BR-SNIS: Bias Reduced Self-Normalized Importance Sampling

Anonymous Author(s)
Affiliation
Address
email

Abstract

Importance Sampling (IS) is a method for approximating expectations under a target distribution using independent samples from a proposal distribution and the associated importance weights. In many applications, the target distribution is known only up to a normalization constant, in which case self-normalized IS (SNIS) can be used. While the use of self-normalization can have a positive effect on the dispersion of the estimator, it introduces bias. In this work, we propose a new method, BR-SNIS, whose complexity is essentially the same as that of SNIS and which significantly reduces bias without increasing the variance. This method is a wrapper in the sense that it uses the same proposal samples and importance weights as SNIS, but makes clever use of iterated sampling–importance resampling (i-SIR) to form a bias-reduced version of the estimator. We furnish the proposed algorithm with rigorous theoretical results, including new bias, variance and high-probability bounds, and these are illustrated by numerical examples.

1 Introduction

Background and previous work: *Importance sampling* [15, 1] (IS) is a classical Monte Carlo technique for estimating expectations under some given probability distribution (the *target*) on the basis of a sample of draws from a different distribution (the *proposal*). In the modern era of artificial intelligence and statistical machine learning, characterized by large computational resources and Bayesian inference, IS technologies are enjoying a revival; see, *e.g.*, [33, 20] and [11] for a recent survey. The method is not only relevant to situations where sampling from the target is intractable; it can also be used to achieve variance reduction [21]. When the proposal is dominating the target—in the sense that the support of the latter is contained in the support of the former—unbiased estimation can be achieved by assigning each draw an *importance weight* given by the likelihood ratio between the target and the proposal. In the very common case where the target is known only up to a normalizing constant, consistent estimation can still be achieved by simply normalizing each importance weight by the total weight of the sample; however, since such *self-normalized importance sampling* (SNIS) involves ratios of random variables, the procedure can only be implemented at the cost of bias, which can be significant in some applications.

More precisely, let $(\mathbb{X}, \mathcal{X})$ be some state space and $\pi(\mathrm{d}x) \propto w(x)\lambda(\mathrm{d}x)$ a given target probability distribution, where w and λ are a positive weight function and a proposal probability distribution on $(\mathbb{X}, \mathcal{X})$, respectively, such that the normalizing constant $\lambda(w) = \int w(x)\lambda(\mathrm{d}x)$ (this will be our generic notation for Lebesgue integrals) of π is finite. The SNIS estimator is given by

$$\Pi_M f(X^{1:M}) = \textstyle\sum_{i=1}^M \omega_M^i f(X^i), \qquad \omega_M^i = w(X^i)/\sum_{\ell=1}^M w(X^\ell) \tag{1}$$

where $X^{1:M} = (X^1, \ldots, X^M)$ are independent draws from λ, and can be used to approximate $\pi(f) = \int f(x)\pi(\mathrm{d}x)$ for any test function f such that $\pi(|f|) < \infty$. The estimator (1) can be calculated without knowledge of the normalizing constant $\lambda(w)$, which is intractable in general.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

The SNIS estimator is known to be biased; provided that $\lambda(w^2) < \infty$, the bias and mean-squared error (MSE) of the SNIS estimator (1) over bounded test functions f satisfying $\|f\|_\infty \leq 1$ are given respectively (see [1, Theorem 2.1]) by

$$|\mathbb{E}[\Pi_M f(X^{1:M})] - \pi(f)| \leq (12/M)\kappa[\pi, \lambda], \quad \mathbb{E}[\{\Pi_M f(X^{1:M}) - \pi(f)\}^2] \leq (4/M)\kappa[\pi, \lambda], \quad (2)$$

where $\kappa[\pi, \lambda] = \lambda(w^2)/\lambda^2(w)$. Although IS is primarily intended to approximate integrals in the form $\pi(f)$, it can also be used to generate unweighted samples being approximately distributed according to π. In this paper, we consider *iterated sampling importance resampling* (i-SIR), proposed in [40]; see [4, 24, 23, 5]. The i-SIR can be seen as an iterative application of the *sampling importance resampling* (SISR) algorithm proposed by [36]; the k-th iteration is defined as follows. Given a state $Y_k \in \mathbb{X}$, (i) set $X_{k+1}^1 = Y_k$ and draw $X_{k+1}^{2:N}$ independently from the proposal distribution λ; (ii) compute, for $i \in \{1, \ldots, N\}$, the normalized importance weights $\omega_{N,k+1}^i = w(X_{k+1}^i)/\sum_{\ell=1}^N w(X_{k+1}^\ell)$; (iii) select Y_{k+1} from the set $X_{k+1}^{1:N}$ by choosing X_{k+1}^i with probability $\omega_{N,k+1}^i$. In the following, Y_{k+1} and $X_{k+1}^{1:N}$ will be referred to as the *state* and the *candidate pool*, respectively. Following [40] (see Section 2.1), i-SIR may be viewed (up to an irrelevant permutation of the samples) as a two-stage Gibbs sampler targeting an extended probability distribution $\boldsymbol{\varphi}_N$ on an enlarged state space including the state as well as the candidate pool. As this extended distribution allows π as a marginal with respect to the state, one can expect the marginal distribution of the generated states $(Y_k)_{k\in\mathbb{N}}$, forming themselves a Markov chain, to approach the target π of interest as k tends to infinity.

This paper: In i-SIR, the only function of the candidate pool is to guide the states selected at stage (iii) towards the target. Thus, since all rejected candidates are discarded, the approach results generally in a large waste of computational work. Thus, in the present paper we propose to recycle *all* the generated samples by incorporating all the proposed candidates $X_k^{1:N}$ into the estimator rather than only the selected candidate Y_k. We proceed in three steps. First, we show that under the stationary distribution $\boldsymbol{\varphi}_N$ of the process $(Y_k, X_k^{1:N})_{k\in\mathbb{N}}$ generated by i-SIR, the expectation of $\Pi_N f(X_k^{1:N})$ (given by (1)) equals $\pi(f)$ for every valid test function f (see Theorem 2). Second, we establish that since i-SIR is nothing but a systematic-scan Gibbs sampler, the two processes $(X_k^{1:N})_{k\in\mathbb{N}}$ and $(Y_k)_{k\in\mathbb{N}}$ are *interleaving* (see Theorem 5); thus, if $(Y_k)_{k\in\mathbb{N}}$ is uniformly geometrically ergodic, so is $(X_k^{1:N})_{k\in\mathbb{N}}$ with the same mixing rate κ_N. Third, as the main result of the present paper, we establish a novel $\mathcal{O}(\kappa_N^k/N)$ bound on the bias of the estimator $\Pi_N f(X_k^{1:N})$ (see Theorem 3), where the exponentially diminishing factor κ_N^k indicates a drastic bias reduction *vis-à-vis* the standard IS estimator (1) based on i.i.d. samples. As a consequence, approximating $\pi(f)$ by the average of $(\Pi_N f(X_k^{1:N}))_{\ell=k_0+1}^k$, where the "burn-in" period k_0 should be chosen proportionally to the mixing time of the process, yields an estimator whose bias can be furnished with a bound which is, roughly, proportional to $\kappa_N^{k_0}$ and inversely proportional to the total number $M = kN$ of samples generated in the algorithm (see Theorem 4). To complete the theoretical analysis of these estimators, we also equip the same with variance bounds. The procedure of recycling, as described above, all the samples generated in the i-SIR and to incorporate, at negligible computational cost, the same into the final estimator, will from now on be referred as BR-SNIS. Finally, we test numerically the proposed estimators and illustrate how a significant bias reduction relatively to the standard i-SIR can be obtained at basically no cost.

To sum up, our contribution is twofold, since we

- propose a new algorithm, BR-SNIS, which makes better use of the available computational resources by recycling the candidate pool generated at each iteration of i-SIR.
- furnish the proposed algorithm with rigorous theoretical results, including novel bias, variance, and high-probability bounds which support our claim that sample recycling may lead to drastic bias reduction without impairing the variance.

2 Main results

2.1 Statements

The i-SIR algorithm can be interpreted as a systematic-scan two-stage Gibbs sampler, alternately sampling from the full conditions of an extended target $\boldsymbol{\varphi}_N$ on the product space of states and candidate pools. Once the extended target $\boldsymbol{\varphi}_N$ is properly defined, these full conditionals can be retrieved from a dual representation of $\boldsymbol{\varphi}_N$ presented in Theorem 1. In order to define $\boldsymbol{\varphi}_N$, we

introduce the Markov kernel (see Appendix A.1 for comments)

$$\boldsymbol{\Lambda}_N(y, \mathrm{d}x^{1:N}) = N^{-1} \sum_{i=1}^N \delta_y(\mathrm{d}x^i) \prod_{j \neq i} \lambda(\mathrm{d}x^j)$$

on $\mathbb{X} \times \mathcal{X}^{\otimes N}$, which describes probabilistically the sampling operation (i) in i-SIR. Using the kernel $\boldsymbol{\Lambda}_N$ we may now define properly the extended target $\boldsymbol{\varphi}_N$ as the probability law

$$\boldsymbol{\varphi}_N(\mathrm{d}(y, x^{1:N})) = \pi(\mathrm{d}y)\boldsymbol{\Lambda}_N(y, \mathrm{d}x^{1:N}) = N^{-1} \sum_{i=1}^N \pi(\mathrm{d}y)\delta_y(\mathrm{d}x^i) \prod_{j \neq i} \lambda(\mathrm{d}x^j)$$

on $(\mathbb{X}^{N+1}, \mathcal{X}^{\otimes(N+1)})$. Note that since for every $A \in \mathcal{X}$, $\boldsymbol{\varphi}_N(1_{A \times \mathbb{X}^N}) = \pi(A)$, the target π coincides with the marginal of $\boldsymbol{\varphi}_N$ with respect to the state. Moreover, it is easily seen that $\boldsymbol{\Lambda}_N$ provides the conditional distribution, under $\boldsymbol{\varphi}_N$, of the candidate pool given the state. Defining the kernels

$$\Gamma_N(x^{1:N}, \mathrm{d}y) = N^{-1} \sum_{i=1}^N w(x^i)\delta_{x^i}(\mathrm{d}y), \quad \Pi_N(x^{1:N}, \mathrm{d}y) = \Gamma_N(x^{1:N}, \mathrm{d}y)/\Gamma_N 1_{\mathbb{X}}(x^{1:N}) \quad (3)$$

on $\mathbb{X} \times \mathcal{X}^{\otimes N}$, the marginal distribution $\boldsymbol{\pi}_N$ of $\boldsymbol{\varphi}_N$ with respect to $x^{1:N}$ is given by

$$\boldsymbol{\pi}_N(\mathrm{d}x^{1:N}) = \lambda(w)^{-1}\Gamma_N 1_{\mathbb{X}}(x^{1:N}) \prod_{j=1}^N \lambda(\mathrm{d}x^j). \quad (4)$$

It is interesting to note that the marginal $\boldsymbol{\pi}_N$ has a probability density function, proportional to $\Gamma_N 1_{\mathbb{X}}(x^{1:N}) = \sum_{i=1}^N w(x^i)/N$, with respect to the product measure $\lambda^{\otimes N}$. Using (4), we immediately obtain the following result.

Theorem 1 (duality of extended target)**.** *For every $N \in \mathbb{N}^*$,*

$$\boldsymbol{\varphi}_N(\mathrm{d}(y, x^{1:N})) = \pi(\mathrm{d}y)\boldsymbol{\Lambda}_N(y, \mathrm{d}x^{1:N}) = \boldsymbol{\pi}_N(\mathrm{d}x^{1:N})\Pi_N(x^{1:N}, \mathrm{d}y). \quad (5)$$

Note that the second identity of the dual representation (5) provides also the conditional distribution, under $\boldsymbol{\varphi}_N$, of the state given the candidates. Consequently, i-SIR is a systematic scan two-stage Gibbs sampler which generates a Markov chain $(X_k, Y_k)_{k \in \mathbb{N}}$ with time-homogeneous Markov kernel

$$\mathbf{P}_N((y_k, x_k^{1:N}), \mathrm{d}(y_{k+1}, x_{k+1}^{1:N})) = \boldsymbol{\Lambda}_N(y_k, \mathrm{d}x_{k+1}^{1:N})\Pi_N(x_{k+1}^{1:N}, \mathrm{d}y_{k+1})$$

on $\mathbb{X}^{N+1} \times \mathcal{X}^{\otimes(N+1)}$. Note that the law $\mathbf{P}_N(y_k, x_k^{1:N}, \cdot)$ does not depend on $x_k^{1:N}$, which means that only the state Y_k needs to be stored from one iteration to the other. Thus, $(Y_k)_{k \in \mathbb{N}}$ is a Markov chain with Markov transition kernel

$$\mathsf{P}_N(y_k, \mathrm{d}y_{k+1}) = \int \boldsymbol{\Lambda}_N(y_k, \mathrm{d}x_{k+1}^{1:N})\Pi_N(x_{k+1}^{1:N}, \mathrm{d}y_{k+1}) = \boldsymbol{\Lambda}_N \Pi_N(y_k, \mathrm{d}y_{k+1}), \quad (6)$$

(where integration is w.r.t. $x_{k+1}^{1:N}$) on $\mathbb{X} \times \mathcal{X}$. Given some probability distribution $\boldsymbol{\xi}$ on $(\mathbb{X}^{N+1}, \mathcal{X}^{\otimes(N+1)})$, we denote by $\mathbb{P}_{\boldsymbol{\xi}}$ the law of the canonical Markov chain $(X_k, Y_k)_{k \in \mathbb{N}}$ with kernel $\mathbf{P}_N$ and initial distribution $\boldsymbol{\xi}$. Our first results establishes the unbiasedness of the estimator $\Pi_N f(X^{1:N})$ under $\boldsymbol{\varphi}_N$.

Theorem 2. *For every $N \in \mathbb{N}^*$ and π-integrable function f,*

$$\int \Pi_N f(x^{1:N})\boldsymbol{\pi}_N(\mathrm{d}x^{1:N}) = \pi(f).$$

The proof of Theorem 2 is postponed to Appendix A.3. Next, we present theoretical bounds on the discrepancy, in terms of bias, MSE and covariance, between $\Pi_N f(X_k^{1:N})$ and $\pi(f)$, for bounded target functions f, when the i-SIR chain is initialized according to an arbitrary distribution $\boldsymbol{\xi}$. We will work under the following assumption.

A1. *It holds that $\omega = \|w\|_\infty/\lambda(w) < \infty$.*

Under **A**1, the state and candidate-pool Markov chains $(Y_k)_{k \in \mathbb{N}}$ and $(X^{1:N})_{k \in \mathbb{N}}$ can be shown to be uniformly geometrically ergodic with mixing rate and mixing-time upper bound

$$\kappa_N = (2\omega - 1)/(2\omega + N - 2), \quad \tau_{mix,N} = \lceil -\ln 4/\ln \kappa_N \rceil, \quad (7)$$

respectively; see Theorem 6 below for details. Here the mixing time $\tau_{mix,N}$ grows logarithmically with the sample size N. The exact value of $\tau_{mix,N}$ is likely to be grossly pessimistic, but we conjecture that the logarithmic dependence in the minibatch size holds true. In addition, under **A**1 we define the constants

$$\begin{aligned} \varsigma^{bias} &= 4(\kappa[\pi, \lambda] + 1 + \omega) \\ \varsigma_i^{mse} &= 4(\kappa[\pi, \lambda]1_{\{0,1\}}(i) + (1+\omega)^2 1_{\{1,2\}}(i)), \quad \varsigma_i^{cov} = \varsigma^{bias}(\varsigma_i^{mse})^{1/2}, \quad i \in \{0, 1, 2\}. \end{aligned} \quad (8)$$

With these definitions, the following holds true.

Theorem 3. *Assume A 1. Then for every initial distribution $\boldsymbol{\xi}$ on $(\mathbb{X}^{N+1}, \mathcal{X}^{\otimes(N+1)})$, bounded measurable function f on $(\mathbb{X}, \mathcal{X})$ such that $\|f\|_\infty \leq 1$, $N \geq 2$, and $(k, \ell) \in (\mathbb{N}^*)^2$,*

(i) $\left|\mathbb{E}_{\boldsymbol{\xi}}[\Pi_N f(X_k^{1:N})] - \pi(f)\right| \leq \varsigma^{bias}(N-1)^{-1}\kappa_N^{k-1}$,

(ii) $\mathbb{E}_{\boldsymbol{\xi}}[\{\Pi_N f(X_k^{1:N}) - \pi(f)\}^2] \leq \sum_{i=0}^{2} \varsigma_i^{mse}(N-1)^{-1-i/2}$,

(iii) $\left|\mathbb{E}_{\boldsymbol{\xi}}[\{\Pi_N f(X_k^{1:N}) - \pi(f)\}\{\Pi_N f(X_{k+\ell}^{1:N}) - \pi(f)\}]\right| \leq \kappa_N^{\ell-1} \sum_{i=0}^{2} \varsigma_i^{cov}(N-1)^{-(3-i/2)/2}$,

where constants are given in (7) *and* (8).

It is worth noting that the bias decreases inversely with the number of candidates and exponentially with the number of iterations (the mixing time of the chain also depends on N). The MSE is also inversely proportional to the number of candidates N. In the light of the previous results, it is natural to consider an estimator formed by an average across the IS estimators $(\Pi_N f(X_k^{1:N}))_{k\in\mathbb{N}}$ associated with the candidate pools generated at the different i-SIR iterations. To mitigate the bias, we remove a "burn-in" period whose length k_0 should be chosen proportional to the mixing time $\tau_{mix,N}$ of the Markov chain $(Y_k)_{k\in\mathbb{N}}$ (which turns out to coincide with that or the chain $(X_k^{1:N})_{k\in\mathbb{N}}$; see Section 2.2). This yields the estimator

$$\Pi_{(k_0,k),N}(f) = (k-k_0)^{-1} \sum_{\ell=k_0+1}^{k} \Pi_N f(X_\ell^{1:N}) \tag{9}$$

of $\pi(f)$. The total number of samples (generated by the proposal λ) underlying this estimator is $M = (N-1)k$. Importantly, all the importance weights included in the estimators are obtained as a by-product of the i-SIR schedule; thus, it is, for a given budget of simulations (*i.e.*, under the constraint that $(k-k_0)N$ is constant), possible to compute $\Pi_{(k_0,k),N}(f)$ for different values of k_0, k and N with a negligible computational cost. We denote by $\upsilon = (k-k_0)/k$ the ratio of the number of candidate pools used in the estimator to the total number of sampled such pools.

Our final main result provides bounds on the bias and the MSE of the estimator (9) as well as a high-probability bound for the same. Define $\zeta^{bias} = 4\tau_{mix,N}\varsigma^{bias}/3$, $\zeta_i^{mse} = \varsigma_{(i+1)\wedge 2}^{mse} 1_{\{0,2\}}(i) + (8/3)\tau_{mix,N}\varsigma_i^{cov}$, $i \in \{0,1,2\}$, $\zeta^{mse} = \zeta_0^{mse} + \zeta_1^{mse}(N-1)^{-1/4} + \zeta_2^{mse}(N-1)^{-1}$, and $\mathsf{MSE}_M^{is} = (4/M)\kappa[\pi,\lambda]$, see (2).

Theorem 4. *Assume A 1. Then the following holds true for every initial distribution $\boldsymbol{\xi}$ on $(\mathbb{X}^{n+1}, \mathcal{X}^{\otimes(n+1)})$, bounded measurable function f on $(\mathbb{X}, \mathcal{X})$ such that $\|f\|_\infty \leq 1$, and $N \geq 2$.*

(i) $\left|\mathbb{E}_{\boldsymbol{\xi}}[\Pi_{(k_0,k),N}(f)] - \pi(f)\right| \leq \zeta^{bias}(\upsilon M)^{-1} 4^{-k_0/\tau_{mix,N}}$

(ii) $\mathbb{E}_{\boldsymbol{\xi}}[\{\Pi_{(k_0,k),N}(f) - \pi(f)\}^2] \leq \mathsf{MSE}_{\upsilon M}^{is} + \zeta^{mse}(\upsilon M)^{-1}(N-1)^{-1/2}$

(iii) For every $\delta \in (0,1)$, $|\Pi_{(k_0,k),N}(f) - \pi(f)| \leq \varsigma^{hpd}(\upsilon M)^{-1/2}(\log(4/\delta))^{1/2}$ with probability at least $1-\delta$, where $\varsigma^{hpd} = 664\omega$.

Bootstrap: As established in Theorem 4, the bias of the BR-SNIS estimator decreases exponentially with the burn-in period k_0, leading to potentially significant bias reduction with respect to SNIS. Still, using a large k_0 is done at a price of increased overall MSE (mainly through the term $\mathsf{MSE}_{\upsilon M}^{is}$ in Theorem 4(ii), which is directly related to k_0 via υ). A natural way to reduce the variance is to use bootstrap. More precisely, we first apply a random permutation to the samples and re-compute BR-SNIS on the basis of the bootstrapped samples. After this, we produce a final estimator by averaging over the bootstrapped BR-SNIS replicates. In most applications, the major computational bottleneck consists of sampling from λ and evaluating w and f at the samples; thus, the additional operations that this bootstrap approach entails are computationally cheap. Therefore, in our experiments, we use bootstrap in combination with the choice $k_0 = k-1$ (in order to minimize the bound in Theorem 4(i)).

2.2 Elements of proofs

Ergodic properties of i-SIR: The systematic scan two-stage Gibbs sampler is a well-studied MCMC algorithmic structure, and we summarize its most important properties in Theorem 5 below; see [27, 3] and [35, Chapter 9] as well as the references therein. In particular, as shown in [27], the state and candidate-pool Markov chains $(Y_k)_{k\in\mathbb{N}}$ and $(X_k^{1:N})_{k\in\mathbb{N}}$ satisfy a duality property referred to as *interleaving* (Theorem 5(iii)).

Theorem 5. *Assume that for every $x \in \mathbb{X}$, $w(x) > 0$, $\lambda(w) < \infty$ and that there exists a set $C \in \mathcal{X}$ such that $\lambda(C) > 0$ and $\sup_{x\in C} w(x)/\lambda(w) < \infty$. Then,*

(i) the Markov kernel $\mathbf{P}_N$ is Harris recurrent and ergodic with unique invariant distribution $\boldsymbol{\varphi}_N$.
(ii) the Markov kernel P_N is π-reversible, Harris recurrent and ergodic.
(iii) the two Markov chains $(Y_k)_{k\in\mathbb{N}}$ and $(X_k^{1:N})_{k\in\mathbb{N}}$ are conjugate of each other with the interleaving property, i.e., for every initial distribution $\boldsymbol{\xi}$ and $k \in \mathbb{N}$, under $\mathbb{P}_{\boldsymbol{\xi}}$,
  (a) $X_k^{1:N}$ and $X_{k+1}^{1:N}$ are conditionally independent given Y_k,
  (b) Y_k and Y_{k+1} are conditionally independent given $X_{k+1}^{1:N}$;
  (c) moreover, under $\mathbb{P}_{\boldsymbol{\varphi}_N}$, $(Y_k, X_{k-1}^{1:N})$ and $(Y_k, X_k^{1:N})$ are identically distributed.

The ergodic behavior of the i-SIR algorithm has been studied in many works; see [23, 25, 5] in particular. The analysis is particularly simple under the assumption that the importance weight function w is bounded, as imposed by **A**1. Recall that the *total variation-distance* between two probability measures ξ and ξ' on $(\mathbb{X}, \mathcal{X})$ is given by $d_{TV}(\xi, \xi') = \sup_{g:\mathrm{osc}(g)\leq 1}\{\xi(g) - \xi'(g)\}$, where $\mathrm{osc}(g) = \sup_{(x,x')\in\mathbb{X}^2} |g(x) - g(x')|$ denotes the oscillator norm of a measurable function g. The following result establishes the uniform geometric ergodicity of the state chain $(Y_k)_{k\in\mathbb{N}}$.

Theorem 6. *Assume **A**1. Then for every $N \geq 2$, $y \in \mathbb{X}$ and $k \in \mathbb{N}$, $d_{TV}(\mathsf{P}_N^k(y, \cdot), \pi) \leq \kappa_N^k$, where κ_N is given in* (7)*.*

The proof is given in [25, 5], but we provide it in Appendix A.5 for completeness. For uniformly ergodic Markov chains, it is often more appropriate to work with the mixing time

$$\min\{k \in \mathbb{N} : \sup_{y\in\mathbb{X}} d_{TV}(\mathsf{P}_N^k(y, \cdot), \pi) \leq 1/4\} \leq \tau_{mix,N}$$

(where $\tau_{mix,N}$ is given in (7)), *i.e.*, the number of time steps required for the distribution of the chain to be within a certain total variation distance from its stationary distribution [2, 14]. An interesting consequence of the interleaving property is that if the Markov chain $(Y_k)_{k\in\mathbb{N}}$ is (geometrically) ergodic, then the Markov chain $(X_k^{1:N})_{k\in\mathbb{N}}$ is (geometrically) ergodic as well with the same mixing time; see [35, Corollary 9.14]).

Bias of the BR-SNIS estimator: As the BR-SNIS estimator $\Pi_N f(X_k^{1:N})$ (where Π_N is defined in (3)) is made up by a ratio of the two unnormalized estimators $\Gamma_N f(X_k^{1:N})$ and $\Gamma_N 1_{\mathbb{X}}(X_k^{1:N})$, a key ingredient in the proof of Theorem 3 is to bound the bias and the pth order moments of statistics defined as ratios of sums of random variables that are not necessarily independent. The basic idea is to reduce the study of these relations to the analysis of the moments of the numerator and the denominator of these statistics and to exploit their concentration around the respective (conditional and unconditional) means. The main results that we will use in the rest of the paper are summarized in Appendix B.

Lemma 7. *For every initial distribution $\boldsymbol{\xi}$ on $(\mathbb{X}^{N+1}, \mathcal{X}^{\otimes(N+1)})$, $k \in \mathbb{N}^*$, and bounded measurable function $f : \mathbb{X} \to \mathbb{R}$, it holds that*

(i) for every $y \in \mathbb{X}$, $\boldsymbol{\Lambda}_N \Gamma_N f(y) = (1 - 1/N)\lambda(wf) + (1/N)w(y)f(y)$,
(ii) $\mathbb{E}_{\boldsymbol{\xi}}\left[\Gamma_N f(X_k^{1:N}) \,\middle|\, Y_{k-1}\right] = \boldsymbol{\Lambda}_N \Gamma_N f(Y_{k-1})$, $\mathbb{P}_{\boldsymbol{\xi}}$-a.s.,
(iii) $\mathbb{E}_{\boldsymbol{\xi}}\left[\{\Gamma_N f(X_k^{1:N}) - \boldsymbol{\Lambda}_N \Gamma_N f(Y_{k-1})\}^2 \,\middle|\, Y_{k-1}\right] = (N-1)/N^2 \lambda(\{wf - \lambda(wf)\}^2)$, $\mathbb{P}_{\boldsymbol{\xi}}$-a.s.

We now have all the elements that allow us to determine the first important result of this work, namely the bias and the MSE of the estimator $\Pi_N f(X_k^{1:N})$ of $\pi(f)$.

Proof of Theorem 3. We establish the bias bound in (i) and postpone the proof of the bounds on the MSE and the covariance in (ii) and (iii) to the supplement. Define the measure $\xi(A) = \boldsymbol{\xi}(A \times \mathbb{X})$, $A \in \mathcal{X}$, and the kernel $\mathsf{P}_N = \boldsymbol{\Lambda}_N \Pi_N$ on $\mathbb{X} \times \mathcal{X}$. Consequently, $\mathsf{P}_N f(Y_{k-1}) = \mathbb{E}_{\boldsymbol{\xi}}[\Pi_N f(X_k^{1:N}) \mid Y_{k-1}]$ and $\boldsymbol{\Lambda}_N \Gamma_N f(Y_{k-1}) = \mathbb{E}_{\boldsymbol{\xi}}[\Gamma_N f(X_k^{1:N}) \mid Y_{k-1}]$, $\mathbb{P}_{\boldsymbol{\xi}}$-a.s. Since $(Y_k)_{k\in\mathbb{N}}$ is, under $\mathbb{P}_{\boldsymbol{\xi}}$, a Markov chain with initial distribution ξ and Markov kernel P_N (see (6)), it holds that

$$\mathbb{E}_{\boldsymbol{\xi}}[\Pi_N f(X_k^{1:N})] = \mathbb{E}_{\boldsymbol{\xi}}[\mathsf{P}_N f(Y_{k-1})] = \mathbb{E}_{\boldsymbol{\xi}}[\mathbb{E}_{\boldsymbol{\xi}}\left[\mathsf{P}_N f(Y_{k-1}) \,|\, Y_0\right]] = \xi \mathsf{P}_N^{k-1} \mathsf{P}_N f.$$

Consequently, the proof is concluded by establishing that for every $k \in \mathbb{N}^*$,

$$\left|\xi \mathsf{P}_N^{k-1} \mathsf{P}_N f - \pi(f)\right| \leq \varsigma^{bias} \kappa_N^{k-1} (N-1)^{-1}. \tag{10}$$

On the other hand, since by Theorem 2, $\pi(\mathsf{P}_N f) = \pi(f)$, we may use Theorem 6 to obtain the bound

$$|\xi \mathsf{P}_N^{k-1} \mathsf{P}_N f - \pi(f)| = |\xi \mathsf{P}_N^{k-1} \mathsf{P}_N f - \pi(\mathsf{P}_N f)| \leq \kappa_N^{k-1} \,\mathrm{osc}(\mathsf{P}_N f).$$

Finally, we establish (10) by bounding $\mathrm{osc}(\mathsf{P}_N f)$. Note that

$$\mathrm{osc}(\mathsf{P}_N f) \leq 2 \left\| \mathsf{P}_N f - \mathbf{\Lambda}_N \Gamma_N f / (\mathbf{\Lambda}_N \Gamma_N 1_{\mathbb{X}}) \right\|_\infty + 2 \left\| \mathbf{\Lambda}_N \Gamma_N f / (\mathbf{\Lambda}_N \Gamma_N 1_{\mathbb{X}}) - \pi(f) \right\|_\infty ,$$

where, for every $y \in \mathbb{X}$, using Theorem 11,

$$\begin{aligned} &|\mathsf{P}_N f(y) - \mathbf{\Lambda}_N \Gamma_N f(y) / \mathbf{\Lambda}_N \Gamma_N 1_{\mathbb{X}}(y)| \\ &\leq \frac{1}{2} \{\mathbf{\Lambda}_N \Gamma_N 1_{\mathbb{X}}(y)\}^{-2} \{\mathbf{\Lambda}_N [\{\Gamma_N f - \mathbf{\Lambda}_N \Gamma_N f(y)\}^2](y) + 3 \mathbf{\Lambda}_N [\{\Gamma_N 1_{\mathbb{X}} - \mathbf{\Lambda}_N \Gamma_N 1_{\mathbb{X}}(y)\}^2](y)\}. \end{aligned}$$

Now, since $\mathbf{\Lambda}_N \Gamma_N 1_{\mathbb{X}}(y) \geq (1 - 1/N)\lambda(w)$, we get, using Lemma 7,

$$\begin{aligned} \left\| \mathsf{P}_N f - \frac{\mathbf{\Lambda}_N \Gamma_N f}{\mathbf{\Lambda}_N \Gamma_N 1_{\mathbb{X}}} \right\|_\infty &\leq (2(N-1))^{-1} \{\lambda(w)\}^{-2} \{\lambda(\{wf - \lambda(wf)\}^2) + 3\lambda(\{w - \lambda(w)\}^2)\} \\ &\leq 2(N-1)^{-1} \lambda(w^2)/(\lambda(w))^2. \end{aligned} \tag{11}$$

On the other hand, using the elementary inequality $a/b - c/d = a(d-b)/bd + (a-c)/d$, we get, as $\pi(f) = \lambda(wf)/\lambda(w)$,

$$\frac{\mathbf{\Lambda}_N \Gamma_N f(y)}{\mathbf{\Lambda}_N \Gamma_N 1_{\mathbb{X}}(y)} - \pi(f) = (1/N) \frac{\mathbf{\Lambda}_N \Gamma_N f(y)}{\mathbf{\Lambda}_N \Gamma_N 1_{\mathbb{X}}(y)} \{1 - w(y)/\lambda(w)\} + (1/N)\{w(y)f(y) - \lambda(wf)\}/\lambda(w).$$

Finally, the bound (10) is established by noting that

$$\left\| \mathbf{\Lambda}_N \Gamma_N f / (\mathbf{\Lambda}_N \Gamma_N 1_{\mathbb{X}}) - \pi(f) \right\|_\infty \leq 2N^{-1}\{1 + w(y)/\lambda(w)\} \leq 2N^{-1}(1 + \omega). \tag{12}$$

□

2.3 Related works

The first use of the IS method, then as a variance reduction technique, dates back to the '50s; see [12, 19] and the references therein. Today, the renewed interest in IS parallels the flurry of activity in the probabilistic ML community and its ever-increasing computational demands; thus, it is impossible to fully present the literature. We therefore limit ourselves to describing results that have inspired our work, and refer the readers to the recent reviews [1, 11] for additional references.

There is clearly a plethora of modern ML applications where the standard SNIS estimator may be substantially improved using the BR-SNIS method. To mention just a selection of examples, SNIS plays a key role for a robust off-policy selection strategy BY [20] (extending [38, 29]), Bayesian problems (see, *e.g.*, [1, Section 3]), Bayesian transfer learning [16, 28], variational autoencoders [9], inference of energy-based models [22], patch-based image restoration [33] and many more.

Despite long-standing interest in SNIS, there are only few theoretical results. For example, [1, Theorem 2.1] provides bounds on the bias and variance of SNIS, results that we extend to BR-SNIS in Theorem 3. Moreover, [29, Proposition D.3] provides a suboptimal variance bound based on a bound for the second-order moment. This result can be compared to the sophisticated sub-Gaussian concentration bound for BR-SNIS obtained in Theorem 4 (a result that can be obtained for SNIS using the same proof mechanism; see Appendix A.8). Finally, [20] obtains a semi-empirical sub-Gaussian concentration inequality using the Efron-Stein estimate of variance and the Harris inequality.

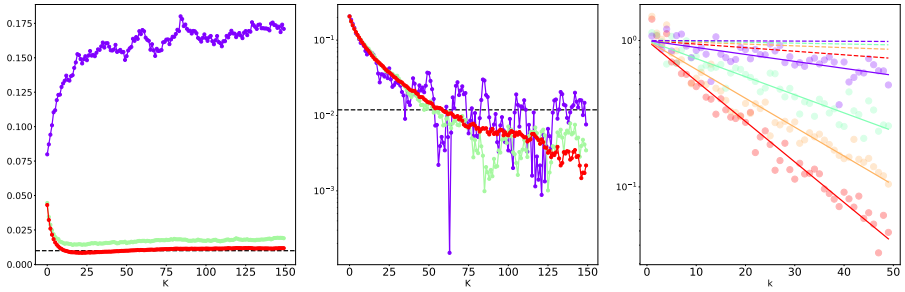
As an MCMC sampling method, the i-SIR algorithm that has been applied successfully in many situations. It was recently used—under the alternative name *conditional importance sampling*—in [30] for *Markovian score climbing*. In the same work, it is mentioned that it is possible to "Rao-Blackwellize" the gradient of the score using the proposed candidates, which is in line with the recycling argument underpinning the estimator suggested by us, but without theoretical justifications. In its most basic form, the i-SIR algorithm appeared in the pioneering work of [40]. The same idea played a key role in the development of the *particle Gibbs sampler* [4, 5, 31], which extends i-SIR principles to *sequential Monte Carlo methods*. An approach very similar to BR-SNIS can be taken also in this context; however, casting BR-SNIS into the framework of particle Gibbs methods is a non-trivial problem which is the subject of ongoing work.

3 Experimental results

In this section we compare numerically the performances of BR-SNIS and SNIS in three different settings: mixture of Gaussians, Bayesian logistic regression and variational autoencoders (VAE). We leave to the supplementary material (Appendix C.1) the detailed numerical verification of the bounds established in Section 2.

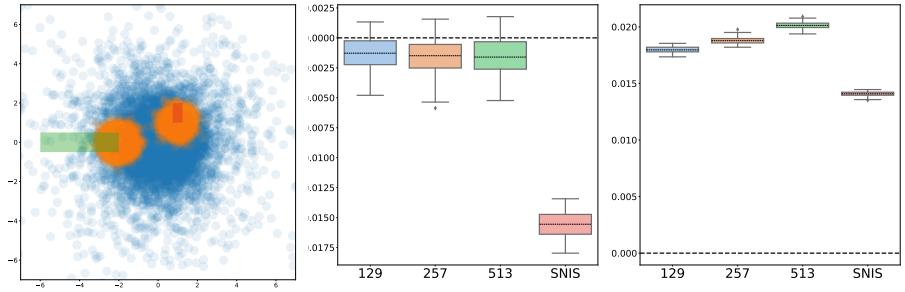
Mixture of Gaussian distributions: We start with an example where the target distribution π is a mixture of two Gaussian distributions of dimension $d = 7$, as shown in Figure 2a. The proposal distribution is a Student distribution with $\nu = 3$ degrees of freedom. The test function is $f = 1_A - 1_B$, where A and B are a d-dimensional rectangle intersecting each of the modes of π (see Appendix C.1 for precise definitions). We verify the positive effect of bootstrap in Figures 1a and 1b by computing the bias and the MSE over 1000 chains for $N = 129$ for several k. The purple, green, and red curves correspond to a number of bootstrap rounds of $1, 21$, and 201, respectively. We illustrate the decay of the mean Sliced Wasserstein distance (according to [7]) with k for different values of N ($N = 8$ purple, $N = 32$ green, $N = 64$ orange, and $N = 128$ red) in Figure 1c. The decay of the Wassertein distance is directly linked to the mixing time of the i-SIR kernel (see (7)), and hence allows us to represent the effective mixing time of the chain. Moreover, we represent the theoretical slopes as dashed lines. This illustrates that the effective value of $\tau_{mix,N}$ is smaller than its theoretical bound. The bias and MSE for SNIS with $M = 25600$ are shown in black dashed lines.

We compare the bias (Figure 2b) and MSE (Figure 2c) of BR-SNIS and SNIS for a fixed budget with a total number of $M = 16384$ samples. We run the experiments 10^6 times; we compute the bias and MSE over batches of 10^4 replications using the true value of $\pi(f)$ computed above (the boxplots in Figure 2 are therefore obtained over 100 replications). For the algorithm BR-SNIS, we used $N \in \{129, 513\}$, $k_0 = k_{max} - 1$ and $k_{max} = M/(N-1)$ bootstrap rounds. As can be seen from



(a) MSE (b) Bias in log scale (c) Sliced Wasserstein

Figure 1

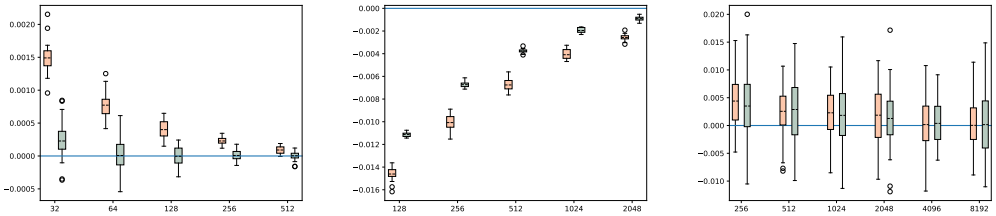


(a) 2d projection (b) Bias (c) MSE

Figure 2: Comparison between SNIS and BR-SNIS for the same budget. In each boxplot the dotted line represents the **mean** value of the samples.

Figure 2b, BR-SNIS significantly reduces bias (by a factor of almost 10) w.r.t. standard SNIS for both configurations, while MSE increases only slightly (at around 20%), as can be seen in Figure 2c. We also show in Appendix C.1 that $k_0 = \lfloor 0.625 k_{max} \rfloor$ can lead to about 3 times less bias w.r.t. standard SNIS while only augmenting the MSE of 10%.

Bayesian Logistic regression: We consider posterior inference in a Bayesian logistic regression model. Let $\mathcal{D}_{train} = (\mathbf{x}_i, y_i)_{i=1}^T$ be a dataset, where each $\mathbf{x}_i \in \mathbb{R}^d$ is a vector of covariates and $y_i \in \{-1, 1\}$ is a binary response. Let $p(y_i \mid \mathbf{x}_i; \theta) = \{1 + \exp(-y_i \, \mathbf{x}_i^\top \theta)\}^{-1}$ be the probability of the ith observation at $\theta \in \Theta \subseteq \mathbb{R}^d$ and $\pi_0(\mathrm{d}\theta)$ be a prior distribution for θ. The Bayesian posterior is



(a) Heart 8 (b) Breast 11 (c) Covertype 6

Figure 3: Visualization of the distribution for each datasets. Each boxplot is grouped by budget, the left one represent SNIS and the right represent BR-SNIS.

given

$$\pi(\mathrm{d}\theta) = \mathrm{Z}^{-1}\pi_0(\mathrm{d}\theta)\exp(\mathcal{L}_T(\theta)), \quad \mathcal{L}_T(\theta) = \textstyle\sum_{i=1}^T \ln p(y_i \mid \mathbf{x}_i;\theta), \quad \mathrm{Z} = \int \exp(\mathcal{L}_T(\theta))\pi_0(\mathrm{d}\theta).$$

For numerical illustration, we use the heart failure clinical records ($d = 13$, $T = 299$), breast cancer detection ($d = 30$, $T = 569$), and Covertype ($d = 55$, $T = 4 \cdot 10^4$) datasets from the UCI machine learning repository. For Covertype, we use Cover type 1 (Spruce/Fir) and Cover type 2 (Lodgepole Pine) classes to define a binary classification problem. As a prior, we use a Gaussian distribution $\mathrm{N}(0, \tau^{-2}\mathbf{I})$ with $\tau^2 = 5 \cdot 10^{-2}$. The importance distribution λ is Gaussian with mean and diagonal covariance learned by variational inference; see Appendix C.2 for details. The boxplots for bias in Figure 3 were constructed in the same way as those in Figure 2. We compare two test

.	CoverType	Breast	Heart
SNIS, M = 32	0.0028 +/- 0.0012	0.00011 +/- 6.04e-5	0.00023 +/- 7.24e-5
BR-SNIS, M= 32	0.0014 +/- 0.0003	7.9e-5 +/- 5.5e-5	0.00012 +/- 6.7e-5
SNIS, M = 512	0.0026 +/- 0.0017	4.3e-5 +/- 3.3e-5	7.8e-5 +/- 6.8e-5
BR-SNIS, M= 512	0.0013 +/- 0.0003	3.5e-5 +/- 2.2e-5	4.9e-5 +/- 5.2e-5

Table 1: Comparison of the TV distance between the posteriors (Lower is better).

functions, $f(\theta) = \theta$, corresponding to evaluation of the posterior mean, and $f(\theta) = p(y \mid \mathbf{x}, \theta)$, where $(\mathbf{x}, y) \in \mathcal{D}_{test}$. This last function allows us to compute a TV distance for the predictive distribution. Indeed, in a classification context, one can compute the TV distance between any two predictive distributions p and $\hat{p}$ as

$$d_{TV}(\hat{p}, p) = T^{-1}\textstyle\sum_{i=1}^T \frac{1}{2}\sum_{j=0}^1 |\hat{p}(y = j \mid \mathbf{x}_i, \mathcal{D}_{train}) - p(y = j \mid \mathbf{x}_i, \mathcal{D}_{train})|,$$

where we compare the predictive distribution $p(y \mid x, \mathcal{D}_{train}) = \int p(y \mid x, \theta)\pi(\theta)\mathrm{d}\theta$ and $\hat{p}$ is the estimation of this quantity, provided in the experiments by SNIS or BR-SNIS. From Figure 3 we can see that for each dataset we have a constant decrease in bias, while the variance increases only slightly. We plot the bias in other components of θ and provide further numerical details in Appendix C.2.

Generative Model: We now extend our methodology to the more complex *deep latent generative models* (DLGM). A DLGM defines a family of probability densities $p_\theta(x)$ over an observation space $x \in \mathbb{R}^P$ by introducing a latent variable $z \in \mathbb{R}^d$, defining the joint density function $p_\theta(x, z)$ (with respect to Lebesgue measure) and aiming to find a parameter θ maximizing the marginal log-likelihood of the model $p_\theta(x) = \int p_\theta(x, z)\mathrm{d}z$. Under simple technical assumptions, by Fisher's identity,

$$\nabla_\theta \log p_\theta(x) = \int \nabla_\theta \log p_\theta(x, z) p_\theta(z \mid x)\mathrm{d}z, \tag{13}$$

In most cases, the conditional density $p_\theta(z \mid x) = p_\theta(x, z)/p_\theta(x)$ is intractable and can only be sampled. The variational autoencoder [18] is based on the introduction of an additional parameter ϕ and a family of variational distributions $q_\phi(z \mid x)$. The joint parameters $\{\theta, \phi\}$ are then inferred by maximizing the *evidence lower bound* (ELBO) defined by

$$\mathcal{L}(\theta, \phi) = \log p_\theta(x) - \mathrm{KL}\big(q_\phi(\cdot \mid x) \parallel p_\theta(\cdot \mid x)\big) \leq \log p_\theta(x).$$

d	VAE	IWAE	BR-IWAE ($k = 4$)	BR-IWAE ($k = 8$)
20	-115.1	-91.5	-90.5	-90.4
50	-96.4	-92.9	-90.1	-90.1

Table 2: Comparison of the mean log likelihood over the MNIST validation set (Higher is better).

This basic setup has been further developed and improved in many directions. Here we consider the *importance weighted autoencoder* (IWAE) [8], which relies on SNIS to design a tighter ELBO on the log-likelihood. The objective of the IWAE is given by

$$\mathcal{L}_M(\theta, \phi) = \int \log \left(M^{-1} \sum_{i=1}^M w_{\theta,\phi,x}(z_i) \right) \prod_{\ell=1}^M q_\phi(z_\ell \mid x) \mathrm{d}z_i,$$

where $w_{\theta,\phi,x}(z) = p_\theta(x, z)/q_\phi(z \mid x)$ denote the importance weights. However, writing, following [8, Eq. (13)],

$$\nabla_\theta \mathcal{L}_M(\theta, \phi) = \int \sum_{i=1}^M \omega_{\theta,\phi,x}^{(i)} \nabla_\theta \log w_{\theta,\phi,x}(z_i) \prod_{\ell=1}^N q_\phi(z_\ell \mid x) \mathrm{d}z_\ell,$$

where $\omega_{\theta,\phi,x}^{(i)} = w_{\theta,\phi,x}(z_i)/\sum_{j=1}^M w_{\theta,\phi,x}(z_j)$ are normalized importance weights, yields an expression of the gradient that corresponds exactly to the biased SNIS approximation of (13). Thus, the optimization problem will suffer from bias. We hence propose to use BR-SNIS for learning IWAE. The proposed algorithm proceeds in two steps, which are repeated during the optimization (details are given in Appendix C.3)

- First, update the parameter ϕ as in the IWAE algorithm (using the reparameterization trick and following the methodology of [8]) according to $\phi^{(t+1)} = \phi^{(t)} - \eta \nabla_\phi \mathcal{L}_M(\theta^{(t)}, \phi^{(t)})$.
- Second, update the parameter θ by estimating (13) using BR-SNIS for $\pi(z) = p_\theta(x, z)$, $f(z) = \nabla_\theta \log p_\theta(x, z)$ and $\lambda(z) = q_\phi(z \mid x)$.

We refer to this model as BR-IWAE. As an illustration, we train the model using the binarized MNIST dataset [37], where $x \in \{0, 1\}^{784}$ are binarized digits images in dimension 784. For both for the encoder q_ϕ and the decoder p_θ, we use a simple fully connected architecture with 2 hidden layers (more details are given in Appendix C.3). For a comparison, we estimate the log-likelihood using the VAE, IWAE and BR-IWAE approaches, and the result is reported in Table 2. All models are run for 100 epochs, using the Adam optimizer [17] and a learning rate of 10^{-3}. The complete experimental details are given in Appendix C.3.

4 Conclusion

In this paper, we have introduced a novel method, BR-SNIS, which improves over SNIS when it comes to producing close to unbiased estimates of expectations taken w.r.t. to distributions known only up to a normalizing constant, a ubiquitous problem in machine learning and statistics. The high performance of BR-SNIS is supported theoretically by non-asymptotic bias, variance and high-probability bounds. We illustrate our method on various examples, which show the practical advantages of BR-SNIS over SNIS. Finally, BR-SNIS is naturally adapted to other IS based methods, for example [39], which use a Hamiltonian (gradient-based) transform [32] as part of the IS proposal. The extension of BR-SNIS to [39] would produce an Hamiltonian based sampler able to recycle all samples, contrarily to other classical Hamiltonian-based methods [32, 13]. BR-SNIS can also be extended to Particle Markov chain Monte Carlo methods such as Particle Gibbs with Ancestor sampling [26].

References

[1] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017.

[2] D. Aldous, L. Lovász, and P. Winkler. Mixing times for uniformly ergodic markov chains. *Stochastic Processes and their Applications*, 71(2):165–185, 1997.

[3] C. Andrieu. On random-and systematic-scan samplers. *Biometrika*, 103(3):719–726, 2016.

[4] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 72(3):269–342, 2010.

[5] C. Andrieu, A. Lee, and M. Vihola. Uniform ergodicity of the iterated conditional smc and geometric ergodicity of particle gibbs samplers. *Bernoulli*, 24(2):842–872, 2018.

[6] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[7] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and Radon Wasserstein Barycenters of Measures. *Journal of Mathematical Imaging and Vision*, 1(51):22–45, 2015.

[8] Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[9] J. Chen, D. Lian, B. Jin, X. Huang, K. Zheng, and E. Chen. Fast variational autoencoder with inverted multi-index for collaborative filtering. In *Proceedings of the ACM Web Conference 2022*, pages 1944–1954, 2022.

[10] R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2018.

[11] V. Elvira and L. Martino. Advances in importance sampling. *arXiv preprint arXiv:2102.05407*, 2021.

[12] T. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

[13] M. D. Hoffman, A. Gelman, et al. The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[14] D. Hsu, A. Kontorovich, D. A. Levin, Y. Peres, C. Szepesvári, and G. Wolfer. Mixing time estimation in reversible markov chains from a single sample path. *The Annals of Applied Probability*, 29(4):2439–2480, 2019.

[15] H. Kahn and A. W. Marshall. Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278, 1953.

[16] A. Karbalayghareh, X. Qian, and E. R. Dougherty. Optimal bayesian transfer learning. *IEEE Transactions on Signal Processing*, 66(14):3724–3739, 2018.

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR 2015*, 2015.

[18] D. P. Kingma and M. Welling. Stochastic gradient vb and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*, volume 19, page 121, 2014.

[19] D. P. Kroese and R. Y. Rubinstein. Monte carlo methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(1):48–58, 2012.

[20] I. Kuzborskij, C. Vernade, A. Gyorgy, and C. Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics*, pages 640–648. PMLR, 2021.

[21] R. Lamberti, Y. Petetin, F. Septier, and F. Desbouvries. A double proposal normalized importance sampling estimator. In *2018 IEEE Statistical Signal Processing Workshop (SSP)*, pages 238–242. IEEE, 2018.

[22] J. Lawson, G. Tucker, B. Dai, and R. Ranganath. Energy-inspired models: Learning with sampler-induced distributions. *Advances in Neural Information Processing Systems*, 32, 2019.

[23] A. Lee. *On auxiliary variables and many-core architectures in computational statistics*. PhD thesis, University of Oxford, 2011.

[24] A. Lee, C. Yau, M. B. Giles, A. Doucet, and C. C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *Journal of computational and graphical statistics*, 19(4):769–789, 2010.

[25] F. Lindsten, R. Douc, and E. Moulines. Uniform ergodicity of the particle gibbs sampler. *Scandinavian Journal of Statistics*, 42(3):775–797, 2015.

[26] F. Lindsten, M. I. Jordan, and T. B. Schön. Particle gibbs with ancestor sampling. 2014.

[27] J. S. Liu, W. H. Wong, and A. Kong. Covariance structure of the gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40, 1994.

[28] O. Maddouri, X. Qian, F. J. Alexander, E. R. Dougherty, and B.-J. Yoon. Robust importance sampling for error estimation in the context of optimal bayesian transfer learning. *Patterns*, page 100428, 2022.

[29] A. M. Metelli, M. Papini, F. Faccio, and M. Restelli. Policy optimization via importance sampling. *Advances in Neural Information Processing Systems*, 31, 2018.

[30] C. Naesseth, F. Lindsten, and D. Blei. Markovian score climbing: Variational inference with kl (p‖ q). *Advances in Neural Information Processing Systems*, 33:15499–15510, 2020.

[31] C. A. Naesseth, F. Lindsten, T. B. Schön, et al. Elements of sequential monte carlo. *Foundations and Trends® in Machine Learning*, 12(3):307–392, 2019.

[32] R. M. Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

[33] M. Niknejad, J. Bioucas-Dias, and M. A. Figueiredo. External patch-based image restoration using importance sampling. *IEEE Transactions on Image Processing*, 28(9):4460–4470, 2019.

[34] D. Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1–32, 2015.

[35] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[36] D. B. Rubin. Comment: A noniterative Sampling/Importance Resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, 82(398):542–543, 1987.

[37] R. Salakhutdinov and I. Murray. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th international conference on Machine learning*, pages 872–879, 2008.

[38] A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28, 2015.

[39] A. Thin, Y. Janati El Idrissi, S. Le Corff, C. Ollion, E. Moulines, A. Doucet, A. Durmus, and C. X. Robert. Neo: Non equilibrium sampling on the orbits of a deterministic transform. *Advances in Neural Information Processing Systems*, 34:17060–17071, 2021.

[40] H. Tjelmeland. Using all Metropolis–Hastings proposals to estimate mean values. Technical report, 2004.

[41] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[42] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work is of theoretical and methodological nature.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Code added to the supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Not applicable
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Not applicable
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Not applicable
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Not applicable
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Not applicable

A Proofs

A.1 i-SIR Algorithm

We analyze a slightly modified version of the i-SIR algorithm, with an extra randomization of the state position. The k-th iteration is defined as follows. Given a state $Y_k \in \mathbb{X}$,

(i) draw $I_{k+1} \in \{1, \dots, N\}$ uniformly at random and set $X_{k+1}^{I_{k+1}} = Y_k$;

(ii) draw $X_{k+1}^{1:N \setminus \{I_{k+1}\}}$ independently from the proposal distribution λ;

(iii) compute, for $i \in \{1, \dots, N\}$, the normalized importance weights

$$\omega_{N,k+1}^i = w(X_{k+1}^i) / \sum_{\ell=1}^N w(X_{k+1}^\ell);$$

(iv) select Y_{k+1} from the set $X_{k+1}^{1:N}$ by choosing X_{k+1}^i with probability $\omega_{N,k+1}^i$.

Thus, compared to the simplified i-SIR algorithm given in the introduction, the state is inserted uniformly at random into the list of candidates instead of being inserted at the first position. Of course, this change has no impact as long as we are interested in integrating functions that are permutation invariant with respect to candidates, which is the case throughout our work. Still, this randomization makes the analysis much more transparent.

A.2 Proof of Theorem 1

We write

$$\begin{aligned}
\boldsymbol{\varphi}_N(\mathrm{d}(y, x_{1:N})) &= \frac{1}{N} \sum_{i=1}^N \pi(\mathrm{d}y) \delta_y(\mathrm{d}x^i) \prod_{j \neq i} \lambda(\mathrm{d}x^j) \\
&= \frac{1}{N\lambda(w)} \sum_{i=1}^N w(x^i) \lambda(\mathrm{d}x^i) \delta_{x^i}(\mathrm{d}y) \prod_{j \neq i} \lambda(\mathrm{d}x^j) \\
&= \frac{1}{\lambda(w)} \prod_{j=1}^N \lambda(\mathrm{d}x^j) \Gamma_N 1_{\mathbb{X}}(x^{1:N}) \sum_{i=1}^N \frac{w(x^i)}{\sum_{\ell=1}^N w(x^\ell)} \delta_{x^i}(\mathrm{d}y),
\end{aligned}$$

where we recognize, and after having recalled definitions (3) and (4) of $\boldsymbol{\pi}_N$ and Π_N, respectively, the right-hand side as $\boldsymbol{\pi}_N(\mathrm{d}x^{1:N}) \Pi_N(x^{1:N}, \mathrm{d}y)$. This completes the proof.

A.3 Proof of Theorem 2

Using (4) we get

$$\begin{aligned}
\int \boldsymbol{\pi}_N(\mathrm{d}x^{1:N}) \Pi_N f(x^{1:N}) &= \int \frac{1}{N\lambda(w)} \sum_{\ell=1}^N w(x^\ell) \Pi_N f(x^{1:N}) \prod_{j=1}^N \lambda(\mathrm{d}x^j) \\
&= \frac{1}{N\lambda(w)} \int \sum_{i=1}^N w(x^i) f(x^i) \prod_{j=1}^N \lambda(\mathrm{d}x^j) = \pi(f),
\end{aligned}$$

and the proof is complete.

A.4 Proof of Theorem 5

Proof. We first check that $\boldsymbol{\varphi}_N$ is an invariant distribution for $\mathbf{P}_N$. For every $A \in \mathcal{X}^{\otimes(N+1)}$, using that π is the marginal of $\boldsymbol{\varphi}_N$ with respect to the state and applying Theorem 1 yields

$$\begin{aligned}
\int \boldsymbol{\varphi}_N(\mathrm{d}(y, x^{1:N})) \mathbf{P}_N(y, x^{1:N}, A) &= \int \pi(\mathrm{d}y) \iint \boldsymbol{\Lambda}_N(y, \mathrm{d}\bar{x}^{1:N}) \Pi_N(\bar{x}^{1:N}, \mathrm{d}\bar{y}) 1_A(\bar{y}, \bar{x}^{1:N}) \\
&= \iiint \boldsymbol{\pi}_N(\mathrm{d}\bar{x}_{1:N}) \Pi_N(\bar{x}^{1:N}, \mathrm{d}y) \Pi_N(\bar{x}^{1:N}, \mathrm{d}\bar{y}) 1_A(\bar{y}, \bar{x}^{1:N}) \\
&= \boldsymbol{\varphi}_N(A),
\end{aligned}$$

which establishes invariance. We now show that P_N is reversible with respect to π. For this purpose, let g and h be two nonnegative measurable functions and write, using Theorem 1 twice,

$$\begin{aligned}\iint \pi(\mathrm{d}y)\mathsf{P}_N(y,\mathrm{d}\bar{y})g(y)h(\bar{y}) &= \int \pi(\mathrm{d}y)\boldsymbol{\Lambda}_N(y,\mathrm{d}x^{1:N})\Pi_N(x^{1:N},\mathrm{d}\bar{y})g(y)h(\bar{y})\\ &= \int \boldsymbol{\pi}_N(\mathrm{d}x^{1:N})\Pi_N(x^{1:N},\mathrm{d}y)\Pi_N(x^{1:N},\mathrm{d}\bar{y})g(y)h(\bar{y})\\ &= \int \pi(\mathrm{d}\bar{y})\boldsymbol{\Lambda}_N(\bar{y},\mathrm{d}x^{1:N})\Pi_N(x^{1:N},\mathrm{d}y)g(y)h(\bar{y})\\ &= \iint \pi(\mathrm{d}\bar{y})\mathsf{P}_N(\bar{y},\mathrm{d}y)g(y)h(\bar{y}).\end{aligned}$$

□

A.5 Proof of Theorem 6

For completeness, we repeat the arguments in [25, 5]. Under **A**1, we have, for $(x,\mathsf{A})\in\mathbb{X}\times\mathcal{X}$,

$$\begin{aligned}\mathsf{P}_N(x,\mathsf{A}) &= \int \delta_x(\mathrm{d}x^1)\sum_{i=1}^N \frac{w(x^i)}{\sum_{j=1}^N w(x^j)}1_{\mathsf{A}}(x^i)\prod_{j=2}^N \lambda(\mathrm{d}x^j)\\ &= \int \frac{w(x)}{w(x)+\sum_{j=2}^N w(x^j)}1_{\mathsf{A}}(x)\prod_{j=2}^N \lambda(\mathrm{d}x^j) + \int \sum_{i=2}^N \frac{w(x^i)}{w(x)+\sum_{j=2}^N w(x^j)}1_{\mathsf{A}}(x^i)\prod_{j=2}^N \lambda(\mathrm{d}x^j)\\ &\geq \sum_{i=2}^N \int \frac{w(x^i)}{w(x)+w(x^i)+\sum_{j=2,j\neq i}^N w(x^j)}1_{\mathsf{A}}(x^i)\prod_{j=2}^N \lambda(\mathrm{d}x^j)\\ &\geq \sum_{i=2}^N \int \pi(\mathrm{d}x^i)1_{\mathsf{A}}(x^i)\int \frac{\lambda(w)}{w(x)+w(x^i)+\sum_{j=2,j\neq i}^N w(x^j)}\prod_{j=2,j\neq i}^N \lambda(\mathrm{d}x^j).\end{aligned}$$

Finally, since the function $f\colon z\mapsto (z+a)^{-1}$ is convex on $\mathbb{R}_+$ and $a>0$, we get for $i\in\{2,\dots,N\}$,

$$\begin{aligned}&\int \frac{\lambda(w)}{w(x)+w(x^i)+\sum_{j=2,j\neq i}^N w(x^j)}\prod_{j=2,j\neq i}^N \lambda(\mathrm{d}x^j)\\ &\geq \frac{\lambda(w)}{\int w(x)+w(x^i)+\sum_{j=2,j\neq i}^N w(x^j)\prod_{j=2,j\neq i}^N \lambda(\mathrm{d}x^j)}\\ &\geq \frac{1}{w(x)/\lambda(w)+w(x^i)/\lambda(w)+N-2} \geq \frac{1}{2\omega+N-2}.\end{aligned}$$

We finally obtain the inequality

$$\mathsf{P}_N(x,\mathsf{A}) \geq \pi(\mathsf{A})\times\frac{N-1}{2\omega+N-2} = \epsilon_N\pi(\mathsf{A}). \tag{14}$$

This means that the whole space $\mathbb{X}$ is $(1,\epsilon_N\pi)$-small (see [10, Definition 9.3.5]). Since $\mathsf{P}_N(x,\cdot)$ and π are probability measures, (14) implies

$$\|\mathsf{P}_N(x,\cdot)-\pi\|_{\mathrm{TV}} = \sup_{\mathsf{A}\in\mathcal{X}}|\mathsf{P}_N(x,\mathsf{A})-\pi(\mathsf{A})| \leq 1-\epsilon_N = \kappa_N.$$

Now the statement follows from [10, Theorem 18.2.4] applied with $m=1$.

A.6 Proof of Theorem 3

Proof of ((ii)). Considering the identity $(a+b)^2 \leq (1+\epsilon^2)a^2+(1+\epsilon^{-2})b^2$, we obtain the decomposition $\{\Pi_N f(X_k^{1:N})-\pi(f)\}^2 \leq (1+(N-1)^{-1/2})I+(1+(N-1)^{1/2})II$, with

$$\begin{aligned}I &:= \{\Pi_N f(X_k^{1:N}) - a_N(Y_{k-1})/b_N(Y_{k-1})\}^2\\ II &:= \{a_N(Y_{k-1})/b_N(Y_{k-1})-\pi(f)\}^2.\end{aligned}$$

By using the identity $a/b - c/d = (1/d)[(a/b)(d-b) - (c-a)]$, we obtain

$$\Pi_N f(X_k^{1:N}) - a_N(Y_{k-1})/b_N(Y_{k-1}) = b_N(Y_{k-1})^{-1}[\Pi_N f(X_k^{1:N})(b_N(Y_{k-1}) - \Gamma_N 1_{\mathbb{X}}(X_k^{1:N})) \\ - (a_N(Y_{k-1}) - \Gamma_N f(X_k^{1:N}))].$$

Therefore, using $(a+b)^2 \leq 2(a^2+b^2)$ we get

$$I \leq \frac{2}{b_N(Y_{k-1})^2}[\Pi_N f(X_k^{1:N})^2\{\Gamma_N 1_{\mathbb{X}}(X_k^{1:N}) - b_N(Y_{k-1})\}^2 + \{\Gamma_N f(X_k^{1:N}) - a_N(Y_{k-1})\}^2].$$

Using the fact that $\Pi_N f(X_k^{1:N})^2 \leq 1$ $\mathbb{P}_{\boldsymbol{\xi}}$-a.s. and $b_N(y) \geq (N-1)/N\lambda(w)$, we get, $\mathbb{P}_{\boldsymbol{\xi}}$-a.s.,

$$I \leq \frac{2N^2}{(N-1)^2\lambda(w)^2}\left[\{\Gamma_N 1_{\mathbb{X}}(X_k^{1:N}) - b_N(Y_{k-1})\}^2 + \{\Gamma_N f(X_k^{1:N}) - a_N(Y_{k-1})\}^2\right]$$

Therefore, using Lemma (7),

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}}[\{\Pi_N f(X_k^{1:N}) - a_N(Y_{k-1})/b_N(Y_{k-1})\}^2] \\
&= \mathbb{E}_{\boldsymbol{\xi}}[\mathbb{E}_{\boldsymbol{\xi}}\left[\{\Pi_N f(X_k^{1:N}) - a_N(Y_{k-1})/b_N(Y_{k-1})\}^2 \,\middle|\, Y_{k-1}\right]] \\
&\leq \frac{2N^2}{(N-1)^2\lambda(w)^2}\left[(N-1)/N^2\lambda(\{w-\lambda(w)\}^2) + (N-1)/N^2\lambda(\{wf-\lambda(wf)\}^2)\right] \\
&\leq 4(N-1)^{-1}\kappa[\pi,\lambda].
\end{aligned}$$

We consider now II. From (12) we have that $II \leq 4N^{-2}(1+\omega)^2$, which completes the proof. □

Proof of ((iii)). Note that

$$\begin{aligned}
III &:= \mathbb{E}_{\boldsymbol{\xi}}[\{\Pi_N f(X_k^{1:N}) - \pi(f)\}\{\Pi_N f(X_{k+\ell}^{1:N}) - \pi(f)\}] \\
&= \mathbb{E}_{\boldsymbol{\xi}}[\{\Pi_N f(X_k^{1:N}) - \pi(f)\}\mathbb{E}_{\boldsymbol{\xi}}\left[\Pi_N f(X_{k+\ell}^{1:N}) - \pi(f) \,\middle|\, Y_{k+\ell-1}\right]].
\end{aligned}$$

As $\mathbb{E}_{\boldsymbol{\xi}}\left[\Pi_N f(X_{k+\ell}^{1:N}) \,\middle|\, Y_{k+\ell-1}\right] = \Phi_N(Y_{k+\ell-1})$ $\mathbb{P}_{\boldsymbol{\xi}}$-a.s., we have that

$$\begin{aligned}
III &= \mathbb{E}_{\boldsymbol{\xi}}[\{\Pi_N f(X_k^{1:N}) - \pi(f)\}\{\Phi_N(Y_{k+\ell-1}) - \pi(f)\}] \\
&= \mathbb{E}_{\boldsymbol{\xi}}[\{\Pi_N f(X_k^{1:N}) - \pi(f)\}\{\mathbb{E}_{\boldsymbol{\xi}}\left[\Phi_N(Y_{k+\ell-1}) \,|\, Y_k\right] - \pi(f)\}].
\end{aligned}$$

By the Markov property, we get that

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\Phi_N(Y_{k+\ell-1}) \,|\, Y_k\right] = \mathsf{P}_N^{\ell-1}\Phi_N(Y_k) = \delta_{Y_k}\mathsf{P}_N^{\ell-1}\Phi_N, \quad \mathbb{P}_{\boldsymbol{\xi}}\text{-a.s.},$$

which, combined with (10), implies that

$$\|\mathsf{P}_N^{\ell-1}\Phi_N - \pi(f)\|_\infty \leq \varsigma^{bias}(N-1)^{-1}\kappa_N^{\ell-1}.$$

Combining the results above, we finally establish that

$$\begin{aligned}
|III| &\leq \varsigma^{bias}(N-1)^{-1}\kappa_N^{\ell-1}\mathbb{E}_{\boldsymbol{\xi}}[\{\Pi_N f(X_k^{1:N}) - \pi(f)\}^2]^{1/2} \\
&\leq \varsigma^{bias}(N-1)^{-1}\kappa_N^{\ell-1}\left\{\sum\nolimits_{i=0}^{2}\varsigma_i^{mse}(N-1)^{-1-i/2}\right\}^{1/2}.
\end{aligned}$$

□

A.7 Proof of Theorem 4

We first consider the bias term. The bound is given by

$$\begin{aligned}
\left|\mathbb{E}_{\boldsymbol{\xi}}[\Pi_{(K_0,K),N}(f)] - \pi(f)\right| &\leq (K-K_0)^{-1}\sum_{\ell=K_0+1}^{K}\left|\mathbb{E}_{\boldsymbol{\xi}}[\Pi_N f(X_\ell^{1:N})] - \pi(f)\right| \\
&\leq (K-K_0)^{-1}(N-1)^{-1}\varsigma^{bias}\sum\nolimits_{\ell=K_0+1}^{K}\kappa_N^{\ell-1}.
\end{aligned}$$

The proof is concluded by noting that

$$\sum\nolimits_{\ell=K_0+1}^{K} \kappa_N^{l-1} \leq \frac{\kappa_N^{K_0}}{1-\kappa_N} \leq \frac{4\tau_{mix,N}(1/4)^{K_0/\tau_{mix,N}}}{3}.$$

We now consider the MSE, using the decomposition

$$\mathbb{E}_{\boldsymbol{\xi}}[(\Pi_{(K_0,K),N}(f)-\pi(f))^2] \leq (K-K_0)^{-2}\left\{\sum\nolimits_{\ell=K_0+1}^{K} \mathbb{E}_{\boldsymbol{\xi}}[\Pi_N f(X_\ell^{1:N})]-\pi(f)\right\}^2$$
$$+2\sum\nolimits_{\ell=K_0+1}^{K}\sum\nolimits_{j=\ell+1}^{K} \mathbb{E}_{\boldsymbol{\xi}}[\{\Pi_N f(X_\ell^{1:N})-\pi(f)\}\{\Pi_N f(X_j^{1:N})-\pi(f)\}].$$

From the MSE bound of Theorem 11, we have that

$$\sum\nolimits_{\ell=K_0+1}^{K} \mathbb{E}_{\boldsymbol{\xi}}[\{\Pi_N f(X_\ell^{1:N})-\pi(f)\}^2] \leq (K-K_0)(N-1)^{-1}\sum\nolimits_{i=0}^{2} \varsigma_i^{mse}(N-1)^{-i/2}.$$

From the covariance bound of Theorem 11, we have that

$$\sum_{\ell=K_0+1}^{K}\sum_{j=\ell+1}^{K} \mathbb{E}_{\boldsymbol{\xi}}[\{\Pi_N f(X_\ell^{1:N})-\pi(f)\}\{\Pi_N f(X_j^{1:N})-\pi(f)\}]$$
$$\leq \sum_{i=0}^{2} \varsigma_i^{cov}(N-1)^{-\frac{3-i/2}{2}}\left[\sum_{\ell=K_0+1}^{K}\sum_{j=\ell+1}^{K} \kappa_N^{(j-\ell)-1}\right].$$

As $\sum_{\ell=K_0+1}^{K}\sum_{j=\ell+1}^{K} \kappa_N^{(j-\ell)-1} \leq (K-K_0)(4/3)\tau_{mix,N}$ we have

$$\mathbb{E}_{\boldsymbol{\xi}}[(\Pi_{(K_0,K),N}(f)-\pi(f))^2] \leq ((K-K_0)(N-1))^{-1}\left[\sum\nolimits_{i=0}^{2} \varsigma_i^{mse}(N-1)^{-i/2}\right]$$
$$+(8/3)(K-K_0)^{-1}(N-1)^{-3/2}\left[\sum_{i=0}^{2} \varsigma_i^{cov}(N-1)^{-i/4}\right].$$

The proof for the MSE is concluded by noting that $(K-K_0)(N-1) = \upsilon M$ and noting that $\mathsf{MSE}_M^{is} := \varsigma_0^{mse} M^{-1}$.

The high-probability bound requires more complex derivations. We use the decomposition

$$\Pi_{(K_0,K),N}(f)-\pi(f) = (K-K_0)^{-1}\sum_{k=K_0+1}^{K} \Pi_N f(X_k^{1:N})-\Phi_N(Y_{k-1})$$
$$+(K-K_0)^{-1}\sum_{k=K_0+1}^{K-1} \Phi_N(Y_{k-1})-\pi(\Phi_N).$$

where we have used $\pi(f)=\pi(\Phi_N)$. Therefore, for any $t \geq 0$, we get that

$$\mathbb{P}_{\boldsymbol{\xi}}(|\Pi_{(K_0,K),N}(f)-\pi(f)| \geq t) \leq \mathbb{P}_{\boldsymbol{\xi}}\left((K-K_0)^{-1}\left|\sum\nolimits_{k=K_0+1}^{K} \Pi_N f(X_k^{1:N})-\Phi_N(Y_{k-1})\right| \geq t/2\right)$$
$$+\mathbb{P}_{\boldsymbol{\xi}}\left((K-K_0)^{-1}\left|\sum\nolimits_{k=K_0+1}^{K-1} \Phi_N(Y_{k-1})-\pi(\Phi_N)\right| \geq t/2\right).$$

We will show that for all $t>0$, and for some absolute constants ζ_I, ζ_{II},

$$I := \mathbb{P}_{\boldsymbol{\xi}}\left((K-K_0)^{-1}\left|\sum\nolimits_{k=K_0+1}^{K} \Pi_N f(X_k^{1:N})-\Phi_N(Y_{k-1})\right| \geq t\right) \leq 2\exp(-t^2\upsilon M/(4\zeta_I)),$$
$$II := \mathbb{P}_{\boldsymbol{\xi}}\left((K-K_0)^{-1}\left|\sum\nolimits_{k=K_0+1}^{K-1} \Phi_N(Y_{k-1})-\pi(\Phi_N)\right| \geq t\right) \leq 2\exp(-t^2\zeta_{II}(K-K_0)(N-1)^2/\tau_{mix,N}).$$

Consider first I. Note first that

$$I = \mathbb{E}_{\boldsymbol{\xi}}\left[\mathbb{P}_{\boldsymbol{\xi}}\left[(K-K_0)^{-1}\left|\sum\nolimits_{k=K_0+1}^{K}\Pi_N f(X_k^{1:N}) - \Phi_N(Y_{k-1})\right| \geq t \,\middle|\, Y_{K_0:K-1}\right]\right].$$

By Theorem 5, the random elements $\{X_k^{1:N}\}_{k=K_0+1}^{K}$ are independent conditionally to $\{Y_k\}_{k=K_0}^{K-1}$. Using the generalized Hoeffding inequality (see [41, Theorem 2.6.2] or [42, Proposition 2.1]), we get, with $\Delta_{N,k} := \Pi_N f(X_k^{1:N}) - \Phi_N(Y_{k-1})$, that

$$\mathbb{P}_{\boldsymbol{\xi}}\left[(K-K_0)^{-1}\left|\sum\nolimits_{k=K_0+1}^{K}\Delta_{N,k}\right| \geq t \,\middle|\, Y_{K_0:K-1}\right] \leq 2\exp\left(-\frac{t^2(K-K_0)^2}{4\sum_{k=K_0+1}^{K}\|\Delta_{N,k}\|_{\psi_2,Y_k}^2}\right),$$

where $\psi_2(x) = \exp(x^2) - 1$ and

$$\|\Delta_{N,k}\|_{\psi_2,Y_{k-1}} = \inf\left\{\lambda > 0 \,:\, \mathbb{E}_{\boldsymbol{\xi}}\left[\psi_2(|\Delta_{N,k}|/\lambda) \,|\, Y_{k-1}\right] \leq 1\right\}.$$

We have to bound $\|\Delta_{N,k}\|_{\psi_2,Y_{k-1}}$. We use the decomposition

$$\Delta_{N,k} = \left\{\frac{\Gamma_N f(X_k^{1:N})}{\Gamma_N 1_{\mathbb{X}}(X_k^{1:N})} - \frac{a_N(Y_{k-1})}{b_N(Y_{k-1})}\right\} + \left\{\frac{a_N(Y_{k-1})}{b_N(Y_{k-1})} - \Phi_N(Y_{k-1})\right\} =: \Delta_{N,k}^{I} + \Delta_{N,k}^{II}.$$

combined with Lemma 9 with $\phi = \psi_2$ and $\chi = \psi_2$ and [41, Proposition 2.6.1]. By (11) and by [41, Equation 2.17] we have that

$$\|\Delta_{N,k}^{II}\|_{\psi_2,Y_{k-1}} \leq 2\log(2)^{-1/2}(N-1)^{-1}\kappa[\lambda,\pi].$$

Using Lemma 9 with $\phi = \chi = \psi_2$ and the fact that $b_N(y) \geq (1-1/N)\lambda(w)$ we have that

$$\begin{aligned}\|\Delta_{N,k}^{I}\|_{\psi_2,Y_{k-1}} &\leq \frac{2}{(1-1/N)\lambda(w)} \\ &\quad\times\left[\|\Gamma_N f(X_k^{1:N}) - a_N(Y_{k-1})\|_{\psi_2,Y_{k-1}} + 2\|\Gamma_N 1_{\mathbb{X}}(X_k^{1:N}) - b_N(Y_{k-1})\|_{\psi_2,Y_{k-1}}\right].\end{aligned}$$

Using [41, Proposition 2.6.1, Eq 2.17], we get that, $\mathbb{P}_{\boldsymbol{\xi}}$-a.s.,

$$\begin{aligned}\|\Gamma_N f(X_{k-1}^{1:N}) - a_N(Y_{k-1})\|_{\psi_2,Y_{k-1}}^2 &\leq 64\mathrm{e}/\log 2N^{-1}\|w(X_k^1)f(X_k^1) - \mathbb{E}_{\boldsymbol{\xi}}\left[w(X_k^1)f(X_k^1)\,\middle|\,Y_{k-1}\right]\|_{\psi_2,Y_{k-1}}^2, \\ &\leq 256\mathrm{e}/\log(2)^2 N^{-1}\|w\|_\infty^2.\end{aligned}$$

The same bound applies to $\|\Gamma_N 1_{\mathbb{X}}(X_k^{1:N}) - b_N(Y_{k-1})\|_{\psi_2,Y_{k-1}}^2$ and we can write

$$\|\Delta_{N,k}^{I}\|_{\psi_2,Y_{k-1}} \leq 96e^{1/2}\log(2)^{-1}(N-1)^{-1/2}\omega.$$

We can now conclude by writing

$$\begin{aligned}\|\Delta_{N,k}\|_{\psi_2,Y_{k-1}}^2 &\leq 2(\|\Delta_{N,k}^{I}\|_{\psi_2,Y_{k-1}}^2 + \|\Delta_{N,k}^{II}\|_{\psi_2,Y_{k-1}}^2) \\ &\leq (N-1)^{-1}(\zeta_{I,1}\omega^2 + \zeta_{I,2}\kappa[\lambda,\pi]^2(N-1)^{-1}),\end{aligned}$$

where $\zeta_{I,1} = 18432\mathrm{e}\log(2)^{-2}$ and $\zeta_{I,2} = 8\log(2)^{-1}$ are absolute constants, which imply

$$\|\Delta_{N,k}\|_{\psi_2,Y_{k-1}}^2 \leq \zeta_I(N-1)^{-1},$$

with $\zeta_I = 1.1\cdot 10^5\omega^2$. This finally concludes $I \leq 2\exp(-t^2\upsilon M/4\zeta_I)$.

Consider now II. We use Lemma 12 with $g_i = \Phi_N(Y_{K_0+i-1}) - \pi(\Phi_N)$. As $\|g_i\|_\infty \leq \mathrm{osc}(\Phi_N) \leq (N-1)^{-1}\varsigma^{bias}$, we obtain

$$II \leq 2\exp\left(-t^2\zeta_{II}(K-K_0)(N-1)^2/\tau_{mix,N}\right)$$

where $\zeta_{II} = \frac{2}{(3\varsigma^{bias})^2}$. Finally, we obtain

$$\mathbb{P}_{\boldsymbol{\xi}}(|\Pi_{(K_0,K),N}(f) - \pi(f)| \geq t) \leq 2\exp(-t^2\upsilon M/4\zeta_I)[1 + \exp(-t^2\upsilon M\{\zeta_{II}(N-1)/\tau_{mix,N} - (4\zeta_I)^{-1}\})].$$

We conclude by noting that, for any $\delta \in (0,1)$, if $(N-1) \geq \tau_{mix,N}(4\zeta_I\zeta_{II})^{-1}$, we have that

$$\mathbb{P}_{\boldsymbol{\xi}}(|\Pi_{(K_0,K),N}(f) - \pi(f)| \geq t) \leq 4\exp(-t^2\upsilon M/4\zeta_I) \leq \delta$$

for all $t \geq 2\zeta_I^{1/2}(\upsilon M)^{-1/2}\log(4/\delta)^{1/2}$. This concludes the proof with $\varsigma^{hpd} = 2\zeta_I^{1/2}$.

A.8 High probability inequality for SNIS

Theorem 8. *Assume that $\omega = \|w\|_\infty/\lambda(w) < \infty$. For any bounded measurable function f on $(\mathbb{X}, \mathcal{X})$ such that $\|f\|_\infty \leq 1$, for any $M > 1$ and for any δ in $(0,1)$,*

$$|\widehat{\pi}_M(f) - \pi(f)| \leq 12\omega[M\log(2)]^{-1/2}\log(2/\delta)^{1/2}$$

with probability larger than $1-\delta$.

Proof. Let $A_M = M^{-1}\sum_{i=1}^M w(X^i)f(X^i)$, $B_M = M^{-1}\sum_{i=1}^M w(X^i)$, $a = \mathbb{E}(A_M) = \lambda(wf)$ and $b = \mathbb{E}(B_M) = \lambda(w)$. Note that $\widehat{\pi}_M(f) = A_M/B_M$ and $\pi(f) = a/b$. By Lemma 9 with $\phi = \chi = \exp(x^2) - 1$, we have that:

$$\|\widehat{\pi}_M(f) - \pi(f)\|_{\psi_2} \leq 2\lambda(w)^{-1}[\|A_M - a\|_{\psi_2} + 2\|B_M - b\|_{\psi_2}].$$

Using [41, Eq 2.17], we get that, $\mathbb{P}_{\boldsymbol{\xi}}$-a.s.

$$\|A_M - a\|_{\psi_2}^2 \leq M^{-1}\|w(X^i)f(X^i) - \lambda(wf)\|_{\psi_2}^2 \leq 4(\log(2)M)^{-1}\|w\|_\infty^2.$$

In the same way, $\|B_M - b\|_{\psi_2}^2 \leq 4(\log(2)M)^{-1}\|w\|_\infty^2$. Therefore, we have that:

$$\|\widehat{\pi}_M(f) - \pi(f)\|_{\psi_2}^2 \leq (12\omega)^2(\log(2)M)^{-1}.$$

Combining it with [41, Proposition 2.5.2], we have that:

$$\mathbb{P}(|\widehat{\pi}_M(f) - \pi(f)| \geq t) \leq 2\exp(-t^2\zeta^{\text{snis}}M)$$

where $\zeta^{\text{snis}} = \log(2)(12\omega)^{-2}$. The high probability inequality follows directly. □

B Moments and high-probability bounds for ratio statistics

Let $(U_i, V_i)_{i\in\{1,\dots,n\}}$ be (possibly dependent) random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that $U_i \geq 0$, $\mathbb{P}$-a.s.. Moreover, let $\hat{A}_n = n^{-1}\sum_{i=1}^n U_iV_i$, $\hat{B}_n = n^{-1}\sum_{i=1}^n U_i$, $\hat{R}_n = \hat{A}_n/\hat{B}_n$ and $a = \mathbb{E}\left[\hat{A}_n\right]$, $b = \mathbb{E}\left[\hat{B}_n\right]$, and $r = a/b$.

A continuous, even, convex function $\phi : \mathbb{R}^+ \to [0, +\infty]$ is a Young function if ϕ is monotonically increasing for $x > 0$, $\phi(0) = 0$, $\lim_{x\to\infty}\phi(x)/x = \infty$, and $\lim_{x\to 0^+}\phi(x)/x = 0$. We denote by ϕ^* the Fenchel-Legendre conjugate of ϕ. Let X be a random variable and ϕ be a Young function. Then the *Orlicz norm* of X is

$$\|X\|_\phi = \inf\{\lambda > 0 \,:\, \mathbb{E}[\phi(|X|/\lambda)] \leq 1\}\,,$$

with the convention that $\inf\emptyset = \infty$. The Orlicz space of random variable $\mathcal{L}_\phi(\Omega)$ is the family of equivalence classes of random variables X such that $\|X\|_\phi < \infty$. $\mathcal{L}_\phi(\Omega)$ is a Banach space. If $\phi_p(x) = |x|^p$ for $p \geq 1$, then $\mathcal{L}_\phi(\Omega) = \mathcal{L}^p(\Omega)$ and we denote $\|\cdot\|_p = \|\cdot\|_{\phi_p}$. If $X \in \mathcal{L}_\phi(\Omega)$, then, for any $x > 0$

$$\mathbb{P}(|X| \geq x) \leq 1/\phi(x/\|X\|_\phi), \quad \text{and} \quad \|1_{\{|X|\geq x\}}\|_\phi = 1/\phi^{-1}(1/\mathbb{P}(|X| \geq x)).$$

Lemma 9. *Let ϕ, χ be Young functions. If $\max_i \|V_i\|_\infty \leq c|r|$, then*

$$\|\hat{R}_n - r\|_\phi/|r| \leq 2\|\hat{A}_n - a\|_\phi/b + 2\|\hat{B}_n - b\|_\phi/b + c/\{(\phi^{-1}\circ\chi)(b/2\|(\hat{B}_n - b)_-\|_\chi)\}\,.$$

Proof. We decompose the computation in two parts: first, when $\hat{B}_n > b/2$, we have

$$\begin{aligned}|\hat{R}_n - r| &= \left|\frac{\hat{A}_n - a}{\hat{B}_n} + a\left(\frac{1}{\hat{B}_n} - \frac{1}{b}\right)\right| \leq \frac{|\hat{A}_n - a|}{\hat{B}_n} + \frac{|a||\hat{B}_n - b|}{\hat{B}_n b} \\ &\leq \frac{|\hat{A}_n - a|}{b/2} + \frac{|a||\hat{B}_n - b|}{(b/2)b} = \frac{2|\hat{A}_n - a|}{b} + \frac{2|r||\hat{B}_n - b|}{b}\,.\end{aligned}$$

Then, when $\hat{B}_n \le b/2$, we have

$$|\hat{R}_n - r| \le |\hat{R}_n| + |r| \le |\hat{R}_n| + \frac{2|r||\hat{B}_n - b|}{b} \le \max_i |V_i| + \frac{2|r||\hat{B}_n - b|}{b}\,,$$

where the second inequality follows from $|\hat{B}_n - b| \ge b/2$. Combining the two previous inequalities, we have

$$|\hat{R}_n - r| \le \frac{2|\hat{A}_n - a|}{b} + \frac{2|r||\hat{B}_n - b|}{b} + \max_i |V_i| 1_{\{\hat{B}_n \le b/2\}}\,.$$

Recall that, if $|X| \le |Y|$ a.s., then $\|X\|_\phi \le \|Y\|_\phi$. Hence, we get that

$$\begin{aligned}
\|\hat{R}_n - r\|_\phi &\le \left\| \frac{2|\hat{A}_n - a|}{b} + \frac{2r|\hat{B}_n - b|}{b} + \max_i |V_i| 1_{\{\hat{B}_n \le b/2\}} \right\|_\phi \\
&\le \left\| \frac{2\hat{A}_n - a}{b} \right\|_\phi + \left\| \frac{2|r||\hat{B}_n - b|}{b} \right\|_\phi + \left\| \max_i |V_i| 1_{\{\hat{B}_n \le b/2\}} \right\|_\phi \\
&= \frac{2\|\hat{A}_n - a\|_\phi}{b} + \frac{2|r|\|\hat{B}_n - b\|_\phi}{b} + \| \max_i |V_i| 1_{\{\hat{B}_n \le b/2\}} \|_\phi \\
&\le \frac{2\|\hat{A}_n - a\|_\phi}{b} + \frac{2|r|\|\hat{B}_n - b\|_\phi}{b} + c|r| \| 1_{\{\hat{B}_n \le b/2\}} \|_\phi \\
&= \frac{2\|\hat{A}_n - a\|_\phi}{b} + \frac{2|r|\|\hat{B}_n - b\|_\phi}{b} + c|r|/\phi^{-1}\left(1/\mathbb{P}(\hat{B}_n \le b/2)\right)\,.
\end{aligned}$$

Finally, we obtain the desired result by noting that, for any Young function χ, $\mathbb{P}(\hat{B}_n \le b/2) = \mathbb{P}(|(\hat{B}_n - b)_-| \ge b/2) \le 1/\chi(b/2\|(\hat{B}_n - b)_-\|_\chi)$. □

Theorem 10. *Let $p \ge 1$. If $\max_i \|V_i\|_\infty \le c|r|$, then*

$$\frac{\|\hat{R}_n - r\|_p}{|r|} \le \frac{2\|\hat{A}_n - a\|_p}{b} + \frac{2(1+c)\|\hat{B}_n - b\|_p}{b}\,.$$

Proof. Apply Lemma 9 with $\chi(x) = \phi(x) = x^p$. □

Theorem 11 (Bias of the estimator)**.** *If $|\hat{A}_n/\hat{B}_n| \le 1$, then*

$$\left|\mathbb{E}[\hat{R}_n] - r\right| \le (2b^2)^{-1}\{3\mathbb{E}[\{\hat{B}_n - b\}^2] + \mathbb{E}[\{\hat{A}_n - a\}^2]\}.$$

Proof. We use the elementary identity

$$\frac{\hat{A}_n}{\hat{B}_n} - \frac{a}{b} = \frac{\hat{A}_n}{\hat{B}_n}\frac{(b - \hat{B}_n)^2}{b^2} + \frac{(\hat{A}_n - a)(b - \hat{B}_n)}{b^2} + \frac{a(b - \hat{B}_n)}{b^2} + \frac{\hat{A}_n - a}{b},$$

which implies that

$$\mathbb{E}[\hat{R}_n] - r = \mathbb{E}\left[\frac{\hat{A}_n}{\hat{B}_n}\frac{(b - \hat{B}_n)^2}{b^2}\right] + \frac{\mathbb{E}[(\hat{A}_n - a)(b - \hat{B}_n)]}{b^2}$$

□

We conclude with a lemma that gives the concentration of a uniformly ergodic Markov chain. We think that this Lemma is of independent interest, and we give it under general conditions.

Lemma 12. *Let $(\mathbb{Z}, \mathcal{Z})$ be a state-space and Q a Markov kernel on $(\mathbb{Z}, \mathcal{Z})$ which is uniformly ergodic with mixing time t_{mix} and stationary distribution π. Let $\{g_i\}_{i=1}^n$ be a family of measurable functions from $\mathbb{Z}$ to $\mathbb{R}^d$ such that $\|g\|_\infty = \max_{i\in\{1,\dots,n\}} \|g_i\|_\infty < \infty$ and $\pi(g_i) = 0$ for any $i \in \{1, \dots, n\}$. Then, for any initial probability ξ on $(\mathbb{Z}, \mathcal{Z})$, $n \in \mathbb{N}$, $t \ge 0$, it holds*

$$\mathbb{P}_\xi\left(\|\sum_{i=1}^n g_i(Z_i)\| \ge t\right) \le 2\exp\left\{-\frac{2t^2}{u_n^2}\right\},\ \textit{where } u_n = 3\|g\|_\infty\sqrt{n}\sqrt{t_{\mathrm{mix}}}. \tag{15}$$

Proof. The function $\varphi(x_1^{1:N}, \dots, x_n^{1:N}) := \| \sum_{i=1}^n g_i(x_i^{1:N}) \|$ on $\mathbb{Z}^n$ satisfies the bounded differences property. Applying [34, Corollary 2.10], we get for $t \geq \mathbb{E}_\xi[\| \sum_{i=1}^n g_i(Z_i) \|]$,

$$\mathbb{P}_\xi\left(\| \sum_{i=1}^n g_i(Z_i) \| \geq t \right) \leq \exp\left\{ -\frac{2(t - \mathbb{E}_\xi[\| \sum_{i=1}^n g_i(Z_i) \|])^2}{9n \|g\|_\infty^2 t_{\mathrm{mix}}} \right\}.$$

It remains to upper bound $\mathbb{E}_\xi[\| \sum_{i=1}^n g_i(Z_i) \|]$. Note that

$$\mathbb{E}_\xi[\| \sum_{i=1}^n g_i(Z_i) \|^2] = \sum_{i=1}^n \mathbb{E}_\xi[\| g_i(Z_i) \|^2] + 2 \sum_{k=1}^{n-1} \sum_{\ell=1}^{n-k} \mathbb{E}_\xi[g_k(Z_k)^\top g_{k+\ell}(Z_{k+\ell})]$$

and, using $\pi(g_{k+\ell}) = 0$, we obtain

$$\left| \mathbb{E}_\xi[g_k(Z_k)^\top g_{k+\ell}(Z_{k+\ell})] \right| = \left| \int_{\mathbb{Z}} g_k(z)^\top \left(\mathrm{Q}^\ell g_{k+\ell}(z) - \pi(g_{k+\ell}) \right) \xi \, \mathrm{Q}^k(\mathrm{d}z) \right| \leq \|g\|_\infty^2 (1/4)^{\lceil \ell / t_{\mathrm{mix}} \rceil}.$$

which implies

$$\sum_{k=1}^{n-1} \sum_{\ell=1}^{n-k} |\mathbb{E}_\xi[g_k(Z_k)^\top g_{k+\ell}(Z_{k+\ell})]| \leq \sum_{k=1}^{n-1} \|g\|_\infty^2 (1/4)^{\lceil \ell / t_{\mathrm{mix}} \rceil} \leq (4/3) \|g\|_\infty^2 t_{\mathrm{mix}} n.$$

Combining the bounds above, we upper bound $\mathbb{E}_\xi[\| \sum_{i=1}^n g_i(Z_i) \|]$ as

$$\mathbb{E}_\xi[\| \sum_{i=1}^n g_i(Z_i) \|] \leq \left\{ \mathbb{E}_\xi[\| \sum_{i=1}^n g_i(Z_i) \|^2] \right\}^{1/2} \leq 2\sqrt{n} \|g\|_\infty \sqrt{t_{\mathrm{mix}}} =: v_n.$$

Plugging this result in (15), we obtain that

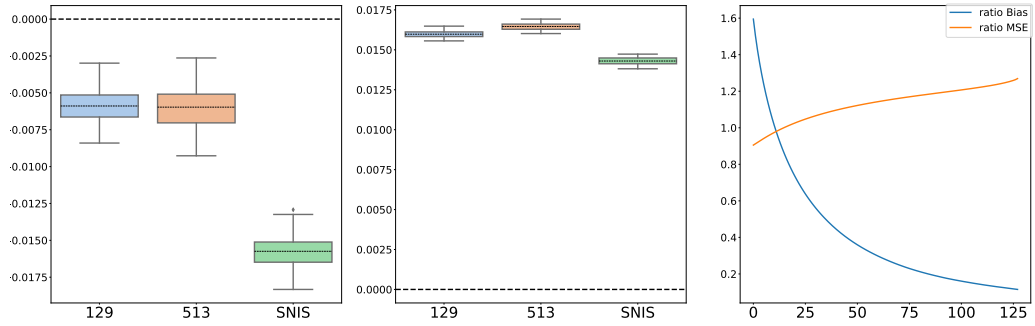
$$\mathbb{P}_\xi\left(\| \sum_{i=1}^n g_i(Z_i) \| \geq t \right) \leq \begin{cases} 1, & t < v_n, \\ \exp\left\{ -\frac{2(t - v_n)^2}{9 v_n^2} \right\}, & t \geq v_n. \end{cases} \tag{16}$$

Now it is easy to see that right-hand side of (16) is upper bounded by $2 \exp\{-2t^2/(9v_n^2)\}$ for any $t \geq 0$, and the statement follows. □

C Experiments

C.1 Toy Problem

Bias MSE trade-off: We display in Figures 4a and 4b the bias and the MSE of the BR-SNIS estimators for the same configuration as in Figure 2 but with $k_0 = \lfloor 0.625 k_{max} \rfloor$. We observe 3 times less bias than the SNIS estimators but only with a 10% increase of the MSE for the $N = 129$ setting. This can be also seen in Figure 4c, where we show the ratio between BR-SNIS and SNIS for bias and MSE with $N = 129$.



(a) Bias (b) MSE (c) Ratios as k_0

Figure 4: Comparison between SNIS and BR-SNIS for the same budget. In each boxplot the dotted line represents the **mean** value of the samples. In Figure 4c we display the ratio between BR-SNIS and SNIS for bias and MSE with $N = 129$.

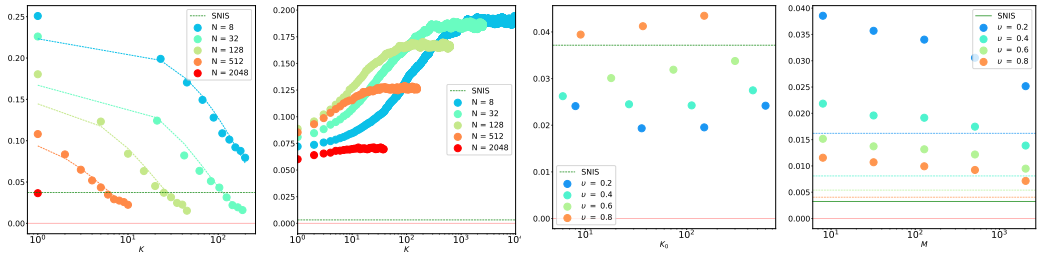
Parameters Gaussian mixture: The π in Section 3 is a Mixture of two Gaussians in dimension 7 with mean vectors $\boldsymbol{\mu}_1 = (1, \dots, 1)^\intercal$ and $\boldsymbol{\mu}_2 = (-2, 0, \dots, 0)^\intercal$ and covariance matrices $\boldsymbol{\Sigma}_1 = d^{-1}\mathbf{I}$ and $\boldsymbol{\Sigma}_2 = d^{-1}\mathbf{I}$, where $p = 1/3$ and $\mathbf{I}$ is the identity matrix In this setting, the quantities $\kappa[\pi, \lambda]$ and ω can be estimated by Monte Carlo and Gradient ascent respectively. Their values are approximately $7 \cdot 10^2$ and $1 \cdot 10^4$, respectively.

The sets A and B used to define the function f are the following:

$$A := [-2, 6] \times [-1, 1]^6, \quad B := [0.75, 1.25] \times [1, 2] \times [-0.1, 0.1]^5.$$

We used this example to illustrate numerically the bounds in Theorems 3 and 4, where each expectation was calculated by Monte Carlo using $2 \cdot 10^4$ samples. We displayed in each figure the equivalent SNIS estimation in a green dashed line. For all the bias related bounds(Theorem 3(i) in Figure 5a, Theorem 4(i) in Figure 5c), we fixed a total budget of $M = 6 \cdot 10^3$. For Figure 5a we added a fit of the type $y = \exp(ak + b)$ to illustrate the exponential decay w.r.t. k.

We then increased the budget to $M = 8 \cdot 10^4$ for the MSE and covariance bounds, in order to fully observe the stabilisation of the MSE in Figure 5b for all the minibatch sizes N. For the true value of $\pi(f)$ needed for calculating the MSE, we use an estimation obtained by Monte Carlo (sampling directly from π) with $4 \cdot 10^7$ samples. In Figure 5d we added dashed lines with the theoretical value of the $\mathsf{MSE}^{is}_{\upsilon M}$ with the same color as υ.



(a) Bias (b) MSE (c) Rolling bias (d) Rolling MSE

Figure 5: Visualization of the theoretical bounds from Theorems 3 and 4.

C.2 Bayesian Logistic regression

The importance distribution used in the Bayesian logistic regression example is given by the mean-field variational distribution [6]. More precisely, given the target π given in Section 3, the proposal λ is a Gaussian distribution with mean μ and diagonal covariance $\mathrm{diag}(\sigma)$, where μ, σ are learnt by maximization of the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\mu, \sigma) = \int \log(\pi(\theta)/\lambda(\theta))\lambda(\theta)\mathrm{d}\theta.$$

In both Figures 3 and 6, the optimal k for a given budget M was chosen by grid search over all the factors of M. The final settings are shown in Table 3.

C.3 Importance Weighted Auto-Encoders

The details of training procedure are given in Table 4. The train ELBO for each latent dimension is shown in Figure 9. For each network we display generated images in Figures 7 and 8. The generation is done by sampling from the prior p_θ.

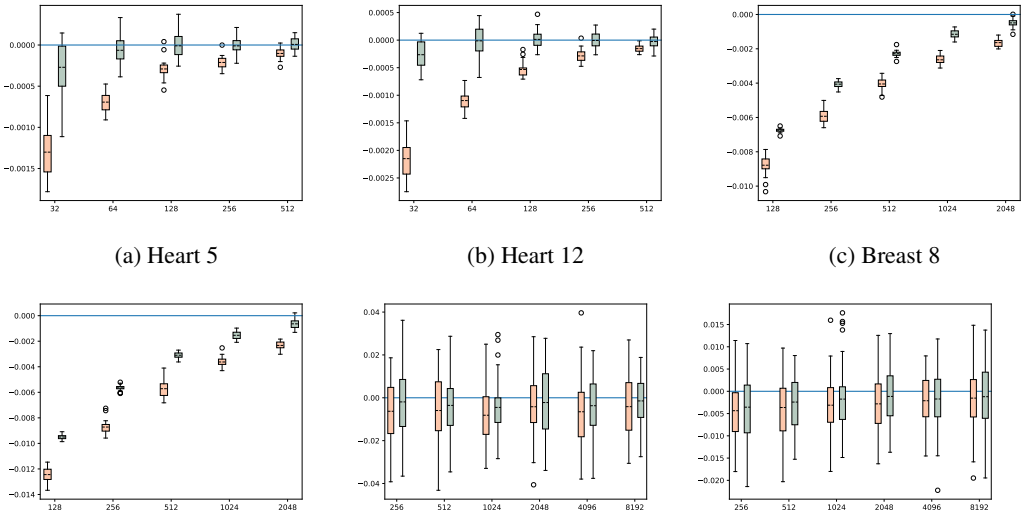
For the log likelihood comparison in Table 2, we use SNIS with the posterior as importance distribution and a total of 10^4 samples. Therefore, the estimation of the log likelihood is:

$$\hat{\mathcal{L}} = T^{-1} \sum_{j=1}^{T} \sum_{i=1}^{M} \omega_{\theta,\phi,x_j} \log p_\theta(x_j \mid z_i^j)$$

with $\omega_{\theta,\phi,x}(z) = p_\theta(x)/q_\phi(z \mid x)$ where z_i^j is sampled from $q_\phi(\cdot \mid x_j)$.

Dataset	component	M	k_{max}	N
breast	8	128	8	17
breast	8	256	16	17
breast	8	512	16	33
breast	8	1024	32	33
breast	8	2048	64	33
breast	11	128	8	17
breast	11	256	16	17
breast	11	512	32	17
breast	11	1024	64	17
breast	11	2048	128	17
breast	14	128	8	17
breast	14	256	16	17
breast	14	512	32	17
breast	14	1024	64	17
breast	14	2048	128	17
covertype	6	512	4	129
covertype	6	1024	8	129
covertype	6	2048	16	129
covertype	6	4096	2	2049
covertype	6	8192	4	2049
covertype	17	512	2	257
covertype	17	1024	2	513
covertype	17	2048	2	1025
covertype	17	4096	2	2049
covertype	17	8192	4	2049
covertype	23	512	2	257
covertype	23	1024	2	513
covertype	23	2048	4	513
covertype	23	4096	16	257
covertype	23	8192	32	257
heart	5	32	4	9
heart	5	64	8	9
heart	5	128	16	9
heart	5	256	32	9
heart	5	512	4	129
heart	8	32	4	9
heart	8	64	4	17
heart	8	128	8	17
heart	8	256	32	9
heart	8	512	8	65
heart	12	32	4	9
heart	12	64	4	17
heart	12	128	8	17
heart	12	256	4	65
heart	12	512	8	65

Table 3: Optimal configurations for Figures 3 and 6



(a) Heart 5 (b) Heart 12 (c) Breast 8

(d) Breast 14 (e) Covertype 17 (f) Covertype 23

Figure 6: Visualisation of the distribution of the bias for the Heart Failure and Breast cancer dataset for other components of θ

Name	Value	Name	Value
Batch size	32	Epochs	100
IS samples(IWAE, Un IWAE)	64	Encoder inner layer size	256
Learning rate	10^{-3}	Decoder inner layer size	256

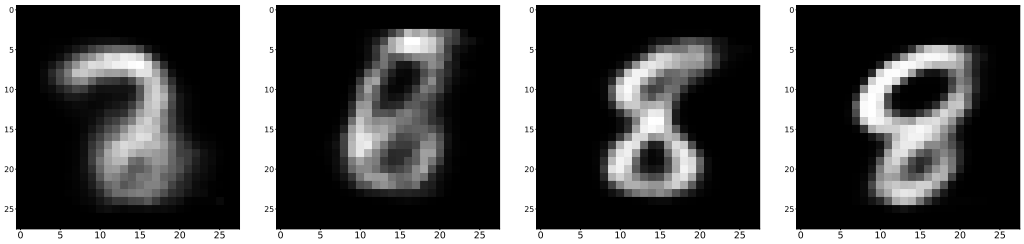
Table 4: Training Parameters

C.4 Resources

All the simulations were done using a server with the following configuration:

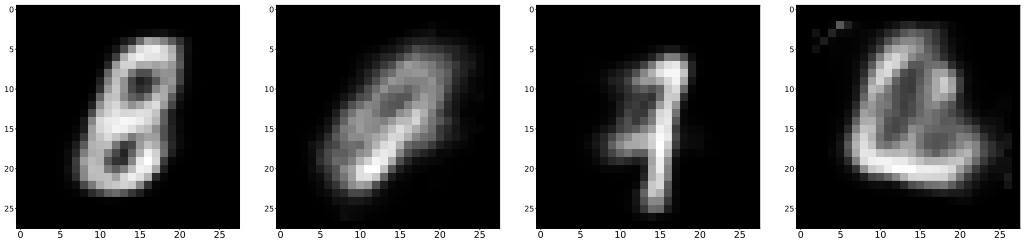
- GPUs: two Tesla V100-PCIE (32Gb RAM)
- CPU: 71 Intel(R) Xeon(R) Gold 6154 CPU @ 3.00GHz
- RAM: 377Gb

locally hosted. We estimate the total number of computing hours for the results presented in this paper to be inferior to 200 hours of GPU usage (All the calculations were done in the GPU).



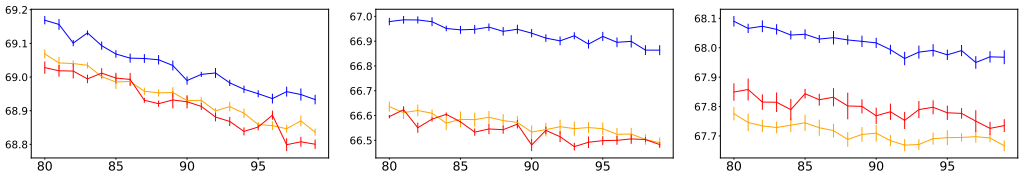
(a) VAE (b) IWAE (c) Un-IWAE ($k = 4$) (d) Un-IWAE ($k = 8$)

Figure 7: Images generated by the autoencoders from a sample of the prior in the latent space for dimension 20.



(a) VAE (b) IWAE (c) Un-IWAE ($k = 4$) (d) Un-IWAE ($k = 8$)

Figure 8: Images generated by the autoencoders from a sample of the prior in the latent space for dimension 50.



(a) Dimension 20 (b) Dimension 50 (c) Dimension 100

Figure 9: Per epoch training loss (ELBO) for the last 25 epochs. Confidence intervals are calculated as $1.96\sigma/\sqrt{n}$ over 4 ($n = 4$) different seeds. Blue corresponds to IWAE, orange to BR-IWAE with $k = 4$ and red to BR-IWAE with $k = 8$.