has density q. This suggests that regularized importance sampling moves from the initial density q to the target density $f^\eta q^{1-\eta}$ (up to a normalization constant), while classical IS corresponding to the case $\eta = 1$, directly targets f.

Remark 2. The latter regularization differs from the one considered in simulated annealing (or tempering) (Zhou et al., 2016), where f^η interpolates between the uniform distribution and f, as (the *temperature*) η increases from 0 to 1. In this setting, the importance weights would be proportional to f^η/q instead of $(f/q)^\eta$ as proposed in this paper.

2.3 Connection with Mirror Descent

The regularization of IS weights introduced in Section 2.2 actually relates to a well-known optimization scheme called *entropic mirror descent* (Beck and Teboulle, 2003). This scheme has been widely used in the machine learning literature, e.g. in variational inference (Dai et al., 2016; Daudel et al., 2020), reinforcement learning (Orabona, 2019) (sometimes referred to continuous exponential weights), optimal transport (Mishchenko, 2019) or to train generative models (Hsieh et al., 2019). Given an initial probability density function q_1, the sequence of proposals $(q_k^\star)_{k\geq 1}$ obtained by entropic mirror descent for Kullback-Leibler minimisation satisfies:

$$q_{k+1}^\star \propto f^{\eta_k}(q_k^\star)^{1-\eta_k}\ , \quad k \geq 1\ , \tag{2}$$

where $(\eta_k)_{k\geq 1}$ is a sequence of positive learning rates (we refer to Appendix B for a derivation of this iterative formula from an optimisation perspective). When $(\eta_k)_{k\geq 1}$ is valued in $(0,1]$, which will be assumed throughout the paper, the iterative scheme defined by (2) is attractive as it satisfies the following lemma, whose proof can be found in Appendix A.2.

Lemma 2. Let $(\eta_k)_{k\geq 1}$ valued in $(0,1]$ and $(q_k^\star)_{k\geq 1}$ be defined by (2) starting from an initial probability density function q_1. Then, for all $n \in \mathbb{N}^\star$,

$$\int |f - q_{n+1}^\star| \leq \sqrt{2\,\mathrm{KL}\,(f\|q_1)}\prod_{k=1}^{n}(1-\eta_k)^{1/2},$$

where $\mathrm{KL}\,(f\|q_1) = \int \log(f/q_1)\, f$ is the Kullback-Leibler divergence between f and q_1.

Remark 3. When $\mathrm{KL}\,(f\|q_1)$ is finite, Lemma 2 yields different convergence rates depending on the regularization schedule. Taking $\eta_k = c/k$ with $0 < c < 1$ yields an $O(n^{-c/2})$ convergence rate, while taking $\eta_k = c/k^\beta$ with $0 < c < 1$ and $\beta \in [0,1)$ yields an $O(\exp(-Cn^{(1-\beta)}))$ convergence rate, for some $C > 0$ (details are given in Appendix A.2 right after the proof of Lemma 2). Note that we recover the case of the constant learning policy when $\beta = 0$. This is in contrast with convergence results from the optimisation literature, as recalled in Lemma 7 of Appendix B. In the latter, convergence is established under stronger assumptions on the iterates $q_k^\star$ (namely, boundedness of the log ratios $\log(q_k^\star/f)$) and on the learning policy (e.g, $\eta_k = c/\sqrt{k}$, i.e. $\beta = 1/2$, which yields a $O(\log(n)/\sqrt{n})$ convergence rate).

Unfortunately, running iteration (2) in practice is not feasible for two reasons: (i) each iteration depends on the whole density f (not some evaluations of f), which is unknown in many applications; (ii) even if f was known, computing $q_{k+1}^\star(x)$ for $x \in \mathbb{R}^d$ would still be difficult due to the normalization following (2) which ensures $q_{k+1}^\star$ is a probability density function. The AIS scheme proposed in the next section might be seen as a feasible stochastic version of iteration (2) that overcomes the two previous problems.

3 SAFE AND REGULARIZED ADAPTIVE IMPORTANCE SAMPLING

This section introduces Safe and Regularized Adaptive Importance Sampling and relates it to the entropic mirror descent algorithm presented in the previous section.

3.1 Proposed Scheme

We propose an *Adaptive Importance Sampling* (AIS) method, which in contrast to classical Importance Sampling (1) that uses a unique proposal distribution q, uses a *policy*, i.e. a sequence of proposals $(q_k)_{k\geq 0}$ in an online manner. We define this notion rigorously below.

Definition 1. Let $(X_k)_{k\geq 1}$ a sequence of random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$, and define the σ-algebra $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$ for $k \geq 1$ and $\mathcal{F}_0 = \emptyset$. The sequence $(X_k)_{k\geq 1}$ is said to have policy $(q_k)_{k\geq 0}$ whenever for all $k \geq 1$, $X_k \sim q_{k-1}$, conditionally to $\mathcal{F}_{k-1}$.

However, in AIS, one should ensure that the proposal q_k sufficiently covers the support of the target f at each iteration $k \geq 0$. Among recent approaches, it has been proposed in Delyon and Portier (2021) to choose the policy $(q_k)_{k\geq 0}$ as a mixture between a certain kernel density estimate f_k of f and some "safe" probability density q_0 with heavy tails compared to the ones of f. Letting $(\lambda_k)_{k\geq 1}$ valued in $[0,1]$ be a decreasing sequence, we seek $(q_k)_{k\geq 0}$ under the form

$$q_k = (1-\lambda_k)f_k + \lambda_k q_0\ , \qquad \forall k \geq 1\ . \tag{3}$$

In this way, the mixture with q_0 shall prevent too small values for q_k and a too high variance on the importance weights Delyon and Portier (2021). In contrast, the other part in the mixture, f_k, is meant to speed-up the convergence of q_k to f.

Let us now define the kernel density estimate f_k of f more precisely, which will differ from the one used in Delyon and Portier (2021) (that does not rely on regularized importance weights). For this purpose, let $K : \mathbb{R}^d \to \mathbb{R}^+$ be a density called *kernel* and for a *bandwith* $h \geq 0$, define $K_h(u) = K(u/h)/h^d$. Let $(h_k)_{k\geq 1}$ valued in $\mathbb{R}^+$ be a sequence of *bandwidths*. The kernel estimate of f at step n denoted f_n, is defined by

$$f_n(x) = \sum_{k=1}^{n} W_{n,k}^{(\eta_k)} K_{h_n}(x - X_k)\ , \qquad \forall x \in \mathbb{R}^d, \tag{4}$$

where for all $k = 1, \dots, n$,

$$W_{n,k}^{(\eta_k)} \propto W_k^{\eta_k} = \left(\frac{f(X_k)}{q_{k-1}(X_k)} \right)^{\eta_k} \quad \text{such that} \quad \sum_{k=1}^{n} W_{n,k}^{(\eta_k)} = 1. \quad (5)$$

In other words, f_n is based on the kernel evaluated at points $(X_k)_{k=1}^n$, where each point X_k is weighted by $W_{n,k}^{(\eta_k)}$ for $k = 1, \dots, n$. The weights $(W_{n,k}^{(\eta_k)})_{k=1}^n$ correspond to the *normalized* counterpart of the *regularized* (i.e., raised to the power η_k) importance weights $(W_k)_{k=1}^n$. The algorithm resulting from the previous description can be written as follows.

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

Safe and Regularized Adaptive Importance Sampling (SRAIS) can be decomposed in three steps: (1) sample a new particle from the current proposal, (2) compute (a) the new importance weight and (b) its normalized, regularized counterpart, (3) update the kernel density estimate of the target, and the proposal as the mixture of this estimate and the safe density.

In Algorithm 1, we recognize at each iteration k a stochastic approximation of the mirror descent scheme (2), as the expectation of $W_k^{\eta_k} K_{h_k}(x - X_k)$ w.r.t. q_k is equal to $(f^{\eta_k} q_{k-1}^{1-\eta_k} \star K_{h_k})(x)$, where $\star$ denotes the convolution operator. From classical convolution results (Tsybakov, 2008), the latter quantity approximates $f^{\eta_k} q_{k-1}^{1-\eta_k}$ when the bandwith h_k is small. Notice also that in Algorithm 1, at each step k, the bandwith h_k is shared across all the generated points $(X_j)_{j=1}^{k+1}$. However, one could choose a different bandwith for each point, as this does not have an impact on our theoretical results (see Remark 5 in the Appendix).

3.2 Uniform Convergence of the Scheme

We now turn to the theoretical analysis of Algorithm 1. Let us first introduce, for all $n \ge 1$, and for any $x \in \mathbb{R}^d$,

$$N_n(x) = \frac{1}{n} \sum_{k=1}^{n} W_k^{\eta_k} K_{h_n}(x - X_k), \; D_n = \frac{1}{n} \sum_{k=1}^{n} W_k^{\eta_k}.$$

Using this notation, the sequence $(f_n)_{n \ge 1}$ from (4) can be rewritten under the form $f_n(x) = N_n(x)/D_n$, for all $n \ge 1$, $x \in \mathbb{R}^d$. Here, the analysis is carried out by examining separately $(D_n)_{n \ge 1}$ and $(N_n)_{n \ge 1}$. In particular the function N_n, which is an unnormalized smooth density estimate of f, can be decomposed as follows:

$$N_n - f = M_n + \left\{ \left(\frac{1}{n} \sum_{k=1}^{n} (f^{\eta_k} q_{k-1}^{1-\eta_k} - f) \right) \star K_{h_n} \right\} + \left\{ f \star K_{h_n} - f \right\}, \quad (6)$$

where the function M_n is defined by, for all $x \in \mathbb{R}^d$,

$$M_n(x) = \frac{1}{n} \sum_{k=1}^{n} \{ W_k^{\eta_k} K_{h_n}(x - X_k) - \{ f^{\eta_k} q_{k-1}^{1-\eta_k} \} \star K_{h_n}(x) \big\}. \quad (7)$$

The previous decomposition (6) sheds some light on the approach taken in the mathematical proof. Indeed, the two last terms on the right hand side of this decomposition represent two different types of bias. The first one with $f^{\eta_k} q_{k-1}^{1-\eta_k} - f$ captures the bias induced by the regularization (it is therefore directly related to the mirror descent update) and should be negligible when $\eta_n \to 1$ as $n \to \infty$. The second one with $f \star K_{h_n} - f$ is well studied in nonparametric statistics Tsybakov (2008) and goes to 0 as soon as $h_n \to 0$ when $n \to \infty$. Finally, the first term M_n in (6) is an average of martingale increments (see Proposition 10 in the Appendix), which will help our theoretical analysis via Freedman-type concentration inequalities (Freedman, 1975; Bercu et al., 2015).

Let us move on to our theoretical results and introduce the following assumptions:

($\mathbf{A}_1$)(i) The sequence $(\lambda_k)_{k \ge 1}$ 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 $(h_k)_{k \ge 1}$ is valued in $\mathbb{R}^+$, nonincreasing, and $\lim_{k \to \infty} h_k = 0$ and $\lim_{k \to \infty} \log(k)/(k h_k^d \lambda_k) = 0$.

(iii) The sequence $(\eta_k)_{k \ge 1}$ 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, bounded by $f_\infty \geq 0$ and is L_f-Lipschitz.

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

In ($\mathbf{A}_1$), the reasonably slow convergence of $(\lambda_k)_{k\geq 1}$ to 0 helps the proposal to visit a large part of the space, while the faster convergence of $(h_k)_{k\geq 1}$ and $(\eta_k)_{k\geq 1}$ to 0 and 1 respectively enables to remove the bias asymptotically. Then, ($\mathbf{A}_2$) enables to control the variance induced by the weights whatever the proposal q_k. In practice, since q_0 is chosen by the user, this assumption is relatively weak. For instance, ($\mathbf{A}_2$) is valid when f has exponential decay and q_0 is any member of the Student's t-distributions family (polynomial decay). Finally, the regularity of f and K required in ($\mathbf{A}_3$)-($\mathbf{A}_4$) is used to ensure the convergence of $(f \star K_{h_k})$ towards f. ($\mathbf{A}_3$) is valid for many families of distributions (e.g., Student's, Gaussian, beta, exponential and any mixture of these densities) and ($\mathbf{A}_4$) is satisfied by most kernels (Gaussian, beta, Epanechnikov,...), see Tsybakov (2008).

Under these assumptions, we first obtain the following uniform convergence result for the estimate (4), whose proof is given in Appendix C.

Proposition 3. Assume ($\mathbf{A}_1$), ($\mathbf{A}_2$), ($\mathbf{A}_3$) and ($\mathbf{A}_4$). Then, for any $r > 0$, we have that almost-surely

$$\sup_{\|x\|\leq n^r} |f_n(x) - f(x)| \to 0 \quad \text{as } n \to \infty.$$

The previous result is now extended to a result over the whole space. If we strenghten our assumptions on q_0 and K as follows, we can obtain a stronger uniform convergence result.

($\mathbf{A}_5$) There exist $(c, \delta) \in (\mathbb{R}^+)^2$, such that for all $x \in \mathbb{R}^d$, $c(1 + \|x\|^\delta)f(x) \leq q_0(x)$.

($\mathbf{A}_6$) There exist $(C_K, r_K) \in (\mathbb{R}^+)^2$, such that for all $x \in \mathbb{R}^d$, $K(x) \leq C_K(1 + \|x\|)^{-r_K}$.

Assumptions ($\mathbf{A}_5$) and ($\mathbf{A}_6$) are not too restrictive, as the examples of f and q_0 as well as the kernel functions given before are still valid under this new framework. Because ($\mathbf{A}_5$) is stronger than ($\mathbf{A}_2$), the latter will no longer be needed in our second result, whose proof is given in Appendix D.

Proposition 4. Assume ($\mathbf{A}_1$), ($\mathbf{A}_3$), ($\mathbf{A}_4$), ($\mathbf{A}_5$) and ($\mathbf{A}_6$). Then, we have that almost-surely

$$\sup_{x\in\mathbb{R}^d} |f_n(x) - f(x)| \to 0 \quad \text{as } n \to \infty.$$

Remark 4. With some additional notations, Algorithm 1 can be extended to a mini-batch version where a fixed number of particles $m_k \geq 1$ is generated at each iteration instead of $m_k = 1$. Such a choice does not impact our proof since the martingale property we use in the proofs of convergence remains valid. This encouraged us to use such a batch of particles to construct the sequence $(\eta_k)_{k\geq 1}$ in an *adaptive* manner, as explained in the following Section 3.3.

3.3 Adaptive Choice of the Regularization Parameter

We have explicited in Section 3.2 the conditions on the hyperparameters of Algorithm 1, so that the latter consistently estimates integrals of interest with respect to the target f. In particular, it requires that the sequence of regularization parameters $(\eta_k)_{k\geq 1}$ converges to 1. Here, we propose a practical and adaptive way to construct such a sequence. It relies on the following idea: when the proposal q_k is equal to f, then all the importance weights are equal to 1, resulting in a uniform distribution over the particles sampled in Algorithm 1. The quality of the proposal at each iteration with respect to the target can thus be reformulated as quantifying how far is the current distribution of the importance weights from the uniform distribution. Intuitively, penalizing the divergence between the distribution over particles reweighted by the importance weights, with respect to the uniform distribution, penalizes a high variance for the importance weights. We choose to consider Renyi's α-divergence (Rényi et al., 1961), as it will enable us to influence how fast the regularization evolves.

At each time k, we draw a batch of m_k i.i.d samples $X_{k,1}, \ldots, X_{k,m_k}$ from q_{k-1}, and compute their associated normalized IS weights $W_{k,1}, \ldots, W_{k,m_k}$, i.e. $W_{k,\ell} \propto f(X_{k,\ell})/q_{k-1}(X_{k,\ell})$ for all $\ell = 1 \ldots m_k$ such that $\sum_{\ell=1}^{m_k} W_{k,\ell} = 1$. Notice that these are not regularized as in (5). Now, let $\mathbb{P} = \sum_{l=1}^{m_k} W_{k,l}\delta_{X_{k,l}}$ and $\mathbb{Q} = \sum_{l=1}^{m_k} {}^1\!/_{m_k}\delta_{X_{k,l}}$ the reweighted and uniform distribution on the particles $(X_{k,l})_{l=1}^{m_k}$ respectively. Renyi's α-divergence (Rényi et al., 1961) of $\mathbb{P}$ from $\mathbb{Q}$ is defined for $\alpha \in \mathbb{R} \setminus \{0, 1\}$ by

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

It can be extended by continuity to the Kullback-Leibler by letting $\alpha \to 1$ (we use the notation $D_1(\mathbb{P}||\mathbb{Q})$).

Notably, for all $\alpha \in [0,1]$, see (van Erven and Harremoes, 2014, Theorem 3), it holds that $D_\alpha(\mathbb{P}||\mathbb{Q})$ is non-decreasing in α. At time k, for $\alpha \in \mathbb{R} \setminus \{0,1\}$, we propose to set the regularisation parameter as

$$\eta_{k,\alpha} := 1 - \frac{D_\alpha(\mathbb{P}||\mathbb{Q})}{\log(m_k)} \,. \tag{8}$$

Choosing the regularization parameter as (8) is particularly convenient, since it does not require to use a pre-defined sequence $(\eta_k)_{k\geq 1}$ as written in Algorithm 1. In contrast, the regularization is chosen adaptively at each iteration, depending on how far is the current estimate to f. The following proposition, whose proof can be found in Appendix E guarantees that our proposal (8) is a valid candidate.

Proposition 5. Let $\alpha \in [0,1]$ and let $(\eta_{k,\alpha})_{k\geq 1}$ be the sequence defined by (8) for all $k \geq 1$. Then, we have:

(i) The sequence $(\eta_{k,\alpha})_{k\geq 1}$ is valued in $[0,1]$, with $\eta_{k,\alpha} = 1$ iff $\mathbb{P} = \mathbb{Q}$;

(ii) $0 \leq \eta_{k,1} \leq \eta_{k,\alpha} \leq 1$;

(iii) 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.

By Proposition 5-(ii), one can possibly increase η_k by decreasing α. The procedure resulting from the previous description can be written as in Algorithm 2.

Algorithm 2 *Renyi's Adaptive Regularization (RAR)*

Input: iteration k, number of samples m, $\alpha \in [0,1]$. For $l = 0, 1, \ldots, m$:

(i) Generate $X_{k,1}, \ldots, X_{k,m}$ from q_{k-1} (from Algorithm 1).

(ii) Compute the normalized importance weights $W_{k,l} \propto f(X_{k,l})/q_{k-1}(X_{k,l})$ such that $\sum_{l=1}^m W_{k,l} = 1$.

(iii) Return $\eta_{k,\alpha} = 1 - D_\alpha(\mathbb{P}||\mathbb{Q})/\log(m)$, where $\mathbb{P} = \sum_{l=1}^m W_{k,l}\delta_{X_{k,l}}$ and $\mathbb{Q} = \sum_{l=1}^m {}^1\!/_m \delta_{X_{k,l}}$.

4 RELATED WORK

Particle Mirror Descent (Dai et al., 2016). Closely related to our scheme Section 3 is the one proposed in Dai et al. (2016), that also approximates entropic mirror descent. Their proposal is to approximate the (intractable) $q^\star_{k+1}$ of the mirror descent update (2) by a weighted kernel density estimate centered at m_k particles sampled from q_k (their previous approximate of $q^\star_k$), and update the weights as (5). Our algorithm is different in nature from the one proposed in Dai et al. (2016), which requires that both the number of iterations k and number of particles at each iteration m_k should grow to infinity. First, we only need one new particle at each iteration, i.e. $m_k = 1$ (but the theoretical guarantees hold even if $m_k \geq 1$, see Remark 4), which means that our approach requires a single evaluation of f at each iteration; and we use all particles generated at times $1 \leq j \leq k$. This is a major improvement making our approach in line with the spirit of *stochastic approximation* (Kushner and Yin, 2003), where each iteration should be of small cost. Second, our proposal is a mixture of f_k, that approximates (2), and a safe density q_0 which allows in practice to visit exhaustively the underlying domain. Finally, our theoretical results also differ from the ones of (Dai et al., 2016, Theorem 6). In our setting, the sequence $(\eta_k)_{k\geq 1}$ is not intended to go to 0 as in Dai et al. (2016), but to 1 in order to realize the bias-variance trade-off described in Section 2. Also in contrast with the results of (Dai et al., 2016, Theorem 6), our results of Proposition 3 and 4 are free from any restriction on the behavior of the sequence $(q_k)_{k\geq 1}$, as we only impose conditions on the safe density $(\mathbf{A}_2)$ and $(\mathbf{A}_5)$. Their results assume that q_k is bounded away from 0, which is unlikely to be satisfied in practice if it does not rely on a safe density.

Safe Adaptive Importance Sampling (SAIS) (Delyon and Portier, 2021). The SAIS algorithm of Delyon and Portier (2021) corresponds to the particular case where the sequence $(\eta_k)_{k\geq 1}$ is constant and equal to 1 in Algorithm 1. In Delyon and Portier (2021), the authors obtain uniform rates of convergence when the sequence $(\lambda_k)_{k\geq 0}$ satisfies $(\mathbf{A}_1)$(i). Given that our proposal SRAIS recovers SAIS as a particular case, the consistency results in Section 3.2 extend in some sense the ones of Delyon and Portier (2021). Moreover our adaptive choice for the regularization given by Renyi's Adaptive Regularization (RAR) is novel, and the final approach combining Algorithm 1 and 2 outperforms SAIS ($\eta_k = 1$) numerically, as will be illustrated in Section 5.

Regularization of IS weights. Nonlinear transformation of the weights have been promoted in the literature of IS previously. For instance, truncated IS (or non linear IS) (Ionides, 2008) Koblents and Míguez (2015), clips the weights W as $W' = \min(W, \tau)$ where τ is some threshold. In Gramacy et al. (2010), it is proposed to elevate IS weights to some regularization parameter η, similarly to our work. This power η is set to maximize the Effective Sample Size (ESS, see eq. 6 and Prop 2.1 therein) instead of Renyi's α-divergence between

the distribution of the importance weights and the uniform. Also these methods are proposed in the naive (i.e., non adaptive) importance sampling framework. In Finke and Thiery (2019) (see Appendix A.3 there) is proposed a similar strategy to Gramacy et al. (2010) but in the AIS setting. It differs at least in two ways from our approach. First, the theoretical guarantees of such a regularization schedule are not investigated in their paper while we establish the convergence of $(\eta_k)_{k\geq 1}$ to 1 in Proposition 5 so that our requirements for uniform convergence are satisfied. Second, in the previous work, the regularization parameters $(\eta_k)_{k\geq 1}$ are chosen again to maximize the ESS, by running, at each iteration, a bisection algorithm. It appears more expensive than the evaluation of the closed form given in Algorithm 1 which is only of complexity $O(m)$ where m is the size of the batch.

Sequential Monte Carlo (SMC). The algorithm proposed in this paper is related to the SMC framework. These methods are usually viewed as performing adaptive importance sampling whilst also allowing for the target distribution to change over iterations. The targets sequence might be chosen as the entropic mirror descent iteration as recommended for instance in Del Moral et al. (2006), eq. 7. The resulting SMC algorithm might benefit from the same advantage as our approach, i.e. a small variance for the estimate of the target. However, the particles in traditional SMC (Chopin, 2004; Del Moral et al., 2006) are moved in a slightly different manner as in the proposed AIS algorithm. When running an SMC algorithm, each particle is moved around its "parent" (a random walk move called mutation) and then, among the new particles obtained, some are removed based on their weights values (a subsampling step called selection). Our algorithm, whose particles distribution is given by the mixture $(q_k)_{k)\geq 1}$, is similar, except that we select the "parents" among the whole set of previous particles (in this way a particle is not removed in our version). The presence of this selection mechanism in SMC suggests some similarities with the popular random walk Metropolis in which a non profitable particle would be removed with high probability. Our algorithm, whose particles distribution is only $(q_k)_{k)\geq 1}$, is simpler. This explains why we obtain different results based on different proofs techniques than the ones used in the SMC literature. While SMC theoretical guarantees have been expressed as central limit theorems (Chopin, 2004), we have obtained almost sure convergence of the sequence of proposals.

5 EXPERIMENTS

In this section, we discuss the performance of Algorithm 1 along with the subroutine Algorithm 2 to approximate integrals of interest w.r.t. a target distribution f, on toy and real-world examples. Since renormalized weights are used in Algorithm 1-2, these experiments only require to know the target distribution up to its normalization constant. A Gaussian kernel is used in all experiments. The code to reproduce the experiment is available at `https://github.com/akorba/Safe_And_Regularized_Importance_Sampling`.

5.1 Toy Examples

Gaussian mixtures. Let us now describe three interesting examples starting by the simplest and finishing with the most difficult case. Denote by ϕ_Σ the multivariate Gaussian density with mean zero and variance Σ. Denote also $\boldsymbol{I}_d$ the identity matrix and $\mathbf{1}_d$ the d-dimensional vector whose coordinates are all equal to 1. The *Cold Start* case is when the target density is a Gaussian given by $f_1(x) = \phi_\Sigma(x - 5\mathbf{1}_d/\sqrt{d})$, $x \in \mathbb{R}^d$; where $\Sigma = (0.16/d)\boldsymbol{I}_d$ and the starting density is the multivariate Student distribution with mean zero and variance $(5/d)\boldsymbol{I}_d$. Note the division by $\sqrt{d}$ which preserves the Euclidean distance between the two centers of the starting distribution and the target, whatever the dimension. The same explanation justifies the normalization of the variances by $1/d$. The *Gaussian mixture* example corresponds to the target density $f_2(x) = 0.5\phi_\Sigma(x - \mathbf{1}_d/(2\sqrt{d})) + 0.5\phi_\Sigma(x + \mathbf{1}_d/(2\sqrt{d}))$, $x \in \mathbb{R}^d$. The initial density is the multivariate student distribution with mean $(1, -1, 0, \ldots, 0)/\sqrt{d}$ and covariance matrix $(5/d)\boldsymbol{I}_d$. The initial mean value differs from zero to prevent the naive algorithm using the initial density from having good results (due to the symmetry of the target). The *Anisotropic Gaussian Mixture* case is similar to the previous example, except that it is unbalanced and that each Gaussian in the mixture is anisotropic. The target is given by $f_3(x) = 0.25\phi_V(x-\mathbf{1}_d/(2\sqrt{d}))+0.75\phi_V(x+\mathbf{1}_d/(2\sqrt{d}))$, $x \in \mathbb{R}^d$; with $V = (.4/\sqrt{d})^2\mathrm{diag}(10, 1, \ldots, 1)$. The starting density is the multivariate student distribution with mean $(1, -1, 0, \ldots, 0)/\sqrt{d}$ and covariance matrix $(5/d)\boldsymbol{I}_d$.

Competitive methods. We focus first on four constant values for η for Algorithm 1, respectively $1, .75, .5, .25$. While $\eta = 1$ is an unbiased estimate with large variance, $\eta = .25$ is biased but has smaller variance. We also run Algorithm 1 when $(\eta_k)_{k\geq 1}$ is tuned via Algorithm 2 (RAR), with parameter $\alpha = .5$. All algorithms are initialised by sampling independently 4×10^4 particles from the initial density. We further sample independently $N = 18 \times 10^3$ particles from q_k at each time $k \geq 1$, for a total computational budget of 4×10^5 particles (i.e $n = 20$ iterations). We follow the recommendation given in Delyon and Portier (2021) which consists in running a subsampling procedure for the KDE estimate of size $\ell_k = \sqrt{N_k}$ when N_k

denotes the total amount of particles at step k. These particles are used to construct the kernel estimate f_k, improving significantly the computing time. We take $\lambda_k \propto \ell_k^{-2/(4+d)}$ and $h_k \propto \ell_k^{-1/(4+d)}$ as recommended in our theoretical study.

Error evaluation. For each method in competition, the evaluation of the mean squared error (MSE) is made by computing the average of $\|\mu_f - \hat{\mu}\|_2^2$ over 50 runs of the method where μ_f is the mean of the target and $\hat{\mu} = \sum_{k=1}^n W_{k+1,j}^{(\eta_j)} X_j$ stands for the estimated mean. Note that the examples where chosen in a way that the error computing the mean reflects the general behavior of the method. In each case, we plot the error at each iteration of the algorithm. Our results are presented in Figures 1 and 2 in dimension $d = 16$ (similar results are have been obtained for $d = 4$ and $d = 8$ but are deferred Appendix F).

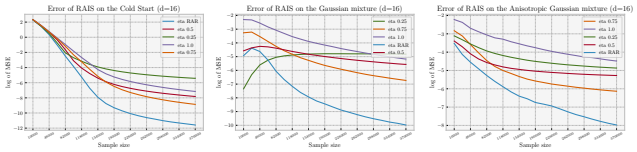

Figure 1: Logarithm of the average squared error for Algorithm 1 for constant values of η or Algorithm 2, computed over 50 replicates.

Figure 1 shows that the sequence $(\eta_k)_{k\geq 1}$ has a strong influence on the outcome of the procedure and an adaptive choice of $(\eta_k)_{k\geq 1}$ (reflecting the quality of the current proposal) leads to a substantial improvement. An explanation of the performance of Algorithm 1 along with Algorithm 2 can be found in Figure 2 where we see that in all cases, 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); and when the policy becomes better the value of η_k converges to 1.

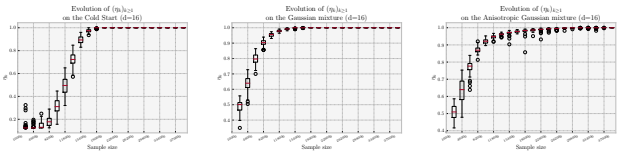

Figure 2: Boxplot of the values of $(\eta_{k,\alpha})_{k\geq 1}$ obtained from Algorithm 2, with $\alpha = 0.5$.

5.2 Bayesian Logistic Regression

We consider the Bayesian Logistic Regression setting of Gershman et al. (2012), also considered in the recent Bayesian inference literature (Liu and Wang, 2016; Daudel et al., 2020). More precisely, we observe a dataset $\mathcal{D} = \{c_i, \boldsymbol{z}_i\}_{i\in I}$ of points $\boldsymbol{z}_i \in \mathbb{R}^L$ and binary class labels $c_i \in \{-1, 1\}$, for $i \in I$. We assume the following model : $p_0(\beta) = \text{Gamma}(\beta; a, b)$, and for any $1 \leq l \leq L$ and $1 \leq i \leq I$,

$$p_0(\omega_\ell|\beta) = \mathcal{N}(\omega_\ell; 0, \beta^{-1}),\ p(c_i = 1|\boldsymbol{z}_i, \boldsymbol{\omega}) = \frac{1}{1 + e^{-\boldsymbol{\omega}^T \boldsymbol{z}_i}},$$

where a and b are hyperparameters (shape and inverse scale) that are fixed to $a = 1$ and $b = 0.01$, and $\Gamma, \mathcal{N}$ denote the Gamma and Gaussian distribution respectively. The parameter vector is then $[\boldsymbol{\omega}, \beta] \in \mathbb{R}^d$, with $\beta \in \mathbb{R}^+$ and $d = L + 1$. Given a new data point $\boldsymbol{z}_{\text{new}}$, we are interested in predicting the label c_{new} using the *posterior predictive distribution* $p(c_{\text{new}}|\boldsymbol{z}_{\text{new}}, \mathcal{D}) = \int p(c_{\text{new}}|\boldsymbol{z}_{\text{new}}, x)p(x|\mathcal{D})\text{d}x$, which plays the role of the target distribution.

Since the posterior predictive distribution is intractable for this choice of model, we resort to Algorithm 1 to approximate this quantity. We consider two main regimes for the regularization scheme in our numerical experiments: (i) constant policy $\eta_k = 1$ for all $k \geq 1$: this case corresponds to Safe Adaptive Importance Sampling algorithm (Delyon and Portier, 2021), with no bias but the highest variance, (ii) adaptive policy $(\eta_{k,\alpha})_{k\geq 1}$ from Algorithm 2, with $\alpha \in \{0.3, 0.2, 0.1, 0.08\}$. Here, the bandwith of the kernel estimate h_k is chosen to be proportional to $k^{-1/(4+d)}$. This corresponds to the optimal choice in nonparametric estimation when the target density f is at least 2-times continuously differentiable and the kernel has order 2 (Stone, 1982). We also choose λ_k proportional to $1/\sqrt{k}$ for all $k \geq 1$. Furthermore, q_0 is chosen as a Gaussian distribution with mean $0.\mathbf{1}_d$ and covariance matrix $5.\boldsymbol{I}_d$, where $\mathbf{1}_d$ is the d-dimensional vector whose coordinates are all equal to 1 and $\boldsymbol{I}_d$ is the identity matrix. Finally, our algorithms are initialised by sampling independently 2000 particles from q_0. We further sample independently 200 particles from q_k at each time $k \geq 1$, for a total computational budget of 61800 particles (i.e $n = 300$ iterations).

We test these algorithms on the Waveform dataset composed of 5000 datapoints in $\mathbb{R}^L$ with $L = 21$, i.e. $d = 22$. Figure 3 displays the average accuracy of the predictions on the testing set and the averaged values of the sequence $(\eta_{k,\alpha})_{k\geq 1}$ for $\alpha \in \{0.3, 0.2, 0.1, 0.08\}$, where the average is made over 100 replicates of the experiment.

As stated in Proposition 5(ii) and can be seen on the right plot of Figure 3, the lower the value of α, the higher the value of $\eta_{k,\alpha}$ at time k. Recalling that the level of confidence in high values of the learning policy is expressed through the parameter α, observe then on the Left plot of Figure 3 that a proper tuning of the

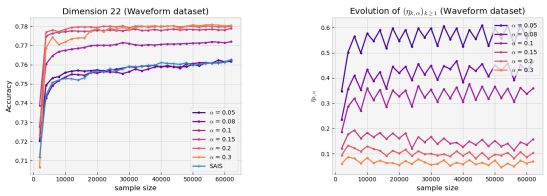

Figure 3: Left plot: Average accuracy over 100 trials of different learning policies $(\eta_{k,\alpha})_{k\geq 1}$ for Bayesian Logistic Regression on the Waveform dataset. Right plot: Averaged values of the learning policy $(\eta_{k,\alpha})_{k\geq 1}$ associated to each choice of α.

parameter α allows us to outperform the Safe Adaptive Importance Sampling case (Algorithm 1 with $(\eta_k)_{k\geq 1}$ constant and equal to 1), the case $\alpha = 0.2$ yielding the best results here overall in terms of speed and accuracy.

6 CONCLUSION

We proposed a new algorithm for Adaptive Importance Sampling, that regularizes the importance weights by raising them to a certain power η, and is related to mirror descent on the space of probability distributions. We prove that it enjoys a uniform convergence guarantee, under mild assumptions on the regularity of the target distribution and safe density, and hyperparameters of the algorithm. Furthermore, by proposing an adaptive way to schedule the regularization, we provide a method that achieves the best numerical performances, compared to a constant regularization or classical safe adaptive importance sampling. The results that we obtained could be very impactful for machine learning tasks where importance sampling is used, e.g. Bayesian inference, but also stochastic optimization or reinforcement learning. Future work includes a non asymptotic analysis of our scheme.

References

Setareh Ariafar, Zelda Mariet, Dana Brooks, Jennifer Dy, and Jasper Snoek. Faster & more reliable tuning of neural networks: Bayesian optimization with importance sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 3961–3969. PMLR, 2021.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3): 167–175, 2003.

Thomas Bengtsson, Peter Bickel, Bo Li, et al. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and statistics: Essays in honor of David A. Freedman*, pages 316–334. Institute of Mathematical Statistics, 2008.

Bernard Bercu, Bernard Delyon, and Emmanuel Rio. *Concentration inequalities for sums and martingales*. Springer, 2015.

Guillaume Bouchard, Théo Trouillon, Julien Perez, and Adrien Gaidon. Accelerating stochastic gradient descent via online learning to sample. *arXiv preprint arXiv:1506.09016*, 2015.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 01 2015. ISSN 1935-8237. doi: 10.1561/2200000050. URL `http://dx.doi.org/10.1561/2200000050`.

Monica F Bugallo, Victor Elvira, Luca Martino, David Luengo, Joaquin Miguez, and Petar M Djuric. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, 2017.

Nicolas Chopin. Central limit theorem for sequential monte carlo methods and its application to bayesian inference. *The Annals of Statistics*, 32(6):2385–2411, 2004.

Jean-Marie Cornuet, Jean-Michel Marin, Antonietta Mira, and Christian P Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, 2012.

Dominik Csiba and Peter Richtárik. Importance sampling for minibatches. *The Journal of Machine Learning Research*, 19(1):962–982, 2018.

Bo Dai, Niao He, Hanjun Dai, and Le Song. Provable bayesian inference via particle mirror descent. In *Artificial Intelligence and Statistics*, pages 985–994, 2016.

Kamélia Daudel, Randal Douc, and François Portier. Infinite-dimensional gradient-based descent for alpha-divergence minimisation. *arXiv preprint arXiv:2005.10618*, 2020.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

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

Randal Douc, Arnaud Guillin, J-M Marin, and Christian P Robert. Minimum variance importance sampling via population monte carlo. *ESAIM: Probability and Statistics*, 11:427–447, 2007.

Ayoub El Hanchi and David Stephens. Adaptive importance sampling for finite-sum optimization and

sampling with decreasing step-sizes. *Advances in Neural Information Processing Systems*, 33, 2020.

Víctor Elvira and Luca Martino. Advances in importance sampling. *arXiv preprint arXiv:2102.05407*, 2021.

Axel Finke and Alexandre H Thiery. On importance-weighted autoencoders. *arXiv preprint arXiv:1907.10477*, 2019.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Samuel Gershman, Matthew D. Hoffman, and David M. Blei. Nonparametric variational inference. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

Geof H Givens and Adrian E Raftery. Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *Journal of the American Statistical Association*, 91(433):132–141, 1996.

Robert Gramacy, Richard Samworth, and Ruth King. Importance tempering. *Statistics and Computing*, 20 (1):1–7, 2010.

Carsten Hartmann and Lorenz Richter. Nonasymptotic bounds for suboptimal importance sampling. *arXiv preprint arXiv:2102.09606*, 2021.

Ya-Ping Hsieh, Chen Liu, and Volkan Cevher. Finding mixed nash equilibria of generative adversarial networks. In *International Conference on Machine Learning*, pages 2810–2819. PMLR, 2019.

Jiawei Huang and Nan Jiang. From importance sampling to doubly robust policy gradient. In *International Conference on Machine Learning*, pages 4434–4443. PMLR, 2020.

Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Tyler B Johnson and Carlos Guestrin. Training deep models faster with robust, approximate importance sampling. *Advances in Neural Information Processing Systems*, 31:7265–7275, 2018.

Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *International conference on machine learning*, pages 2525–2534. PMLR, 2018.

Eugenia Koblents and Joaquín Míguez. A population monte carlo scheme with transformed weights and its application to stochastic kinetic models. *Statistics and Computing*, 25(2):407–425, 2015.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 2016.

Yao Liu, Pierre-Luc Bacon, and Emma Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. In *International Conference on Machine Learning*, pages 6184–6193. PMLR, 2020.

Luca Martino, Victor Elvira, David Luengo, and Jukka Corander. An adaptive population importance sampler: Learning from uncertainty. *IEEE Transactions on Signal Processing*, 63(16):4422–4437, 2015.

Konstantin Mishchenko. Sinkhorn algorithm as a special case of stochastic mirror descent. *arXiv preprint arXiv:1909.06918*, 2019.

Jan C Neddermeyer. Computationally efficient nonparametric importance sampling. *Journal of the American Statistical Association*, 104(486):788–802, 2009.

Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27:1017–1025, 2014.

Man-Suk Oh and James O. Berger. Adaptive importance sampling in Monte Carlo integration. *J. Statist. Comput. Simulation*, 41(3-4):143–168, 1992.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Art Owen and Yi Zhou. Safe and effective importance sampling. *J. Amer. Statist. Assoc.*, 95(449):135–143, 2000.

Art B Owen. Monte carlo theory, methods and examples. 2013.

Thomas Peel, Sandrine Anthoine, and Liva Ralaivola. Empirical bernstein inequalities for u-statistics. In *Neural Information Processing Systems (NIPS)*, number 23, pages 1903–1911, 2010.

François Portier and Bernard Delyon. Asymptotic optimality of adaptive importance sampling. In *Advances in Neural Information Processing Systems*, 2018.

Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

Sebastian U Stich, Anant Raj, and Martin Jaggi. Safe adaptive importance sampling. *Advances in Neural Information Processing Systems*, 2017.

Charles J. Stone. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4):1040 – 1053, 1982. doi: 10.1214/aos/1176345969. URL `https://doi.org/10.1214/aos/1176345969`.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

Tim van Erven and Peter Harremoes. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, Jul 2014. ISSN 1557-9654. doi: 10.1109/tit.2014.2320500. URL `http://dx.doi.org/10.1109/TIT.2014.2320500`.

Mike West. Approximating posterior distributions by mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2):409–422, 1993.

Ping Zhang. Nonparametric importance sampling. *Journal of the American Statistical Association*, 91 (435):1245–1253, 1996.

Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *international conference on machine learning*, pages 1–9. PMLR, 2015.

Yan Zhou, Adam M Johansen, and John AD Aston. Toward automatic model comparison: an adaptive sequential monte carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726, 2016.

A Proofs of the lemmas

A.1 Proof of Lemma 1

Because q dominates f, we have $\mathbb{E}[W(X)] = 1$. The first inequality is due to Jensen's inequality: $1 = \mathbb{E}[W(X)]^\eta \geq \mathbb{E}[W(X)^\eta]$. When $W(X)$ is not a constant, equality holds if and only if $\eta = 1$.

For the second inequality, write

$$\mathbb{V}\mathrm{ar}[W(X)^\eta] \leq \mathbb{V}\mathrm{ar}[W(X)^\eta] + (\mathbb{E}[W(X)^\eta] - 1)^2 = \mathbb{E}[(W(X)^\eta - 1)^2] \leq \mathbb{E}[(W(X) - 1)^2].$$

The first inequality is obvious. The second inequality holds because $|w^\eta - 1| \leq |w - 1|$ for all $w \geq 0$. □

A.2 Proof of Lemma 2

Recall that the general definition of the Kullback-Leibler divergence is given by

$$\mathrm{KL}\left(f \| q\right) = \int f \log\left(\frac{f}{q}\right) + \int q - \int f \ .$$

Note that it extends the definition given in Lemma 2 to unnormalized densities. Let q be a probability density function and set $\tilde{q} = f^\eta q^{1-\eta}$. Then, using that $\log(u) \leq u - 1$ for all $u > 0$, we have that

$$\begin{aligned}
\mathrm{KL}\left(q \| f\right) \left(f \middle\| \frac{\tilde{q}}{\int \tilde{q}}\right) &= \int f \log\left(\frac{f}{\tilde{q}} \cdot \int \tilde{q}\right) \\
&= \int f \log\left(\frac{f}{\tilde{q}}\right) + \log\left(\int \tilde{q}\right) \\
&\leq \mathrm{KL}\left(f \| \tilde{q}\right) \ . \qquad (9)
\end{aligned}$$

Furthermore, by definition of $\tilde{q}$, it holds that

$$\begin{aligned}
\mathrm{KL}\left(f \| \tilde{q}\right) &= \int \log\left(\frac{f}{f^\eta q^{1-\eta}}\right) f + \int f^\eta q^{1-\eta} - 1 \\
&= (1-\eta) \int \log(f/q) f + \int f^\eta q^{1-\eta} - 1 \\
&= (1-\eta) KL(f \| q) + \int f^\eta q^{1-\eta} - 1 \\
&\leq (1-\eta)\, \mathrm{KL}\left(f \| q\right) \ ,
\end{aligned}$$

where the last inequality results from Jensen's inequality applied to the convex function $u \mapsto u^\eta$:

$$\int f^\eta q^{1-\eta} = \int q \left(\frac{f}{q}\right)^\eta \leq \left(\int f\right)^\eta = 1 \ .$$

Combining with (9) and letting $(q_k^*)_{k \geq 1}$ be defined by (2) starting from an initial probability density function q_1, by recursion we have for all $n \in \mathbb{N}^\star$,

$$\mathrm{KL}\left(f \| q_{n+1}^*\right) \leq \mathrm{KL}\left(f \| q_1\right) \prod_{k=1}^{n} (1 - \eta_k).$$

By applying Pinsker's inequality, we finally obtain

$$\int |f - q_{n+1}^*| \leq \sqrt{2\, \mathrm{KL}\left(f \| q_1\right)} \prod_{k=1}^{n} (1 - \eta_k)^{1/2}.$$

□

Convergence rates obtained from Lemma 2. Assuming that $\mathrm{KL}\left(f\|q_1\right) < +\infty$ and noticing that

$$\log\left(\prod_{k=1}^{n}(1-\eta_k)^{1/2}\right) \leq -\frac{1}{2}\sum_{k=1}^{n}\eta_k,$$

we get the following convergence rates:

- taking $\eta_k = c/k$ with $0 < c < 1$ yields $\int |f - q^*_{n+1}| = O(n^{-c/2})$,
- taking $\eta_k = c/k^\beta$ with $0 < c < 1$ and $\beta \in [0, 1)$ yields $\int |f - q^*_{n+1}| = O(\exp(-Cn^{(1-\beta)})$, with $C = c/(2(1-\beta))$.

B Deriving (2) from an optimisation perspective

One way to approximate an unknown probability density is to formulate an optimisation problem over a certain space of distributions, as it is typically done in *variational inference*. The common choice in variational inference then often corresponds selecting the Kullback-Leibler divergence and to try to find

$$q^\star = \mathrm{arginf}_{q\in\mathcal{Q}} \ \mathrm{KL}\left(q\|f\right) \ ,$$

where $\mathcal{Q}$ is a valid set of probability densities on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $\mathcal{B}(\mathbb{R}^d)$ denoting the Borel σ-field of $\mathbb{R}^d$, and where $\mathrm{KL}\left(f\|q\right)$ stands for the Kullback-Leibler divergence between f and q, i.e $\mathrm{KL}\left(f\|q\right) = \int \log\left(f/q\right) f$.

Following the approach of Dai et al. (2016), one way to solve this optimisation problem is to resort to the *entropic mirror descent* algorithm applied to the objective function $q \mapsto \mathrm{KL}\left(q\|f\right)$. When $\mathcal{Q}$ corresponds to the set of probability density functions on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, this algorithm admits a closed-form solution and generates a sequence $(q_k)_{k\geq 1}$ in $\mathcal{Q}$ satisfying (2).

To see this, let us start with a preliminary result. Let h be a real-valued measurable function defined on $\mathbb{R}^d$ such that $\int \exp(-h)\mathrm{d}\lambda < \infty$. For any probability density function f on $\mathbb{R}^d$ such that $\int |h + \log(f)| f \mathrm{d}\lambda < \infty$, define

$$\Psi(f) = \int (h + \log(f)) f \mathrm{d}\lambda.$$

Lemma 6. The minimum of the function Ψ is attained for $f \propto \exp(-h)$.

Proof. By applying Jensen's inequality to the convex function $u \mapsto \exp(-u)$, we obtain

$$\exp(-\Psi(f)) \leq \int \exp\left(-\left[h + \log(f)\right]\right) f \mathrm{d}\lambda \ .$$

Thus, we have that

$$\Psi(f) \geq -\log\left(\int \exp\left(-h\right) \mathrm{d}\lambda\right) \ ,$$

where the r.h.s does not depend on f and equality is attained whenever $f \propto \exp(-h)$ almost everywhere. □

Next, we rewrite (2) as an Entropic Mirror Descent step. For any $x \in \mathbb{R}^d$ probability density $q \in \mathcal{Q}$, we set $h_q(x) = \log(q(x)/f(x)) + 1$. Given a probability density q^*_k and $\eta_k > 0$, one iteration of the (Infinite-Dimensional) Entropic Mirror Descent algorithm applied to the objective function $q \mapsto \mathrm{KL}\left(q\|f\right)$ with a learning rate η_k corresponds to finding

$$q^*_{k+1} = \mathrm{argmin}_{q\in\mathcal{Q}} \ \eta_k \int h_{q^*_k}(x) q(x) \mathrm{d}x + \mathrm{KL}\left(q\|q^*_k\right) \ .$$

In this expression, which is called the proximal form of the Entropic Mirror Descent, the function $h_{q^*_k}$ plays the role of the gradient of $\mathrm{KL}\left(q\|f\right)$ w.r.t the probability density q^*_k (here, it corresponds to its Fréchet differential).

Based on the previous paragraph, we deduce that if $\mathcal{Q}$ corresponds to the set of probability density functions on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, then

$$q_{k+1}^* = \frac{q_k^*(x)e^{-\eta_k h_{q_k^*}(x)}}{\int q_k^*(x')e^{-\eta_k h_{q_k^*}(x')}\mathrm{d}x'} \propto f^{\eta_k}(x)q_k^*(x)^{1-\eta_k} \ ,$$

that is, we recover (2).

Under minimal assumptions, the convergence towards f can be established with a known convergence rate for an appropriate choice of learning policy $(\eta_k)_{k\geq 1}$.

Lemma 7. Let $(\eta_k)_{k\geq 1}$ be a sequence of positive learning rates and let $\mathcal{Q}$ be set of probability density functions on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Let $q_1 \in \mathcal{Q}$ and let the sequence $(q_k^*)_{k\geq 1}$ be defined by (2). Assume that $x \mapsto h_q(x)$ is bounded by a positive constant L for all $x \in \mathbb{R}^d$ and $q \in \mathcal{Q}$. Then, for all $n \in \mathbb{N}^\star$, we have

$$\mathrm{KL}\left(\sum_{k=1}^n \frac{\eta_k q_k^*}{\sum_{k'=1}^n \eta_{k'}} \middle\| f\right) \leq \frac{\sum_{k=1}^n \eta_k^2 L^2/2}{\sum_{k=1}^n \eta_k} + \frac{\mathrm{KL}\left(f\|q_1\right)}{\sum_{k=1}^n \eta_k} \ .$$

In particular, taking $\eta_k = c_0/\sqrt{k}$ with $c_0 > 0$ yields an $O(\log(n)/\sqrt{n})$ convergence rate. If the total number of iterations n is known in advance, setting $\eta_k = c_0/\sqrt{n}$ for all $k = 1 \dots n$ with $c_0 > 0$ yields an $O(1/\sqrt{n})$ convergence rate.

The proof of this result can be adapted from (Bubeck, 2015, Theorem 4.2). It is provided here for the sake of completeness.

Proof. For all $k \geq 1$, set $\Delta_k = \mathrm{KL}\left(q_k^*\|f\right)$. By convexity of the function $u \mapsto u\log u$, we have

$$\begin{aligned}
\Delta_k &= \int \log\left(\frac{q_k^*(x)}{f(x)}\right) q_k^*(x)\mathrm{d}x \\
&\leq \int \left[\log\left(\frac{q_k^*(x)}{f(x)}\right) + 1\right](q_k^*(x) - f(x))\mathrm{d}x \ , \\
&= \int h_{q_k^*}(x)(q_k^*(x) - f(x))\mathrm{d}x \ .
\end{aligned}$$

Since the integral of any constant w.r.t $q_k^* - f$ is null, we deduce

$$\begin{aligned}
\eta_k\Delta_k &\leq \int \log\left(\frac{q_k^*(x)}{q_{k+1}^*(x)}\right)(q_k^*(x) - f(x))\mathrm{d}x \\
&= \int \log\left(\frac{q_k^*(x)}{q_{k+1}^*(x)}\right) q_k^*(x)\mathrm{d}x - \int \log\left(\frac{q_k^*(x)}{q_{k+1}^*(x)}\right) f(x)\mathrm{d}x \\
&= \int \log\left(\frac{q_k^*(x)}{q_{k+1}^*(x)}\right)(q_k^*(x) - q_{k+1}^*(x))\mathrm{d}x - \mathrm{KL}\left(q_{k+1}^*\|q_k^*\right) \\
&\quad + \left[\mathrm{KL}\left(f\|q_k^*\right) - \mathrm{KL}\left(f\|q_{k+1}^*\right)\right]
\end{aligned}$$

Let us consider the first term of the r.h.s. of the latter inequality. We have that

$$\begin{aligned}
\int \log\left(\frac{q_k^*(x)}{q_{k+1}^*(x)}\right)(q_k^*(x) - q_{k+1}^*(x))\mathrm{d}x &= \eta_k \int h_{q_k^*}(x)(q_k^*(x) - q_{k+1}^*(x))\mathrm{d}x \\
&\leq \eta_k L \int |q_k^* - q_{k+1}^*| \ .
\end{aligned}$$

since by assumption $h_{q_k^*}$ is bounded by L. Additionally, we have by Pinsker's inequality that

$$-\mathrm{KL}\left(q_{k+1}^*\|q_k^*\right) \leq -\frac{1}{2}\left(\int |q_k^* - q_{k+1}^*|\right)^2 \ .$$

Now combining with the fact that $\eta_k L a - a^2/2 \leq (\eta_k L)^2/2$ for all $a \geq 0$, we get:

$$\int \log\left(\frac{q_k^*(x)}{q_{k+1}^*(x)}\right)(q_k^*(x) - q_{k+1}^*(x))\mathrm{d}x - \mathrm{KL}\left(q_{k+1}^* \| q_k^*\right) \leq \frac{(\eta_k L)^2}{2}$$

and as a consequence we deduce

$$\eta_k \Delta_k \leq \frac{(\eta_k L)^2}{2} + \left[\mathrm{KL}\left(f\|q_k^*\right) - \mathrm{KL}\left(f\|q_{k+1}^*\right)\right] .$$

Finally, as we recognize a telescoping sum in the right-hand side, we have

$$\sum_{k=1}^n \eta_k \Delta_k \leq \sum_{k=1}^n \eta_k^2 L^2/2 + \mathrm{KL}\left(f\|q_1\right)$$

that is, by convexity of the mapping $q \mapsto \mathrm{KL}\left(q\|f\right)$,

$$\mathrm{KL}\left(\sum_{k=1}^n \frac{\eta_k q_k^*}{\sum_{k'=1}^n \eta_{k'}} \middle\| f\right) - \mathrm{KL}\left(f\|f\right) \leq \frac{\sum_{k=1}^n \eta_k^2 L^2/2}{\sum_{k=1}^n \eta_k} + \frac{\mathrm{KL}\left(f\|q_1\right)}{\sum_{k=1}^n \eta_k}$$

Then, notice that taking $\eta_k = \eta_0/\sqrt{k}$ with $\eta_0 > 0$ yields an $O(\log(n)/\sqrt{n})$ convergence rate and that setting $\eta_k = \eta_0/\sqrt{n}$ for all $k = 1 \dots n$ yields an $O(1/\sqrt{n})$ convergence rate. □

C Proof of Proposition 3

The proof is organized in three parts. First we provide high-level results related to Freedman's inequality for martingales. Then we provide some intermediary technical results, and finally we conclude with the proof of Proposition 3.

C.1 Bernstein inequalities for martingale processes

The two following propositions can be found in Delyon and Portier (2021).

Proposition 8. Let $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k\geq 1}, \mathbb{P})$ be a filtered space. Let $(Y_k)_{1\leq k\leq n}$ be real valued random variables such that

$$\mathbb{E}[Y_k|\mathcal{F}_{k-1}] = 0, \quad \text{for all } k = 1, \dots, n .$$

Then, for all $t \geq 0$ and all $v, m > 0$,

$$\mathbb{P}\left(\Big|\sum_{k=1}^n Y_k\Big| \geq t, \max_{k=1,\dots,n} |Y_k| \leq m, \sum_{k=1}^n \mathbb{E}[Y_k^2|\mathcal{F}_{k-1}] \leq v\right) \leq 2\exp\left(-\frac{t^2}{2(v + tm/3)}\right) .$$

Proposition 9. Let $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k\geq 1}, \mathbb{P})$ be a filtered space. Let $(Y_k)_{k\geq 1}$ be a sequence of real valued stochastic processes defined on $\mathbb{R}^d$, adapted to $(\mathcal{F}_k)_{k\geq 1}$, such that for any $x \in \mathbb{R}^d$,

$$\mathbb{E}[Y_k(x)|\mathcal{F}_{k-1}] = 0, \quad \text{for all } k \geq 1 .$$

Consider $\epsilon > 0$ and let $(\tilde{Y}_k)_{k\geq 1}$ be another $(\mathcal{F}_k)_{k\geq 1}$-adapted sequence of nonnegative stochastic processes defined on $\mathbb{R}^d$ such that for all $k \geq 1$ and $x \in \mathbb{R}^d$

$$\sup_{\|y\|\leq\epsilon} |Y_k(x+y) - Y_k(x)| \leq \tilde{Y}_k(x) .$$

Let $n \geq 1$ and assume that for some $A \geq 0$ and some set $\Omega_1 \subset \Omega$, there exist $m, v, \tau \in \mathbb{R}^+$ such that for all $\omega \in \Omega_1$ and $\|x\| \leq A$,

$$\max_{k=1,\dots,n} |Y_k(x)| \leq m \tag{10}$$

$$\sum_{k=1}^n \mathbb{E}\big[Y_k(x)^2|\mathcal{F}_{k-1}\big] \leq v \tag{11}$$

$$\sum_{k=1}^n \mathbb{E}\big[\tilde{Y}_k(x)|\mathcal{F}_{k-1}\big] \leq \tau . \tag{12}$$

Then, for all $t \geq 0$,

$$\mathbb{P}\Big(\sup_{\|x\| \leq A} \big| \sum_{k=1}^{n} Y_k(x) \big| > t + \tau, \Omega_1 \Big) \leq 4(1 + 2A/\varepsilon)^d \exp\Big(-\frac{t^2}{8(\tilde{v} + 2mt/3)} \Big) \ ,$$

with $\tilde{v} = \max(v, 2m\tau)$.

Recall also that the *predictable quadratic variation* Bercu et al. (2015) of a martingale $\sum_{k=1}^{n} \beta_k$ is given by

$$\sum_{k=1}^{n} \mathbb{E}[\beta_k^2 | \mathcal{F}_{k-1}].$$

This quantity is important as it appears as an essential factor in the Bernstein inequalities above. In particular, the predictable quadratic variation of the function M_n defined in (7) can be found in Proposition 10 below.

Proposition 10. Suppose that the kernel function $K : \mathbb{R}^d \to \mathbb{R}^+$ is bounded. For $n \geq 1$, let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ be the σ-algebra generated by the random variables $X_1, \ldots, X_n$, and $\mathcal{F}_0 = \emptyset$. Then, for each $x \in \mathbb{R}^d$, $(nM_n(x))_{n \geq 1}$ is a $(\mathcal{F}_n)_{n \geq 1}$-martingale with predictable quadratic variation $\sum_{k=1}^{n} V_{\eta_k, h_n}(x)$ where $V_{\eta,h}(x) = \int K_h(x-y)^2 f^{2\eta}(y) q^{1-2\eta}(y) dy - f \star K_h(x)^2$.

Proof. Because q_k is positive on $\mathbb{R}^d$ and $\mathcal{F}_k$-measurable,

$$\mathbb{E}[W_k^{\eta_k} K_{h_n}(x - X_k) \mid \mathcal{F}_{k-1}] = \int f^{\eta_k}(y) q_{k-1}^{1-\eta_k}(y) K_{h_n}(x-y) = (f^{\eta_k} q_{k-1}^{1-\eta_k}) \star K_{h_n}(x).$$

The formula for the predictable quadratic variation follows in the same way. □

C.2 Technical results

We start with some notation.

Notation. Recall from Section 3.2 that for all $n \geq 1$ and all $x \in \mathbb{R}^d$,

$$\begin{aligned} f_n(x) &= \frac{N_n(x)}{D_n} \\ N_n(x) &= n^{-1} \sum_{k=1}^{n} W_k^{\eta_k} K_{h_n}(x - X_k) \\ D_n &= n^{-1} \sum_{k=1}^{n} W_k^{\eta_k} \ . \end{aligned}$$

In addition, for all $k \geq 1$ and all $x \in \mathbb{R}^d$, we introduce

$$\begin{aligned} &\tilde{f}_k(x) = f^{\eta_k}(x) q_{k-1}^{1-\eta_k}(x), \\ &Z_k^{(1)}(x) = W_k^{\eta_k} K_{h_k}(x - X_k) - (\tilde{f}_k \star K_{h_k})(x), \\ &Z_k^{(2)} = W_k^{\eta_k} - \int \tilde{f}_k. \end{aligned}$$

The three following propositions focus on convergence results for $Z_k^{(1)}$ and $Z_k^{(2)}$.

Proposition 11. Under $(\mathbf{A}_1)$ and $(\mathbf{A}_2)$ we have that almost-surely,

$$\forall k = 1, \ldots, n, \qquad 0 \leq W_k^{\eta_k} \leq (c\lambda_n)^{-1} \ .$$

Proof. Note that by definition of q_k in (3) combined with $(\mathbf{A}_2)$, we have that for all $k = 1, \ldots, n$ and all $x \in \mathbb{R}^d$,

$$q_k(x) \geq \lambda_k q_0(x) \geq \lambda_k c f(x) \ . \tag{13}$$

As a consequence, $0 \leq W_k \leq 1/(c\lambda_{k-1}) \leq 1/(c\lambda_n)$ using that $(\lambda_k)_{k \geq 1}$ is nonincreasing under $(\mathbf{A}_1)$. Note also that (13) implies that $\lambda_n \in (0, 1/c]$. Thus, we can write, for all $k = 1, \ldots, n$, $(c\lambda_n)^{-\eta_k} \leq (c\lambda_n)^{-1}$, which yields the stated result. □

Proposition 12. Under $(\mathbf{A}_1)$ and $(\mathbf{A}_2)$ we have that almost-surely

$$n^{-1}\sum_{k=1}^{n} Z_k^{(2)} = O\left(\sqrt{\frac{\log(n)}{n\lambda_n}}\right).$$

Proof. We first show that without loss of generality, we can derive the proof assuming that the sequence $(\eta_k)_{k\geq 1}$ is valued in $(1/2, 1]$. From Proposition 11, it is easy to see that for all $k = 1, \ldots, n$,

$$|Z_k^{(2)}| = |W_k^{\eta_k} - \mathbb{E}[W_k^{\eta_k}|\mathcal{F}_{k-1}]| \leq (c\lambda_n)^{-1}\,. \tag{14}$$

Since the sequence $(\eta_k)_{k\geq 1}$ goes to 1 under Assumption $(\mathbf{A}_1)$(iii), there is an integer $k_0 \geq 1$ such that $\eta_k \in (1/2, 1]$ for all $k > k_0$. Hence, whenever $n > k_0$, we have

$$\sum_{k=1}^{n} Z_k^{(2)} = \sum_{k=1}^{k_0} Z_k^{(2)} + \sum_{k=k_0+1}^{n} Z_k^{(2)}.$$

By (14), the first term is bounded by $k_0/(c\lambda_{k_0})$ and hence has a negligible contribution in the bound we need to establish. The only term that matters is then the second one. Hence, from now on we assume that $(\eta_k)_{k\geq 1}$ is valued in $(1/2, 1]$. The goal will be to apply Proposition 8 to $Y_k(x) = Z_k^{(2)}(x)$. Using that $0 < 2\eta_k - 1 \leq 1$ and Proposition 11, we can write

$$\begin{aligned}
\mathbb{E}[Z_k^{(2)2}|\mathcal{F}_{k-1}] &\leq \mathbb{E}[f(X_k)/q_{k-1}(X_k))^{2\eta_k}|\mathcal{F}_{k-1}] \\
&= \int f(x)^{2\eta_k} q_{k-1}(x)^{1-2\eta_k}\,\mathrm{d}x \\
&= \int (f(x)/q_{k-1}(x))^{2\eta_k - 1} f(x)\,\mathrm{d}x \\
&\leq (1/(c\lambda_{k-1}))^{2\eta_k - 1} \\
&\leq (1/(c\lambda_n))^{2\eta_k - 1} \\
&\leq (c\lambda_n)^{-1}\,.
\end{aligned}$$

It follows that

$$\sum_{k=1}^{n} \mathbb{E}[Z_k^{(2)2}|\mathcal{F}_{k-1}] \leq n(c\lambda_n)^{-1}.$$

Consequently, and using (14), we can apply Proposition 8 with $m = (c\lambda_n)^{-1}$, $v = n(c\lambda_n)^{-1}$ and we get that for all n large enough such that $\log(n)/(n\lambda_n) \leq c$; this is made possible by $(\mathbf{A}_1)$; and all $\gamma \geq 9$,

$$\begin{aligned}
\mathbb{P}\left(\Big|\sum_{k=1}^{n} Z_k^{(2)}\Big| \geq \sqrt{\gamma(c\lambda_n)^{-1} n\log(n)}\right) &\leq 2\exp\left(-\frac{\gamma n\log(n)}{2(n + \sqrt{\gamma(c\lambda_n)^{-1}n\log(n)}/3)}\right) \\
&\leq 2\exp\left(-\frac{\gamma}{2(1+\sqrt{\gamma}/3)}\log(n)\right) \\
&= 2\exp\left(-\frac{\sqrt{\gamma}}{2}\frac{\sqrt{\gamma}}{1+\sqrt{\gamma}/3}\log(n)\right) \\
&\leq 2\exp\left(-\frac{9}{4}\log(n)\right) \\
&\leq 2\exp\left(-2\log(n)\right) = 2n^{-2},
\end{aligned}$$

which series is convergent. We obtain the desired result by invoking the Borel-Cantelli lemma.

□

Proposition 13. Under $(\mathbf{A}_1)$, $(\mathbf{A}_2)$, $(\mathbf{A}_3)$ and $(\mathbf{A}_4)$, we have that for any $r > 0$,

$$\sup_{\|x\| \leq n^r} \left| n^{-1} \sum_{k=1}^{n} Z_k^{(1)}(x) \right| = O\left(\sqrt{\frac{\log(n)}{n h_n^d \lambda_n}} \right) .$$

Proof. Using similar arguments as in the beginning of the proof of Proposition 12, we can assume that $(\eta_k)_{k \geq 1}$ is valued in $(1/2, 1]$. The proof consists in applying Proposition 9 to $Y_k(x) = Z_k^{(1)}(x)$ that is

$$Y_k(x) = W_k^{\eta_k} K_{h_n}(x - X_k) - \mathbb{E}[W_k^{\eta_k} K_{h_n}(x - X_k) | \mathcal{F}_{k-1}] .$$

In the next few lines, we derive the quantities m, v, τ that appear in Proposition 9. Under Assumption $(\mathbf{A}_4)$ combined with Proposition 11, we have that

$$|Y_k(x)| \leq \frac{K_\infty}{c \lambda_n h_n^d} .$$

The previous bound corresponds to m in Proposition 9. Moreover, using Proposition 11,

$$\begin{aligned}
\mathbb{E}[Y_k(x)^2 | \mathcal{F}_{k-1}] &\leq \mathbb{E}[W_k^{2\eta_k} K_{h_n}(x - X_k)^2 | \mathcal{F}_{k-1}] \\
&= h_n^{-2d} \int \left(\frac{f(y)}{q_{k-1}(y)} \right)^{2\eta_k - 1} f(y) K((x-y)/h_n)^2 \mathrm{d}y \\
&\leq h_n^{-2d} (1/(c\lambda_{k-1}))^{2\eta_k - 1} \int f(y) K((x-y)/h_n)^2 \mathrm{d}y \\
&= h_n^{-d} (1/(c\lambda_n))^{2\eta_k - 1} \int f(x - h_n u) K(u)^2 \mathrm{d}u \\
&\leq h_n^{-d} (c\lambda_n)^{-1} U K_\infty
\end{aligned}$$

where we used a variable change $u = (x - y)/h_n$ in the penultimate inequality, and the last inequality results from Assumption $(\mathbf{A}_4)$, $0 < 2\eta_k - 1 \leq 1$, and Assumption $(\mathbf{A}_3)$.

Hence, we get

$$\sum_{k=1}^{n} \mathbb{E}[Y_k(x)^2 | \mathcal{F}_{k-1}] \leq n h_n^{-d} (c\lambda_n)^{-1} U K_\infty .$$

The previous bound corresponds to v in Proposition 9. Under $(\mathbf{A}_4)$, $|K_h(x+y-X_k) - K_h(x-X_k)| \leq L_K \|y/h\| h^{-d}$ and it holds that for all $\|y\| \leq \epsilon$,

$$|Y_k(x+y) - Y_k(x)| \leq (W_k^{\eta_k} + \mathbb{E}[W_k^{\eta_k} | \mathcal{F}_{k-1}]) L_K \epsilon h_n^{-d-1}.$$

The l.h.s. of the previous inequality corresponds to $\tilde{Y}_k(x)$ in Proposition 9. We have

$$\sum_{k=1}^{n} \mathbb{E}[\tilde{Y}_k(x) | \mathcal{F}_{k-1}] \leq 2 n L_K \epsilon h_n^{-d-1}.$$

where we have used that $\mathbb{E}[W_k^{\eta_k} | \mathcal{F}_{k-1}] \leq 1$.

Taking $\epsilon = h_n^{d+1}/n$, the value for τ in Proposition 9 is $2L_K$. Let us now summarize the different factors taken to apply Proposition 9:

$$\begin{aligned}
m &= \frac{K_\infty}{c \lambda_n h_n^d} \\
v &= n h_n^{-d} (c\lambda_n)^{-1} U K_\infty \\
\tau &= 2 L_K \\
\tilde{v} &= \max(v, 2m\tau) \leq C \max(n/(\lambda_n h_n^d), \lambda_n^{-1} h_n^{-d}) = C n/(\lambda_n h_n^d)
\end{aligned}$$

where C is a positive constant. Let $\gamma > 1$. We have, taking $t = \sqrt{\gamma n \log(n)/(h_n^d \lambda_n)}$, $A = n^r$ and $\Omega_1 = \Omega$, for n large enough ($t \geq \tau$ and $\tilde{v}\sqrt{\gamma} \geq 2mt/3$; this is made possible by $(\mathbf{A}_1)$),

$$\begin{aligned}
\mathbb{P}\Big(\sup_{\|x\| \leq n^r} \big| \sum_{k=1}^n Y_k(x) \big| > 2t \Big) &\leq \mathbb{P}\Big(\sup_{\|x\| \leq n^r} \big| \sum_{k=1}^n Y_k(x) \big| > t + \tau \Big) \\
&\leq 4(1 + 2n^{r+1}/h_n^{d+1})^d \exp\left(-\frac{t^2}{8(1+\sqrt{\gamma})\tilde{v}} \right) \\
&\leq 4(1 + 2n^{r+1}/h_n^{d+1})^d \exp\left(-\frac{\gamma \log(n)}{16C} \right)
\end{aligned}$$

The last inequality holds because $\gamma > 1$. It remains to choose γ large enough in order to ensure the summability condition in the Borel Cantelli lemma. □

C.3 End of the proof of Proposition 3

Since $f_n(x) = N_n(x)/D_n$ for all $n \geq 1$ and all $x \in \mathbb{R}^d$, it is enough to show that $|D_n - 1| = o(1)$, and $\sup_{\|x\| \leq n^r} |N_n(x) - f(x)| = o(1)$. Both results are obtained independently, starting with D_n.

Proof for D_n. First note that for all $n \geq 1$, we can write the following decomposition for D_n

$$D_n = n^{-1} \sum_{k=1}^n \int \tilde{f}_k + n^{-1} \sum_{k=1}^n Z_k^{(2)} \ .$$

Furthermore, under $(\mathbf{A}_1)$ and $(\mathbf{A}_2)$, Proposition 12 implies that the second term of the r.h.s. is $O(\sqrt{\log(n)/(n\lambda_n)})$ almost surely. Hence, by $(\mathbf{A}_1)$ the sequence $(D_n)_{n\geq 1}$ converges to 1 as soon as $(n^{-1}\sum_{k=1}^n \int \tilde{f}_k)_{n\geq 1}$ does. By the Cesaro lemma, this will be a consequence of having proven that $(\int \tilde{f}_k)_{k\geq 1}$ goes to 1, which is what we set out to do next. For all $k \geq 1$, Jensen's inequality yields $\int \tilde{f}_k \leq 1$ and setting $\lambda_0 = 1$, we deduce using $(\mathbf{A}_2)$ that

$$\int \tilde{f}_k = \int f^{\eta_k} q_{k-1}^{1-\eta_k} \geq \lambda_{k-1}^{1-\eta_k} \int f^{\eta_k} q_0^{1-\eta_k} \geq (\lambda_{k-1} c)^{1-\eta_k} \ .$$

Thus, $(\int \tilde{f}_k)_{k\geq 1}$ goes to 1 under $(\mathbf{A}_1)$ and we can conclude that $(D_n)_{n\geq 1}$ converges to 1.

Proof for N_n. For the numerator N_n, we follow a similar approach except that we need to deal with some convolution operator. For all $n \geq 1$ and all $x \in \mathbb{R}^d$, we can write

$$N_n(x) = n^{-1} \sum_{k=1}^n (\tilde{f}_k \star K_{h_n})(x) + n^{-1} \sum_{k=1}^n Z_k^{(1)},$$

where given $r > 0$, the second term of the r.h.s is $O(\sqrt{\log(n)/(nh_n^d\lambda_n)})$ as a consequence of Proposition 13.

To treat the first term, use $(\mathbf{A}_2)$ and $(\mathbf{A}_4)$, to obtain that for all $x \in \mathbb{R}^d$,

$$\lambda_k c f(x) \leq q_k(x) \leq h_k^{-d} K_\infty + U_{q_0} := h_k^{-d} C,$$

for some $C > 0$. It follows that

$$\left(n^{-1} \sum_{k=1}^n f_k^- \right) \star K_{h_n}(x) \leq n^{-1} \sum_{k=1}^n (\tilde{f}_k \star K_{h_n})(x) \leq \left(n^{-1} \sum_{k=1}^n f_k^+ \right) \star K_{h_n}(x)$$

with $f_k^-(x) = f(x)(c\lambda_k)^{1-\eta_k}$ and $f_k^+(x) = f(x)^{\eta_k}(h_k^{-d}C)^{1-\eta_k}$.

It remains to show that the previous lower and upper bounds converge to f uniformly. For $1 \leq p \leq +\infty$, let $\|\cdot\|_p$ denote the $L_p(\lambda)$-norm. Using that $\|g \star \tilde{g}\|_\infty \leq \|g\|_\infty \|\tilde{g}\|_1$, we find, for all collection of bounded functions

$(g_1, \dots, g_n)$,

$$
\begin{aligned}
|\left(n^{-1}\sum_{k=1}^{n} g_k\right) \star K_{h_n}(x) - f(x)| &\leq |\left(n^{-1}\sum_{k=1}^{n}(g_k - f)\right) \star K_{h_n}(x)| + |f \star K_{h_n}(x) - f(x)| \\
&\leq \|n^{-1}\sum_{k=1}^{n}(g_k - f)\|_\infty \|K_{h_n}\|_1 + \|f \star K_{h_n} - f\|_\infty \\
&\leq n^{-1}\sum_{k=1}^{n}\|g_k - f\|_\infty + \|f \star K_{h_n} - f\|_\infty
\end{aligned}
$$

Hence, in virtue of the Cesaro lemma, the fact that f_k^- and f_k^+ both converge uniformly to f enables to conclude that the first term in the latter upper bound goes to 0. The fact that $\|f \star K_{h_n} - f\|_\infty$ goes to 0 is an easy consequence of $(\mathbf{A}_4)$.

□

Remark 5. Notice that in the latter proof, a different bandwith h_k for each point X_k, $k = 1, \dots, n$ could have been set (instead of $h_k = h_n$ for all k). Indeed, in the latter inequalities, the term $\|f \star K_{h_n} - f\|_\infty$ would be replaced by $1/n \sum_{k=1}^{n} \|f \star K_{h_k} - f\|_\infty$, which also goes to 0 by the Cesaro Lemma as soon as $\|f \star K_{h_n} - f\|_\infty$ goes to 0 .

D Proof of Proposition 4

For the sake of completeness, we first recall the following basic lemma (whose proof can be found in (Peel et al., 2010, Lemma 1)).

Lemma 14. Let X be a random variable and let $a, b > 0$, $c, d \geq 0$ be such that

$$
\forall t > 0 \ , \quad \mathbb{P}\left(X \geq t\right) \leq a \exp\left(-\frac{bt^2}{c + dt}\right) \ .
$$

Then, with probability at least $1 - \delta$,

$$
|X| \leq \sqrt{\frac{c}{b}\ln\frac{a}{\delta}} + \frac{d}{b}\ln\frac{a}{\delta} \ .
$$

The proof of Proposition 4 is an easy consequence of the following Lemma.

Lemma 15. Under $(\mathbf{A}_1)$, $(\mathbf{A}_2)$, $(\mathbf{A}_5)$ and $(\mathbf{A}_6)$, there exists $s_0 \in \mathbb{N}$ large enough such that

$$
\sup_{\|x\| > n^{s_0}} f_n(x) = o(1) \ , \ a.s \ , \tag{15}
$$

$$
\sup_{\|x\| > n^{s_0}} f(x) = o(1) \ . \tag{16}
$$

Proof. We start with (15). Let $n \geq 1$ and set $A = n^{s_0}/2$ with $s_0 \in \mathbb{N}$. For all $x \in \mathbb{R}^d$, we have the following decomposition

$$
f_n(x) = \sum_{k=1}^{n} W_{n,k}^{(\eta_k)} K_{h_n}(x - X_k)\mathbb{I}_{\{\|X_k\| \leq A\}} + \sum_{k=1}^{n} W_{n,k}^{(\eta_k)} K_{h_n}(x - X_k)\mathbb{I}_{\{\|X_k\| > A\}} \ . \tag{17}
$$

Our goal is to prove that for s_0 large enough, both terms on the r.h.s of (17) go to 0. We start by studying the first term of the r.h.s.

(i) Proof for the first term of the r.h.s of (17). For any $\|x\| > n^{s_0}$, we can write

$$\begin{aligned}\sum_{k=1}^{n} W_{n,k}^{(\eta_k)} K_{h_n}(x - X_k)\mathbb{I}_{\{\|X_k\|\leq A\}} &\leq \sup_{\|y\|\leq A} K_{h_n}(x-y) \\ &\leq C_K h_n^{-d} \sup_{\|y\|\leq A,\, \|x\|>n^{s_0}} (1 + \|x-y\|/h_n)^{-r_K} \\ &\leq C_K h_n^{-d} \sup_{\|y\|\leq A,\, \|x\|>n^{s_0}} (1 + \|x-y\|/h_1)^{-r_K} \\ &\leq C_K h_n^{-d} (1 + n^{s_0}/(2h_1))^{-r_K},\end{aligned}$$

where the last inequality follows from the fact that for all $x, y \in \mathbb{R}^d$, $|\|x\| - \|y\|| \leq \|x - y\|$. We can then ensure that the previous term goes to 0 by letting s_0 be large enough.

(ii) Proof for the second term of the r.h.s of (17). For the second term of the r.h.s, we have

$$\begin{aligned}\sum_{k=1}^{n} W_{n,k}^{(\eta_k)} K_{h_n}(x - X_k)\mathbb{I}_{\{\|X_k\|>A\}} &= \left(\sum_{k=1}^{n} W_k^{\eta_k}\right)^{-1} \sum_{k=1}^{n} W_k^{\eta_k} K_{h_n}(x - X_k)\mathbb{I}_{\{\|X_k\|>A\}} \\ &\leq \left(\sum_{k=1}^{n} W_k^{\eta_k}\right)^{-1} K_\infty h_n^{-d} \sum_{k=1}^{n} W_k^{\eta_k}\mathbb{I}_{\{\|X_k\|>A\}}\,, \end{aligned} \tag{18}$$

and we are thus interested in studying the r.h.s of (18). A first remark is that using Proposition 12, we obtain that almost surely

$$\left(\sum_{k=1}^{n} W_k^{\eta_k}\right)^{-1} = n^{-1}(1 + o(1))^{-1}\,. \tag{19}$$

We now move on to the study of $\sum_{k=1}^{n} W_k^{\eta_k}\mathbb{I}_{\{\|X_k\|>A\}}$ in (18). To do so, for all $k \geq 1$, let us define $p_k(A) = \mathbb{E}[W_k^{\eta_k}\mathbb{I}_{\{\|X_k\|>A\}}|\mathcal{F}_{k-1}]$ and $Z_k^{(3)}(A) = W_k^{\eta_k}\mathbb{I}_{\{\|X_k\|>A\}} - p_k(A)$ so that

$$\sum_{k=1}^{n} W_k^{\eta_k}\mathbb{I}_{\{\|X_k\|>A\}} = \sum_{k=1}^{n} Z_k^{(3)}(A) + \sum_{k=1}^{n} p_k(A)\,.$$

Then, for all $k \geq 1$, using Jensen inequality, it holds that

$$\begin{aligned}p_k(A) &= \int_{\|x\|>A} f(x)^{\eta_k} q_{k-1}(x)^{1-\eta_k}\,\mathrm{d}x \\ &= \int q_{k-1}(f(x)\mathbb{I}_{\|x\|>A}/q_{k-1}(x))^{\eta_k}\,\mathrm{d}x \\ &\leq \left(\int f(x)\mathbb{I}_{\|x\|>A}\,\mathrm{d}x\right)^{\eta_k} \\ &= p(A)^{\eta_k}\,,\end{aligned}$$

with $p(A) := \int_{\|x\|>A} f(x)\,\mathrm{d}x$. Using $(\mathbf{A}_1)$, we deduce that there exists a constant $\eta_0 = \inf_{n\geq 1}\eta_n \in (0, 1]$ such that

$$\sum_{k=1}^{n} p_k(A) \leq n p(A)^{\eta_0}\,.$$

Furthermore, under $(\mathbf{A}_5)$, Markov's inequality yields

$$p(A) \leq A^{-\delta}\int \|x\|^\delta f(x)\,\mathrm{d}x \tag{20}$$

and as a consequence, we obtain

$$\sum_{k=1}^{n} p_k(A) \leq n A^{-\delta\eta_0}\left(\int \|x\|^\delta f(x)\,\mathrm{d}x\right)^{\eta_0}\,. \tag{21}$$

Additionally, observe that $\sum_{k=1}^n Z_k^{(3)}(A)$ is a sum of martingale increments so our next step will be to apply Proposition 8. For this purpose, note that under $(\mathbf{A}_2)$ and $(\mathbf{A}_5)$, we can write for all $k = 1, \dots, n$,

$$W_k \leq \lambda_{k-1}^{-1} C_0 (1 + \|X_k\|^\delta)^{-1} \leq \lambda_n^{-1} C_0 (1 + \|X_k\|^\delta)^{-1}$$

for some constant $C_0 \geq 1$. Then it holds that

$$|Z_k^{(3)}(A)| \leq \lambda_n^{-\eta_k} C_0^{\eta_k} \sup_{\|x\|>A} (1 + \|x\|^\delta)^{-\eta_k} \leq \lambda_n^{-1} C_0 (1 + A^\delta)^{-\eta_0} \leq C_0 \lambda_n^{-1} A^{-\delta\eta_0} \, . \tag{22}$$

We now treat the two case $k \geq k_0$ and $k < k_0$ separately.

- When $k \geq k_0$ (such that $0 < 2\eta_k - 1 \leq 1$), we can write

$$\begin{aligned}
\mathbb{E}[Z_k^{(3)}(A)^2 | \mathcal{F}_{k-1}] &\leq \mathbb{E}[(f(X_k)/q_{k-1}(X_k))^{2\eta_k} \mathbb{I}_{\{\|X_k\|>A\}} | \mathcal{F}_{k-1}] \\
&= \int_{\|x\|>A} f(x)^{2\eta_k} q_{k-1}(x)^{1-2\eta_k} \, \mathrm{d}x \\
&= \int_{\|x\|>A} (f(x)/q_{k-1}(x))^{2\eta_k - 1} f(x) \, \mathrm{d}x \\
&\leq (1/(c\lambda_{k-1}))^{2\eta_k - 1} p(A) \\
&\leq (c\lambda_n)^{-1} p(A) \\
&\leq (c\lambda_n)^{-1} A^{-\delta} \int \|x\|^\delta f(x) \, \mathrm{d}x \, .
\end{aligned}$$

  where we have used (20) in the last inequality.

- When $k < k_0$, we have $\mathbb{E}[Z_k^{(3)}(A)^2 | \mathcal{F}_{k-1}] \leq M A^{-\eta_0 \delta}$ for some constant $M > 0$ that can be deduced from the almost sure bound (22) given just before.

It follows that, when n is large enough, for all $k = 1, \dots, n$,

$$\mathbb{E}[Z_k^{(3)}(A)^2 | \mathcal{F}_{k-1}] \leq (c\lambda_n)^{-1} A^{-\eta_0\delta} \int \|x\|^\delta f(x) \, \mathrm{d}x \, .$$

and therefore,

$$\sum_{k=1}^n \mathbb{E}[Z_k^{(3)}(A)^2 | \mathcal{F}_{k-1}] \leq n(c\lambda_n)^{-1} A^{-\eta_0\delta} \int \|x\|^\delta f(x) \, \mathrm{d}x \, .$$

Consequently, we can apply Proposition 8 with $m = C_0 \lambda_n^{-1} A^{-\eta_0\delta}$, $v = n(c\lambda_n)^{-1} A^{-\eta_0\delta} \int \|x\|^\delta f(x) \, \mathrm{d}x$ and we obtain that for all $t > 0$

$$\mathbb{P}\left(\Big| \sum_{k=1}^n Z_k^{(3)}(A) \Big| \geq t \right) \leq 2 \exp\left(-\frac{t^2}{2(v + tm/3)} \right) \, .$$

Inverting this inequality using Lemma 14, we get that, with probability $1 - 1/n^2$,

$$\begin{aligned}
\left| \sum_{k=1}^n Z_k^{(3)}(A) \right| &\leq \sqrt{4v \log(2n)} + (2m/3) \log(2n) \\
&= \sqrt{4n(c\lambda_n)^{-1} A^{-\eta_0\delta} \left(\int \|x\|^\delta f(x) \, \mathrm{d}x \right) \log(2n)} + (2/3) C_0 \lambda_n^{-1} A^{-\eta_0\delta} \log(2n).
\end{aligned} \tag{23}$$

Invoking the Borel-Cantelli lemma we obtain that the previous bound is an almost sure rate.

Putting together (19), (21) and (23) in (18), we obtain the almost-sure bound

$$\left|\sum_{k=1}^{n} W_{n,k}^{(\eta_k)} K_{h_n}(x - X_k)\mathbb{I}_{\{\|X_k\|>A\}}\right| \\ = O\left(n^{-1}h_n^{-d}\left(\sqrt{nA^{-\eta_0\delta}\lambda_n^{-1}\log(2n)} + (A^{-\eta_0\delta}\lambda_n)^{-1}\log(2n) + nA^{-\eta_0\delta}\right)\right) .$$

We easily obtain that the previous bound goes to 0 provided that s_0 is large enough, which concludes the proof of (15).

As for (16), notice that by $(\mathbf{A}_5)$, for all $\|x\| > n^{s_0}$

$$f(x) \leq \frac{c^{-1}q_0(x)}{1+\|x\|^\delta} \leq \frac{c^{-1}q_0(x)}{1+n^{s_0\delta}} \leq \frac{c^{-1}\|q_0\|_\infty}{1+n^{s_0\delta}}.$$

The previous upper bound does not depend on x and goes to 0 as $n \to \infty$.

□

E Proof of Proposition 5

Proof. We start by proving (i) and (ii).

Proof of (i) and (ii). First note that for our choice of $\mathbb{P}$ and $\mathbb{Q}$, we can write

$$D_1(\mathbb{P}||\mathbb{Q}) = \sum_{\ell=1}^{m_k} W_{k,\ell}\log\left(\frac{W_{k,\ell}}{1/m_k}\right) = \sum_{\ell=1}^{m_k} W_{k,\ell}\log\left(W_{k,\ell}\right) + \log(m_k) \\ \leq \sum_{\ell=1}^{m_k} W_{k,\ell}\left(W_{k,\ell}-1\right) + \log(m_k)$$

where we have used that $\log(x) \leq x-1$ for $x>0$. Thus, we have that

$$D_1(\mathbb{P}||\mathbb{Q}) \leq \log(m_k) . \tag{24}$$

In addition, (van Erven and Harremoes, 2014, Theorem 3) implies that for all $\alpha \in [0,1]$

$$0 \leq D_\alpha(\mathbb{P}||\mathbb{Q}) \leq D_1(\mathbb{P}||\mathbb{Q})$$

where equality is reached if and only if $\mathbb{P}=\mathbb{Q}$. Now combining with (24) and by definition of $\eta_{k,\alpha}$ in (8), we deduce that for all $\alpha \in [0,1]$,

$$0 \leq \eta_{k,1} \leq \eta_{k,\alpha} \leq 1 ,$$

and that $\eta_{k,\alpha}=1$ if and only if $\mathbb{P}=\mathbb{Q}$.

Proof of (iii). A first remark is that thanks to (ii), it is enough to prove that $\lim_{k\to\infty}\eta_{k,1}=1$ in L_1 to obtain that for all $\alpha\in[0,1]$, $\lim_{k\to\infty}\eta_{k,\alpha}=1$ in L_1, that is $\lim_{k\to\infty}\mathbb{E}[|\eta_{k,\alpha}-1|]=0$.

Since that for all $k\geq 1$,

$$\eta_{k,1} = 1 - \frac{D_1(\mathbb{P}||\mathbb{Q})}{\log(m)} = -\frac{\sum_{\ell=1}^m W_{k,\ell}\log\left(W_{k,\ell}\right)}{\log(m)}$$

the proof is concluded if we can prove that, in L_1

$$\lim_{k\to\infty}\sum_{\ell=1}^{m} W_{k,\ell}\log\left(W_{k,\ell}\right) = -\log(m) . \tag{25}$$

To see this, let us define the two maps $g_1 : (w_1,\ldots,w_m) \mapsto \sum_{\ell=1}^m w_\ell\log(w_\ell)$ and $g_2 : (w_1,\ldots,w_m) \mapsto \left(\sum_{\ell'=1}^m w'_\ell\right)^{-1}(w_1,\ldots,w_m)$. Observe then that the map g_1 is a continuous transformation (defined on the

simplex) and g_2 is a continuous transformation on the space of nonnegative weights. If we further denote by $(\tilde{W}_{k,\ell})_{\ell=1}^m$ the unnormalised weights (i.e $\tilde{W}_{k,\ell} = f(X_{k,\ell})/q_{k-1}(X_{k,\ell})$ for all $\ell = 1 \dots m$), then, by virtue of the continuous mapping theorem, it is enough to show that $(\tilde{W}_{k,1}, \dots, \tilde{W}_{k,m})$ converges to $(1, \dots, 1)$ in L_1 to prove (25). Since for all $\ell = 1, \dots, m$,

$$\mathbb{E}[|\tilde{W}_{k,\ell} - 1|] = \int |f - q_{k-1}| \,,$$

this follows from Scheffé's lemma and the proof is concluded.

□

F Additional Experiments

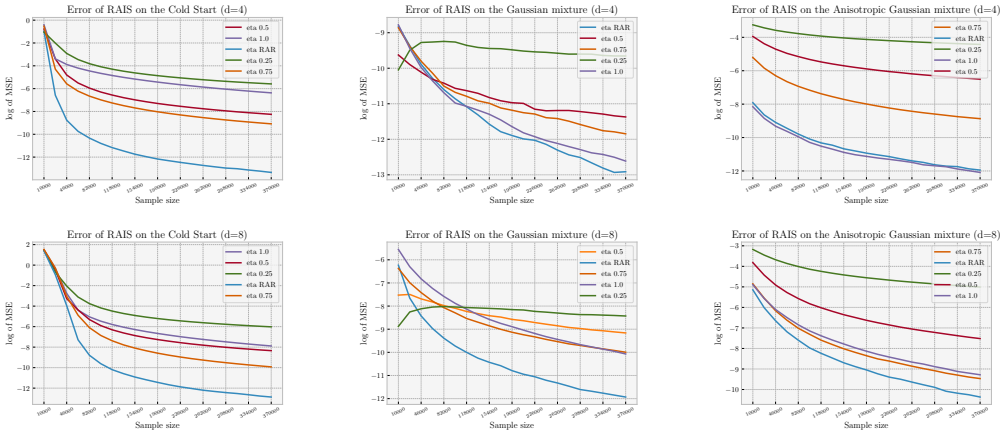

Figure 4: Logarithm of the average squared error, computed over 50 replicates.

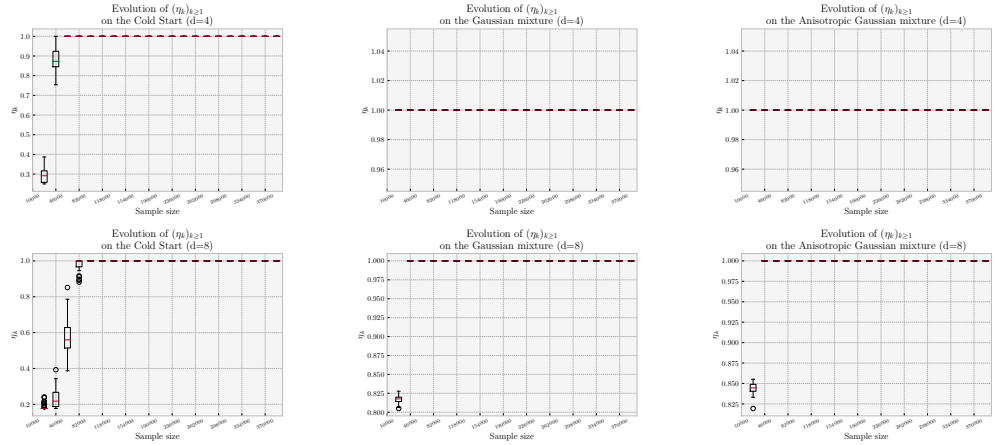

Figure 5: Boxplot of the values of η obtained from the ADA strategy.

We report in Figure 4 the results of the proposed method, and in Figure 5 the evolution of η in smaller dimensions. We can see in Figure 4 that Algorithm 1 along with the subroutine Algorithm 2 always outperform the competitive schedules for the regularization, and in Figure 5 that Algorithm 2 converges to 1.

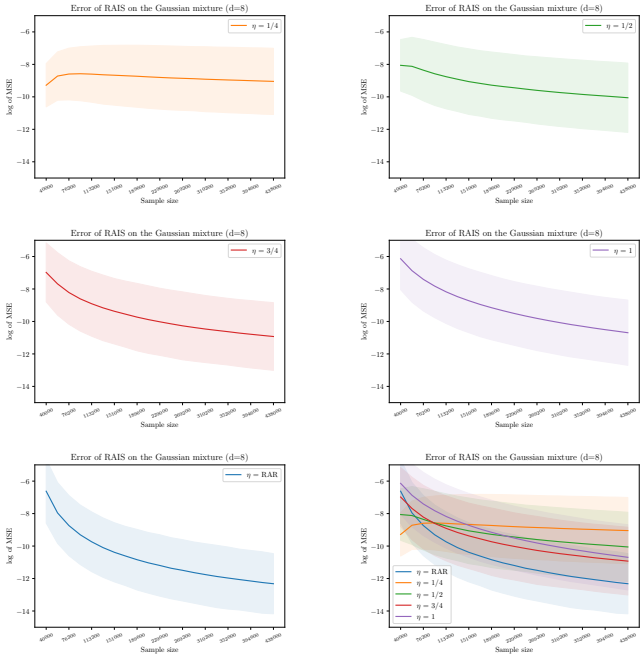

Figure 6: Logarithm of the average squared error with confidence intervals, computed over 500 replicates.