Statistics and Computing (2021) 31:12
https://doi.org/10.1007/s11222-020-09983-1

# Convergence rates for optimised adaptive importance samplers

**Ömer Deniz Akyildiz**[1,2] · **Joaquín Míguez**[3]

Received: 8 October 2019 / Accepted: 26 September 2020 / Published online: 21 January 2021


**Abstract**
Adaptive importance samplers are adaptive Monte Carlo algorithms to estimate expectations with respect to some target distribution which *adapt* themselves to obtain better estimators over a sequence of iterations. Although it is straightforward to show that they have the same $\mathcal{O}(1/\sqrt{N})$ convergence rate as standard importance samplers, where $N$ is the number of Monte Carlo samples, the behaviour of adaptive importance samplers over the number of iterations has been left relatively unexplored. In this work, we investigate an adaptation strategy based on convex optimisation which leads to a class of adaptive importance samplers termed *optimised adaptive importance samplers* (OAIS). These samplers rely on the iterative minimisation of the $\chi^2$-divergence between an exponential family proposal and the target. The analysed algorithms are closely related to the class of adaptive importance samplers which minimise the variance of the weight function. We first prove non-asymptotic error bounds for the mean squared errors (MSEs) of these algorithms, which explicitly depend on the number of iterations and the number of samples together. The non-asymptotic bounds derived in this paper imply that when the target belongs to the exponential family, the $L_2$ errors of the optimised samplers converge to the optimal rate of $\mathcal{O}(1/\sqrt{N})$ and the rate of convergence in the number of iterations are explicitly provided. When the target does *not* belong to the exponential family, the rate of convergence is the same but the asymptotic $L_2$ error increases by a factor $\sqrt{\rho^\star} > 1$, where $\rho^\star - 1$ is the minimum $\chi^2$-divergence between the target and an exponential family proposal.

**Keywords** Adaptive importance sampling · Convex optimization · Variational Inference · Importance sampling

## 1 Introduction

The class of adaptive importance sampling (AIS) methods is a key Monte Carlo methodology for estimating integrals that cannot be obtained in closed form (Robert and Casella 2004). This problem arises in many settings, such as Bayesian signal processing and machine learning (Bugallo et al. 2015, 2017) or optimal control, (Kappen and Ruiz 2016) where the quantities of interest are usually defined as intractable expectations. Adaptive importance samplers are versions of classical importance samplers (IS) which iteratively improve the proposals to generate samples better suited to the estimation problem at hand. Its variants include, for example, *population Monte Carlo* methods (Cappé et al. 2004) and adaptive mixture importance sampling (Cappé et al. 2008). Since there has been a surge of papers on the topic of AIS recently, a comprehensive review is beyond the scope of this article; see e.g. Bugallo et al. (2017) for a recent review.

Due to the popularity of the adaptive importance samplers, their theoretical performance has also received attention in the past few years. The same as conventional IS methods, AIS schemes enjoy the classical $\mathcal{O}(1/\sqrt{N})$ convergence rate of the $L_2$ error, where $N$ is the number of Monte Carlo samples used in the approximations, see e.g. Robert and Casella (2004) and Agapiou et al. (2017). However, since an adaptation is performed over the iterations and the goal of this

Ö.D.A. is funded by the Lloyds Register Foundation programme on Data Centric Engineering through the London Air Quality project. This work was supported by The Alan Turing Institute for Data Science and AI under EPSRC Grant EP/N510129/1. J.M. acknowledges the support of the Spanish *Agencia Estatal de Investigación* (awards TEC2015-69868-C2-1-R ADVENTURE and RTI2018-099655-B-I00 CLARA) and the Office of Naval Research (Award No. N00014-19-1-2226).

✉ Ömer Deniz Akyildiz
omer.akyildiz@warwick.ac.uk

Joaquín Míguez
joaquin.miguez@uc3m.es

[1] University of Warwick, Coventry, UK

[2] The Alan Turing Institute, London, UK

[3] Universidad Carlos III de Madrid & Instituto de Investigación Sanitaria Gregorio Marañón, Leganés, Spain

adaptation is to improve the proposal quality, an insightful convergence result would provide a bound which explicitly depends on the number of iterations, $t$, (which sometimes we refer to as *time*) and the number of samples, $N$. Although there are convergence results of adaptive methods (see Douc et al. (2007) for a convergence theory for population Monte Carlo based on minimizing Kullback–Leibler divergence), none of the available results yields an explicit bound of the error in terms of the number of iterations and the number of particles at the same time.

One difficulty of proving such a result for adaptive mixture samplers is that the adaptive mixtures form an interacting particle system, and it is unclear what kind of adaptation they perform or whether the adapted proposals actually get closer to the target for some metric. An alternative to adaptation using mixtures is the idea of minimizing a cost function in order to adapt the proposal. This idea has been popular in the literature, in particular, minimizing the variance of the weight function has received significant attention, see, e.g. Arouna (2004a, b); Kawai (2008); Lapeyre and Lelong (2011); Ryu and Boyd (2014); Kawai (2017, 2018). Relevant to us, in particular, is the work of Ryu and Boyd (2014), who have proposed an algorithm called Convex Adaptive Monte Carlo (Convex AdaMC). This scheme is based on minimizing the variance of the IS estimator, which is a quantity related to the $\chi^2$ divergence between the target and the proposal. Ryu and Boyd (2014) have shown that the variance of the IS estimator is a convex function of the parameters of the proposal when the latter is chosen within the exponential family. Based on this observation, Ryu and Boyd (2014) have formulated Convex AdaMC, which draws one sample at each iteration and construct the IS estimator, which requires access to the normalised target. They proved a central limit theorem (CLT) for the resulting sampler. The idea has been further extended for self-normalised importance samplers by Ryu (2016), who considered minimising the $\alpha$-divergence between the target and an exponential family. Similarly, Ryu (2016) proved a CLT for the resulting sampler. Similar ideas were also considered by Kawai (2017, 2018), who also aimed at minimizing the variance expression. Similarly, Kawai (2018) showed that the variance of the weight function is convex when the proposal family is suitably chosen and provided general conditions for such proposals. Kawai (2018) has also developed an adaptation technique based on the stochastic approximation, which is similar to the scheme we analyse in this paper. There have been other results also considering $\chi^2$ divergence and relating it to the necessary sample size of the IS methods, see, e.g. Sanz-Alonso (2018). Following the approach of Chatterjee et al. (2018), Sanz-Alonso (2018) considers and ties the necessary sample size to $\chi^2$-divergence, in particular, shows that the necessary sample size grows with $\chi^2$-divergence, hence implying that minimizing it can lead to more efficient importance sampling procedures.

In this work, we develop and analyse a family of adaptive importance samplers, coined *optimised adaptive importance samplers* (OAIS), which relies on a particular adaptation strategy based on convex optimisation. We adapt the proposal with respect to a quantity (essentially the $\chi^2$-divergence between the target and the proposal) that also happens to be the constant in the error bounds of the IS [see, e.g. (Agapiou et al. 2017)]. Assuming that proposal distributions belong to the exponential family, we recast the adaptation of the proposal as a convex optimisation problem and then develop a procedure which essentially optimises the $L_2$ error bound of the algorithm. By using results from convex optimisation, we obtain error rates depending on the number of iterations, denoted as $t$, and the number of Monte Carlo samples, denoted as $N$, together. In this way, we explicitly display the trade-off between these two essential quantities. To the best of our knowledge, none of the papers on the topic provides convergence rates depending explicitly on the number of iterations and the number of particles together, as we do herein.

The paper is organised as follows. In Sect. 2, we introduce the problem definition, the IS and the AIS algorithms. In Sect. 3, we introduce the OAIS algorithms. In Sect. 4, we provide the theoretical results regarding optimised AIS and show its convergence using results from convex optimisation. Finally, we make some concluding remarks in Sect. 5.

## Notation

For $L \in \mathbb{N}$, we use the shorthand $[L] = \{1, \ldots, L\}$. We denote the state space as $\mathsf{X}$ and assume $\mathsf{X} \subseteq \mathbb{R}^{d_x}$, $d_x \geq 1$. The space of bounded real-valued functions and the set of probability measures on space $\mathsf{X}$ are denoted as $B(\mathsf{X})$ and $\mathcal{P}(\mathsf{X})$, respectively. Given $\varphi \in B(\mathsf{X})$ and $\pi \in \mathcal{P}(\mathsf{X})$, the expectation of $\varphi$ with respect to (w.r.t.) $\pi$ is written as $(\varphi, \pi) = \int \varphi(x)\pi(\mathrm{d}x)$ or $\mathbb{E}_\pi[\varphi(X)]$. The variance of $\varphi$ w.r.t. $\pi$ is defined as $\mathrm{var}_\pi(\varphi) = (\varphi^2, \pi) - (\varphi, \pi)^2$. If $\varphi \in B(\mathsf{X})$, then $\|\varphi\|_\infty = \sup_{x \in \mathsf{X}} |\varphi(x)| < \infty$. The unnormalised density associated with $\pi$ is denoted with $\Pi(x)$. We denote the proposal as $q_\theta \in \mathcal{P}(\mathsf{X})$, with an explicit dependence on the parameter $\theta \in \Theta$. The parameter space is assumed to be a subset of $d_\theta$-dimensional Euclidean space, i.e. $\Theta \subseteq \mathbb{R}^{d_\theta}$.

Whenever necessary, we denote both the probability measures, $\pi$ and $q_\theta$, and their densities with the same notation. To be specific, we assume that both $\pi(\mathrm{d}x)$ and $q_\theta(\mathrm{d}x)$ are absolutely continuous with respect to the Lebesgue measure, and we denote their associated densities as $\pi(x)$ and $q_\theta(x)$. The use of either the measure or the density will be clear from both the argument (sets or points, respectively) and the context.

## 2 Background

In this section, we review importance and adaptive importance samplers.

### 2.1 Importance sampling

Consider a target density $\pi \in \mathcal{P}(\mathsf{X})$ and a bounded function $\varphi \in B(\mathsf{X})$. Often, the main interest is to compute an integral of the form

$$(\varphi, \pi) = \int_{\mathsf{X}} \varphi(x)\pi(x)\mathrm{d}x. \tag{2.1}$$

While perfect Monte Carlo can be used to estimate this expectation when it is possible to sample exactly from $\pi(x)$, this is in general not tractable. Hereafter, we consider the cases when the target can be evaluated exactly and up to a normalising constant, respectively.

Importance sampling (IS) uses a proposal distribution which is easy to sample and evaluate. The method consists in weighting these samples, in order to correct the discrepancy between the target and the proposal, and finally constructing an estimator of the integral. To be precise, let $q_\theta \in \mathcal{P}(\mathsf{X})$ be the proposal which is parameterised by the vector $\theta \in \Theta$. The unnormalised target density is denoted as $\Pi : \mathsf{X} \to \mathbb{R}_+$. Therefore, we have

$$\pi(x) = \frac{\Pi(x)}{Z_\pi},$$

where $Z_\pi := \int_{\mathsf{X}} \Pi(x)\mathrm{d}x < \infty$. Next, we define functions $w_\theta, W_\theta : \mathsf{X} \times \Theta \to \mathbb{R}_+$ as

$$w_\theta(x) = \frac{\pi(x)}{q_\theta(x)} \quad \text{and} \quad W_\theta(x) = \frac{\Pi(x)}{q_\theta(x)},$$

respectively. For a chosen proposal $q_\theta$, the IS proceeds as follows. First, a set of independent and identically distributed (iid) samples $\{x^{(i)}\}_{i=1}^N$ is generated from $q_\theta$. When $\pi(x)$ can be evaluated, one constructs the empirical approximation of the probability measure $\pi$, denoted $\pi_\theta^N$, as

$$\pi_\theta^N(\mathrm{d}x) = \frac{1}{N}\sum_{i=1}^N w_\theta(x^{(i)})\delta_{x^{(i)}}(\mathrm{d}x),$$

where $\delta_{x'}(\mathrm{d}x)$ denotes the Dirac delta measure that places unit probability mass at $x = x'$. For this case, the IS estimate of the integral in (2.1) can be given as

$$(\varphi, \pi_\theta^N) = \frac{1}{N}\sum_{i=1}^N w_\theta(x^{(i)})\varphi(x^{(i)}). \tag{2.2}$$

However, in most practical cases, the target density $\pi(x)$ can only be evaluated up to an unknown normalizing proportionality constant (i.e. we can evaluate $\Pi(x)$ but not $Z_\pi$). In this case, we construct the empirical measure $\pi_\theta^N$ as

$$\pi_\theta^N(\mathrm{d}x) = \sum_{i=1}^N \mathsf{w}_\theta^{(i)}\delta_{x^{(i)}}(\mathrm{d}x),$$

where

$$\mathsf{w}_\theta^{(i)} = \frac{W_\theta(x^{(i)})}{\sum_{j=1}^N W_\theta(x^{(j)})}.$$

Finally, this construction leads to the so-called self-normalizing importance sampling (SNIS) estimator

$$(\varphi, \pi_\theta^N) = \sum_{i=1}^N \mathsf{w}_\theta^{(i)}\varphi(x^{(i)}). \tag{2.3}$$

Although the IS estimator (2.2) is unbiased, the SNIS estimator (2.3) is in general biased. However, the bias and the MSE vanish with a rate $\mathcal{O}(1/N)$, therefore providing guarantees of convergence as $N \to \infty$. Crucially for us, the MSE of both estimators. For clarity, below, we present an MSE bound for the (more general) SNIS estimator (2.3) which is adapted from Agapiou et al. (2017).

**Theorem 1** *Assume that* $(W_\theta^2, q_\theta) < \infty$. *Then for any* $\varphi \in B(\mathsf{X})$, *we have*

$$\mathbb{E}\left[\left((\varphi, \pi) - (\varphi, \pi_\theta^N)\right)^2\right] \leq \frac{c_\varphi \rho(\theta)}{N}, \tag{2.4}$$

*where* $c_\varphi = 4\|\varphi\|_\infty^2$ *and the function* $\rho : \Theta \to [\rho^\star, \infty)$ *is defined as*

$$\rho(\theta) = \mathbb{E}_{q_\theta}\left[\frac{\pi^2(X)}{q_\theta^2(X)}\right], \tag{2.5}$$

*where* $\rho^\star := \inf_{\theta\in\Theta} \rho(\theta) \geq 1$.

***Proof*** See Appendix A.1 for a self-contained proof. □

***Remark 1*** For the IS estimator (2.2), this bound can be improved so that $c_\varphi = \|\varphi\|_\infty^2$. However, this improvement does not effect our results in this paper; hence, we present a single bound of the form in (2.4) for the estimators (2.2) and (2.3) for conciseness. □

***Remark 2*** As pointed out by Agapiou et al. (2017), the function $\rho$ is essentially the $\chi^2$ divergence between $\pi$ and $q_\theta$, i.e.

$$\rho(\theta) := \chi^2(\pi||q_\theta) + 1.$$

Note that $\rho(\theta)$ can also be expressed in terms of the variance of the weight function $w_\theta$, which coincides with the $\chi^2$-divergence, i.e.

$$\rho(\theta) = \mathrm{var}_{q_\theta}(w_\theta(X)) + 1.$$

Therefore, minimizing $\rho(\theta)$ is equivalent to minimizing $\chi^2$-divergence and the variance of the weight function $w_\theta$, i.e. $\mathrm{var}_{q_\theta}(w_\theta(X))$. □

***Remark 3*** Remark 2 implies that when both $\pi$ and $q_\theta$ belong to the same parametric family (i.e. there exists $\theta \in \Theta$ such that $\pi = q_\theta$), one readily obtains

$$\rho^\star := \inf_{\theta\in\Theta} \rho(\theta) = 1. \quad \square$$

***Remark 4*** For the IS estimator (2.2), the bound in Theorem 1 can be modified so that it holds for unbounded test functions $\varphi$ as well; see, e.g. Ryu and Boyd (2014). Therefore, a similar quantity to $\rho(\theta)$, which includes $\varphi$ while still retaining convexity, can be optimised for this case. Unfortunately, obtaining such a bound is not straightforward for the SNIS estimator (2.3) as shown by Agapiou et al. (2017). In order to significantly simplify the presentation, we restrict ourselves to the class of bounded test functions, i.e. we assume $\|\varphi\|_\infty < \infty$. □

Finally, we present a bias result from Agapiou et al. (2017).

**Theorem 2** *Assume that* $(W_\theta^2, q_\theta) < \infty$. *Then for any* $\varphi \in B(\mathsf{X})$, *we have*

$$\left|\mathbb{E}\left[(\varphi, \pi_\theta^N)\right] - (\varphi, \pi)\right| \leq \frac{\bar{c}_\varphi \rho(\theta)}{N},$$

*where* $\bar{c}_\varphi = 12\|\varphi\|_\infty^2$ *and the function* $\rho : \Theta \to [\rho^\star, \infty)$ *is the same as in Theorem 1.*

***Proof*** See Theorem 2.1 in Agapiou et al. (2017). □

## 2.2 Parametric adaptive importance samplers

Standard importance sampling may be inefficient in practice when the proposal is poorly calibrated with respect to the target. In particular, as implied by the error bound provided in Theorem 1, the error made by the IS estimator can be high if the $\chi^2$-divergence between the target and the proposal is large. Therefore, it is more common to employ an iterative version of importance sampling, also called as *adaptive importance sampling* (AIS). The AIS algorithms are importance sampling methods which aim at iteratively improving the proposal distributions. More specifically, the AIS methods specify a sequence of proposals $(q_t)_{t\geq 1}$ and perform importance sampling at each iteration. The aim is to improve the proposal so that the samples are better matched with the target, which results in less variance and more accuracy in the estimators. There are several variants, the most popular one being population Monte Carlo methods (Cappé et al. 2004) which uses previous samples in the proposal.

**Algorithm 1** Parametric AIS

1: Choose a parametric proposal $q_\theta$ with initial parameter $\theta = \theta_0$.
2: **for** $t \geq 1$ **do**
3: Adapt the proposal,

$$\theta_t = \mathcal{T}_t(\theta_{t-1}),$$

4: Sample,

$$x_t^{(i)} \sim q_{\theta_t}, \quad \text{for } i = 1, \ldots, N,$$

5: Compute weights,

$$\mathsf{w}_{\theta_t}^{(i)} = \frac{W_{\theta_t}(x_t^{(i)})}{\sum_{i=1}^N W_{\theta_t}(x_t^{(i)})}, \quad \text{where} \quad W_{\theta_t}^{(i)} = \frac{\Pi(x_t^{(i)})}{q_{\theta_t}(x^{(i)})}.$$

6: Report the point-mass probability measure

$$\pi_{\theta_t}^N(\mathrm{d}x) = \sum_{i=1}^N \mathsf{w}_{\theta_t}^{(i)} \delta_{x_t^{(i)}}(\mathrm{d}x),$$

and the estimator

$$(\varphi, \pi_{\theta_t}^N) = \sum_{i=1}^N \mathsf{w}_{\theta_t}^{(i)} \varphi(x_t^{(i)}).$$

7: **end for**

In this section, we review one particular AIS, which we refer to as *parametric AIS*. In this variant, the proposal distribution is a parametric distribution, denoted $q_\theta$. Over time, this parameter $\theta$ is updated (or *optimised*) with respect to a predefined criterion resulting in a sequence $(\theta_t)_{t\geq 1}$. This yields a sequence of proposal distributions denoted as $(q_{\theta_t})_{t\geq 1}$.

One iteration of the algorithm goes as follows. Assume at time $t-1$ we are given a proposal distribution $q_{\theta_{t-1}}$. At time $t$, we first update the parameter of this proposal,

$$\theta_t = \mathcal{T}_t(\theta_{t-1}),$$

where $\{\mathcal{T}_t : \Theta \to \Theta, t \geq 1\}$, is a sequence of (deterministic or stochastic) maps, e.g. gradient mappings, constructed so that they minimise a certain cost function. Then, in the same way, we have done in conventional IS, we sample

$$x_t^{(i)} \sim q_{\theta_t}(\mathrm{d}x), \quad \text{for } i = 1, \ldots, N,$$

compute weights

$$\mathsf{w}_{\theta_t}^{(i)} = \frac{W_{\theta_t}(x_t^{(i)})}{\sum_{i=1}^N W_{\theta_t}(x_t^{(i)})},$$

and finally construct the empirical measure

$$\pi_{\theta_t}^N(\mathrm{d}x) = \sum_{i=1}^N \mathrm{w}_{\theta_t}^{(i)} \delta_{x_t^{(i)}}(\mathrm{d}x).$$

The estimator of the integral (2.1) is then computed as in Eq. (2.3).

The full procedure of the parametric AIS method is summarized in Algorithm 1. Since this is a valid IS scheme, this algorithm enjoys the same guarantee provided in Theorem 1. In particular, we have the following theorem.

**Theorem 3** *Assume that given a sequence of proposals* $(q_{\theta_t})_{t\geq 1} \in \mathcal{P}(\mathsf{X})$*, we have* $(W_{\theta_t}^2, q_{\theta_t}) < \infty$ *for every* $t$*. Then for any* $\varphi \in B(\mathsf{X})$*, we have*

$$\mathbb{E}\left[\left|(\varphi,\pi) - (\varphi,\pi_{\theta_t}^N)\right|^2\right] \leq \frac{c_\varphi \rho(\theta_t)}{N},$$

*where* $c_\varphi = 4\|\varphi\|_\infty^2$ *and the function* $\rho(\theta_t) : \Theta \to [\rho^\star, \infty)$ *is defined as in Eq.* (2.5)*.*

***Proof*** The proof is identical to the proof of Theorem 1. We have just re-stated the result to introduce the iteration index $t$. □

However, this theorem does not give an insight of what happens as the number of iterations increases, i.e. when $t \to \infty$, with the bound. Ideally, the adaptation of the AIS should improve this bound with time. In other words, in the ideal case, the error should decrease as $t$ grows. Fortunately, Theorem 3 suggests that the maps $\mathcal{T}_t : \Theta \to \Theta$ can be chosen so that the function $\rho$ is minimised over time. More specifically, the sequence $(\theta_t)_{t\geq 1}$ can be chosen so that it leads to a decreasing sequence (at least in expectation) $(\rho(\theta_t))_{t\geq 1}$. In the following sections, we will summarize the deterministic and stochastic strategies to achieve this aim.

***Remark 5*** We define the unnormalised version of $\rho(\theta)$ and denote it as $R(\theta)$. It is characterised as follows:

$$\rho(\theta) = \frac{R(\theta)}{Z_\pi^2} \quad \text{where} \quad Z_\pi = \int_{\mathsf{X}} \Pi(x)\mathrm{d}x < \infty.$$

Hence, $R(\theta)$ can also be expressed as

$$R(\theta) = \mathbb{E}_{q_\theta}\left[\frac{\Pi^2(X)}{q_\theta^2(X)}\right]. \tag{2.6}$$

□

### 2.3 AIS with exponential family proposals

Following Ryu and Boyd (2014), we note that when $q_\theta$ is chosen as an exponential family density, the function $\rho(\theta)$ is convex. In particular, we define

$$q_\theta(x) = \exp(\theta^\top T(x) - A(\theta))h(x), \tag{2.7}$$

where $A : \mathbb{R}^{d_\theta} \to \mathbb{R} \cup \{\infty\}$ is the log of the normalization constant, i.e.

$$A(\theta) = \log \int \exp(\theta^\top T(x))h(x)\mathrm{d}x,$$

while $T : \mathbb{R}^{d_x} \to \mathbb{R}^{d_\theta}$ and $h : \mathbb{R}^{d_x} \to \mathbb{R}_+$. Then, we have the following lemma adapted from Ryu and Boyd (2014).

**Lemma 1** *Let* $q_\theta$ *be chosen as in* (2.7)*. Then,* $\rho : \Theta \to [\rho^\star, \infty)$ *is convex, i.e. for any* $\theta_1, \theta_2 \in \Theta$ *and* $\lambda \in [0, 1]$*, the following inequality holds*

$$\rho(\lambda\theta_1 + (1-\lambda)\theta_2) \leq \lambda\rho(\theta_1) + (1-\lambda)\rho(\theta_2).$$

***Proof*** See Appendix A.2 for a self-contained proof. □

Lemma 1 shows that $\rho$ is a convex function; therefore, optimising it could give us provably convergent algorithms (as $t$ increases). Next lemma, borrowed from Ryu and Boyd (2014), shows that $\rho$ is differentiable and its gradient can indeed be computed as an expectation.

**Lemma 2** *The gradient* $\nabla\rho(\theta)$ *can be written as*

$$\nabla\rho(\theta) = \mathbb{E}_{q_\theta}\left[(\nabla A(\theta) - T(X))\frac{\pi^2(X)}{q_\theta^2(X)}\right]. \tag{2.8}$$

***Proof*** The proof is straightforward since $q_\theta$ is from an exponential family and $A(\theta)$ is differentiable. □

***Remark 6*** Note that Eqs. (2.6) and (2.8) together imply that

$$\nabla R(\theta) = \mathbb{E}_{q_\theta}\left[(\nabla A(\theta) - T(X))\frac{\Pi^2(X)}{q_\theta^2(X)}\right]. \tag{2.9}$$

We also note (see Remark 5) that

$$\nabla R(\theta) = Z_\pi^2 \nabla\rho(\theta). \tag{2.10}$$

□

In the following sections, we assume that $\rho(\theta)$ is a convex function. Thus, Lemma 1 constitutes an important motivation for our approach. We leave general proposals which lead to nonconvex $\rho(\theta)$ for future work.

## 3 Algorithms

In this section, we describe adaptation strategies based on minimizing $\rho(\theta)$. In particular, we design maps $\mathcal{T}_t : \Theta \to \Theta$, for $t \geq 1$, for scenarios where

(i) the gradient of $\rho(\theta)$ can be exactly computed,

(ii) an unbiased estimate of the gradient of $\rho(\theta)$ can be obtained, and
(iii) an unbiased estimate of the gradient of $R(\theta)$ can be obtained.

Scenario (i) is unrealistic in practice but gives us a guideline in order to further develop the idea. In particular, the error bounds for the more complicated cases follow the same structure as this case. Therefore, the results obtained in case (i) provide a good qualitative understanding of the results introduced later. Scenario (ii) can be realized in cases where it is possible to evaluate $\pi(x)$, in which case the IS leads to unbiased estimators. Scenario (iii) is what a practitioner would most often encounter: the target can only be evaluated up to the normalizing constant, i.e. $\Pi(x)$ can be evaluated but $\pi(x)$ cannot.

We finally remark that for the cases where we assume a stochastic gradient can be obtained for $\rho$ and $R$ (namely, the case (ii) and the case (iii) respectively), we consider two possible algorithms to perform adaptation. The first method is a *vanilla* SGD algorithm (Bottou et al. 2016) and the second method is a SGD scheme with iterate averaging (Schmidt et al. 2017). While vanilla SGD is easier to implement and algorithmically related to population-based Monte Carlo methods, iterate averaged SGD results in a better theoretical bound and it has some desirable variance reduction properties.

### 3.1 Exact gradient OAIS

We first introduce the OAIS scheme where we assume that the exact gradients of $\rho(\theta)$ are available. Since $\rho$ is defined as an expectation (an integral), this assumption is unrealistic. However, the results we can prove for this procedure shed light onto the results that will be proved for practical scenarios in the following sections.

In particular, in this scheme, given $\theta_{t-1}$, we specify $\mathcal{T}_t$ as

$$\theta_t = \mathcal{T}_t(\theta_{t-1}) = \mathsf{Proj}_{\Theta}(\theta_{t-1} - \gamma \nabla \rho(\theta_{t-1})), \tag{3.1}$$

where $\gamma > 0$ is the step-size parameter of the map and $\mathsf{Proj}_{\Theta}$ denotes projection onto the compact parameter space $\Theta$. This is a classical gradient descent scheme on $\rho(\theta)$. In Sect. 4.1, we provide non-asymptotic results for this scheme. However, as we have noted, this idea does not lead to a practical scheme and cannot be used in most cases in practice as the gradients of $\rho$ in exact form are rarely available.

***Remark 7*** We use a projection operator in Eq. (3.1) because we assume throughout the analysis in Sect. 4 that the parameter space $\Theta$ is compact. □

### 3.2 Stochastic gradient OAIS

Although it has a nice and simple form, exact-gradient OAIS is often intractable as, in most practical cases, the gradient can only be estimated. In this section, we first look at the case where $\pi(x)$ can be evaluated, which means that an unbiased estimate of $\nabla\rho(\theta)$ can be obtained. Then, we consider the general case, where one can only evaluate $\Pi(x)$ and can obtain an unbiased estimate of $\nabla R(\theta)$.

In the following subsections, we consider an algorithm where the gradient is estimated using samples which can also be used to construct importance sampling estimators. The procedure is outlined in Algorithm 3 for the case in which only $\Pi(x)$ can be evaluated and $\nabla R(\theta)$ is estimated.

#### 3.2.1 Normalised case

If we assume that the density $\pi(x)$ can be evaluated exactly, then the algorithm can be described as follows. Given $(\theta_k)_{1\le k\le t-1}$, at iteration $t$, we compute the next parameter iterate as

$$\theta_t = \mathsf{Proj}_{\Theta}(\theta_{t-1} - \gamma_t g_t),$$

where $g_t$ is an unbiased estimator of $\nabla\rho(\theta_{t-1})$. We note that due to the analytical form of $\nabla\rho$ (see Eq. (2.8)), the samples and weights generated at iteration $t-1$, i.e. $\left\{x_{t-1}^{(i)}, w_{\theta_{t-1}}(x_{t-1}^{(i)})\right\}_{i=1}^{N}$ can be reused to estimate the gradient. This makes an algorithmic connection to the population Monte Carlo methods where previous samples and weights are used to adapt the proposal (Cappé et al. 2004).

Given the updated parameter $\theta_t$, the algorithm first samples from the updated proposal $x_t^{(i)} \sim q_{\theta_t}$, $i = 1, \ldots, N$, and then proceeds to construct the IS estimator as in (2.2). Namely,

$$(\varphi, \pi_{\theta_t}^N) = \frac{1}{N}\sum_{i=1}^{N} w_{\theta_t}(x_t^{(i)})\varphi(x_t^{(i)}). \tag{3.2}$$

#### 3.2.2 Self-normalised case

For the general case, where we can only evaluate $\Pi(x)$, the algorithm proceeds similarly. Given $(\theta_k)_{1\le k\le t-1}$, the method proceeds by first updating the parameter

$$\theta_t = \mathsf{Proj}_{\Theta}(\theta_{t-1} - \gamma_t \tilde{g}_t),$$

where $\tilde{g}_t$ is an unbiased estimator of $\nabla R(\theta_{t-1})$. Given the updated parameter, we first sample $x_t^{(i)} \sim q_{\theta_t}, i = 1, \ldots, N$,

**Algorithm 2** Stochastic gradient OAIS

1: Choose a parametric proposal $q_\theta$ with initial parameter $\theta = \theta_0$.
2: **for** $t \geq 1$ **do**
3: Update the proposal parameter,

$$\theta_t = \mathsf{Proj}_\Theta(\theta_{t-1} - \gamma_t \tilde{g}_t)$$

where $\tilde{g}_t$ is computed by approximating the expectation in Eq. (2.9) using the samples $x_{t-1}^{(i)}$ and weights $\mathsf{w}_{\theta_{t-1}}^{(i)} = \Pi(x_{t-1}^{(i)}) q_{\theta_{t-1}}(x_{t-1}^{(i)})^{-1}$, $i = 1, \ldots, N$.
4: Sample,

$$x_t^{(i)} \sim q_{\theta_t}, \quad \text{for } i = 1, \ldots, N,$$

5: Compute weights,

$$\mathsf{w}_{\theta_t}^{(i)} = \frac{W_{\theta_t}(x_t^{(i)})}{\sum_{i=1}^N W_{\theta_t}(x_t^{(i)})}.$$

6: Report,

$$\pi_{\theta_t}^N(\mathrm{d}x) = \sum_{i=1}^N \mathsf{w}_{\theta_t}^{(i)} \delta_{x_t^{(i)}}(\mathrm{d}x),$$

and

$$(\varphi, \pi_{\theta_t}^N) = \sum_{i=1}^N \mathsf{w}_{\theta_t}^{(i)} \varphi(x_t^{(i)}).$$

7: **end for**

and then construct the SNIS estimate as in (2.3), i.e.

$$(\varphi, \pi_{\theta_t}^N) = \sum_{i=1}^N \mathsf{w}_{\theta_t}^{(i)} \varphi(x_t^{(i)}).$$

where

$$\mathsf{w}_{\theta_t}^{(i)} = \frac{W_{\theta_t}(x^{(i)})}{\sum_{j=1}^N W_{\theta_t}(x^{(j)})},$$

## 3.3 Stochastic gradient OAIS with averaged iterates

Next, we describe a variant of the stochastic gradient OAIS that uses averages of the iterates generated by the SGD scheme (Schmidt et al. 2017) in order to compute the proposal densities, generate samples and compute weights. In Sect. 4, we show that the convergence rate for this method is better than the rate that can be guaranteed for Algorithm 2.

### 3.3.1 Normalised case

We assume first that the density $\pi(x)$ can be evaluated. At the beginning of the $t$-th iteration, the algorithm has generated the sequence $(\theta_k)_{1 \leq k \leq t-1}$. First, in order to perform the adaptive importance sampling steps, we set

$$\bar{\theta}_t = \frac{1}{t} \sum_{k=0}^{t-1} \theta_k \tag{3.3}$$

and sample $\bar{x}_t^{(i)} \sim q_{\bar{\theta}_t}$ for $i = 1, \ldots, N$. Following the standard parametric AIS procedure (Algorithm 1), we obtain the estimate of $(\varphi, \pi)$ as,

$$(\varphi, \pi_{\bar{\theta}_t}^N) = \frac{1}{N} \sum_{i=1}^N w_{\bar{\theta}_t}(\bar{x}_t^{(i)}) \varphi(\bar{x}_t^{(i)}).$$

Next, we update the parameter vector using the projected stochastic gradient step

$$\theta_t = \mathcal{T}_t(\theta_{t-1}) = \mathsf{Proj}_\Theta(\theta_{t-1} - \gamma_t g_t), \tag{3.4}$$

where $g_t$ is an unbiased estimate of $\nabla \rho(\theta_{t-1})$, i.e. $\mathbb{E}[g_t] = \nabla \rho(\theta_{t-1})$ and $\mathsf{Proj}_\Theta$ denotes projection onto the set $\Theta$. Note that in order to estimate this gradient using (2.8), we sample $x_t^{(i)} \sim q_{\theta_{t-1}}$ for $i = 1, \ldots, N$, and estimate the expectation in (2.8). It is worth noting that the samples $\{x_t^{(i)}\}_{i=1}^M$ are different from the samples $\{\bar{x}_t^{(i)}\}_{i=1}^N$ used to estimate $(\varphi, \pi)$.

### 3.3.2 Self-normalised case

In general, $\pi(x)$ cannot be evaluated exactly; hence, a stochastic unbiased estimate of $\nabla \rho(\theta)$ cannot be obtained. When the target can only be evaluated up to a normalisation constant, i.e. only $\Pi(x)$ can be computed, we can use the SNIS procedure as explained in Sect. 2. Therefore, we introduce here the most general version of the stochastic method, coined *stochastic gradient OAIS*, which uses the averaged iterates in (3.3) to construct the proposal functions. The scheme is outlined in Algorithm 3.

To run this algorithm, given the parameter vector $\bar{\theta}_t$ in (3.3), we first generate a set of samples $\{\bar{x}_t^{(i)}\}_{i=1}^N$ from the proposal $q_{\bar{\theta}_t}$. Then, the integral estimate given by the SNIS can be written as,

$$(\varphi, \pi_{\bar{\theta}_t}^N) = \sum_{i=1}^N \mathsf{w}_{\bar{\theta}_t}^{(i)} \varphi(\bar{x}_t^{(i)}),$$

where

$$\mathsf{w}_{\bar{\theta}_t}^{(i)} = \frac{W_{\bar{\theta}_t}(\bar{x}^{(i)})}{\sum_{j=1}^N W_{\bar{\theta}_t}(\bar{x}^{(j)})}.$$

Finally, for the adaptation step, we obtain the unbiased estimate of the gradient $\nabla R(\theta)$ and adapt the parameter as

$$\theta_t = \mathsf{Proj}_\Theta(\theta_{t-1} - \gamma_t \tilde{g}_t) \tag{3.5}$$

**Algorithm 3** Stochastic gradient OAIS with averaged iterates

1: Choose a parametric proposal $q_\theta$ with initial parameter $\theta = \theta_0$.
2: **for** $t \geq 1$ **do**
3: Compute the average parameter vector

$$\bar{\theta}_t = \frac{1}{t}\sum_{k=0}^{t-1}\theta_k$$

4: Sample,

$$\bar{x}_t^{(i)} \sim q_{\bar{\theta}_t}, \quad \text{for } i = 1, \ldots, N,$$

5: Compute weights,

$$\mathsf{w}_{\bar{\theta}_t}^{(i)} = \frac{W_{\bar{\theta}_t}(\bar{x}_t^{(i)})}{\sum_{i=1}^N W_{\bar{\theta}_t}(\bar{x}_t^{(i)})}.$$

6: Report the point-mass probability measure

$$\pi_{\bar{\theta}_t}^N(\mathrm{d}x) = \sum_{i=1}^N \mathsf{w}_{\bar{\theta}_t}^{(i)}\delta_{\bar{x}_t^{(i)}}(\mathrm{d}x),$$

and the estimator

$$(\varphi, \pi_{\bar{\theta}_t}^N) = \sum_{i=1}^N \mathsf{w}_{\bar{\theta}_t}^{(i)}\varphi(\bar{x}_t^{(i)}).$$

7: Update the parameter vector,

$$\theta_t = \mathsf{Proj}_\Theta(\theta_{t-1} - \gamma_t \tilde{g}_t)$$

where $\tilde{g}_t$ is an estimate of $\nabla R(\theta_{t-1})$ computed by approximating the expectation in Eq. (2.9) using a set of iid samples $x_t^{(i)} \sim q_{\theta_{t-1}}$, $i = 1, \ldots, N$.
8: **end for**

where $\tilde{g}_t$ is an unbiased estimate of $\nabla R(\theta_{t-1})$, i.e. $\mathbb{E}[\tilde{g}_t] = \nabla R(\theta_{t-1})$. Note that as in the normalised case, this gradient is estimated by approximating the expectation in (2.9) using iid samples $x_t^{(i)} \sim q_{\theta_{t-1}}$, $i = 1, \ldots, N$. These samples are different, again, from the set $\{\bar{x}_t^{(i)}\}_{i=1}^N$ employed to estimate $(\varphi, \pi)$.

***Remark 8*** In Algorithm 3, the samples $\{\bar{x}_t^{(i)}\}_{i=1}^N$ drawn from the proposal distribution $q_{\bar{\theta}_{t-1}}(\mathrm{d}x)$ are *not* used to compute the gradient estimator $\tilde{g}_t$ which, in turn, is needed to generate the next iterate $\theta_t$. Therefore, if we can afford to generate $T$ iterates, $\theta_0, \ldots, \theta_{T-1}$, with $T$ known before hand, and we are only interested in the estimator $(\varphi, \pi_{\bar{\theta}_T}^N)$ obtained at the last iteration (once the proposal function has been optimized), then it is be possible to skip steps 3–6 in Algorithm 3 up to time $T-1$. Only at time $t = T$, we would compute the average parameter vector $\bar{\theta}_T$, sample $\bar{x}_T^{(i)}$ from the proposal $q_{\bar{\theta}_T}(\mathrm{d}x)$ and generate the point-mass probability measure $\pi_{\bar{\theta}_T}^N$ and the estimator $(\varphi, \pi_{\bar{\theta}_T}^N)$.

## 4 Analysis

Theorem 1 yields an intuitive result about the performance of IS methods in terms of the divergence between the target $\pi$ and the proposal $q_\theta$. We now apply ideas from convex optimisation theory in order to minimize $\rho(\theta)$ and obtain finite-time, finite-sample convergence rates for the AIS procedures outlined in Sect. 3.

### 4.1 Convergence rate with exact gradients

Let us first assume that we can compute the gradient of $\rho(\theta)$ exactly. In particular, we consider the update rule in Eq. (3.1). For the sake of the analysis, we impose some regularity assumptions on the $\rho(\theta)$.

**Assumption 1** Let $\rho(\theta)$ be a convex function with Lipschitz derivatives in the compact space $\Theta$. To be specific, $\rho$ is convex and differentiable, and there exists a constant $L < \infty$ such that

$$\|\nabla\rho(\theta) - \nabla\rho(\theta')\|_2 \leq L\|\theta - \theta'\|_2$$

for any $\theta, \theta' \in \Theta$.

***Remark 9*** Assumption 1 holds when the density $q_\theta(x)$ belongs to an exponential family (see Sect. 2.3) and $\Theta$ is compact (Ryu and Boyd 2014), even if it may not hold in general for $\theta \in \mathbb{R}^{d_\theta}$. □

**Lemma 3** *If Assumption 1 holds and we set a step-size* $\gamma \leq 1/L$*, then the inequality*

$$\rho(\theta_t) - \rho^\star \leq \frac{\|\theta_0 - \theta^\star\|^2}{2\gamma t}, \tag{4.1}$$

*is satisfied for the sequence* $\{\theta_t\}_{t\geq 0}$ *generated by the recursion (3.1) where* $\theta^\star$ *is a minimum of* $\rho$*.*

***Proof*** See, e.g. Nesterov (2013). □

This rate in (4.1) is one of the most fundamental results in convex optimisation. Lemma 3 enables us to prove the following result for the MSE of the AIS estimator adapted using exact gradient descent in Eq. (3.2).

**Theorem 4** *Let Assumption 1 hold and construct the sequence* $(\theta_t)_{t\geq 1}$ *using recursion (3.1), where* $(q_{\theta_t})_{t\geq 1}$ *is the sequence of proposal distributions. Then, the inequality*

$$\mathbb{E}\left[\left((\varphi, \pi) - (\varphi, \pi_{\theta_t}^N)\right)^2\right] \leq \frac{c_\varphi\|\theta_0 - \theta^\star\|_2^2}{2\gamma t N} + \frac{c_\varphi\rho^\star}{N} \tag{4.2}$$

*is satisfied, where* $c_\varphi = 4\|\varphi\|_\infty^2$, $0 < \gamma \leq 1/L$ *and* $L$ *is the Lipschitz constant of the gradient* $\nabla\rho(\theta)$ *in Assumption 1.*

***Proof*** See Appendix A.3. □

***Remark 10*** Theorem 4 sheds light onto several facts. We first note that $\rho^\star$ in the error bound (4.2) can be interpreted as an indicator of the quality of the parametric proposal. We recall that $\rho^\star = 1$ when both $\pi$ and $q_\theta$ belong to the same exponential family. For this special case, Theorem 4 implies that

$$\lim_{t\to\infty} \left\| (\varphi,\pi) - (\varphi,\pi^N_{\theta_t}) \right\|_2 \leq \mathcal{O}\left(\frac{1}{\sqrt{N}}\right).$$

In other words, when the target and the proposal are both from the exponential family, this adaptation strategy is leading to an *asymptotically optimal* Monte Carlo estimator (optimal meaning that we attain the same rate as a Monte Carlo estimator with $N$ iid samples from $\pi$). On the other hand, when $\pi$ and $q_\theta$ do not belong to the same family, we obtain

$$\lim_{t\to\infty} \left\| (\varphi,\pi) - (\varphi,\pi^N_{\theta_t}) \right\|_2 \leq \mathcal{O}\left(\sqrt{\frac{\rho^\star}{N}}\right),$$

i.e. the $L_2$ rate is again asymptotically optimal, but the constant in the error bound is worse (bigger) by a factor $\sqrt{\rho^\star} > 1$. □

This bound shows that as $t \to \infty$, what we are left with is essentially the minimum attainable IS error for a given parametric family $\{q_\theta\}_{\theta\in\Theta}$. Intuitively, when the proposal $q_\theta$ is from a different parametric family than $\pi$, the gradient OAIS optimises the error bound in order to obtain the best possible proposal. In particular, the MSE has two components: First an $\mathcal{O}(1/tN)$ component which can be made to vanish over time by improving the proposal and a second $\mathcal{O}(1/N)$ component which is related to $\rho^\star$. The quantity $\rho^\star$ is related to the minimum $\chi^2$-divergence between the target and proposal. This means that the discrepancy between the target and *optimal* proposal (according to the $\chi^2$-divergence) can only be tackled by increasing $N$. This intuition is the same for the schemes we analyse in the next sections, although the rate with respect to the number of iterations necessarily worsens because of the uncertainty in the gradient estimators.

***Remark 11*** When $\gamma = 1/L$, Theorem 4 implies that if $t = \mathcal{O}(L/\rho^\star)$ and $N = \mathcal{O}(\rho^\star/\varepsilon)$, for some $\varepsilon > 0$, then we have

$$\mathbb{E}\left[\left((\varphi,\pi) - (\varphi,\pi^N_{\theta_t})\right)^2\right] \leq \mathcal{O}(\varepsilon).$$

We remark that once we choose the number of samples $N = \mathcal{O}(\rho^\star/\varepsilon)$, the number of iterations $t$ for adaptation is independent of $N$ and $\varepsilon$. □

***Remark 12*** One can use different maps $\mathcal{T}_t$ for optimisation. For example, one can use Nesterov's accelerated gradient descent (which has more parameters than just a step size), in which case, one could prove (by a similar argument) the inequality (Nesterov 2013)

$$\mathbb{E}\left[\left((\varphi,\pi) - (\varphi,\pi^N_{\theta_t})\right)^2\right] \leq \mathcal{O}\left(\frac{1}{t^2N} + \frac{\rho^\star}{N}\right).$$

This is an improved convergence rate, going from $\mathcal{O}(1/t)$ to $\mathcal{O}(1/t^2)$ in the first term of the bound. □

## 4.2 Convergence rate with averaged SGD iterates

While, for the purpose of analysis, it is convenient to assume that the minimization of $\rho(\theta)$ can be done deterministically, this is rarely the case in practice. The 'best' realistic case is that we can obtain an unbiased estimator of the gradient. Hence, we address this scenario, under the assumption that the actual gradient functions $\nabla\rho$ and $\nabla R$ are bounded in $\Theta$ (i.e. $\rho(\theta)$ is Lipschitz in $\Theta$).

**Assumption 2** The gradient functions $\nabla\rho(\theta)$ and $\nabla R(\theta)$ are bounded in $\Theta$. To be specific, there exist finite constants $G_\rho$ and $G_R$ such that

$$\sup_{\theta\in\Theta} \|\nabla\rho(\theta)\|_2 < G_\rho < \infty \quad \text{and}$$
$$\sup_{\theta\in\Theta} \|\nabla R(\theta)\|_2 < G_R < \infty.$$

We note that this is a mild assumption in the case of interest in this paper, where $\Theta \subset \mathbb{R}^{d_\theta}$ is assumed to be compact.

### 4.2.1 Normalised target

First, we assume that we can evaluate $\pi(x)$, which means that at iteration $t$, we can obtain an unbiased estimate of $\nabla\rho(\theta_{t-1})$, denoted $g_t$. We use the optimisation algorithms called *stochastic gradient* methods, which use stochastic and unbiased estimates of the gradients to optimise a given cost function (Robbins and Monro 1951). Particularly, we focus on optimised samplers using stochastic gradient descent (SGD) as an adaptation strategy.

We start proving that the stochastic gradient estimates $(g_t)_{t\geq 0}$ have a finite mean-squared error (MSE) w.r.t. the true gradients. To prove this result, we need an additional regularity condition.

**Assumption 3** The normalised target and proposal densities satisfy the inequality

$$\sup_{\theta\in\Theta} \mathbb{E}_{q_\theta}\left[\left|\frac{\pi^2(X)}{q^2_\theta(X)}\frac{\partial \log q_\theta}{\partial\theta_j}(X)\right|^2\right] =: D^j_\pi < \infty.$$

for $j = 1,\dots,d_\theta$. We denote $D_\pi := \max_{j\in\{1,\dots,d_\theta\}} D^j_\pi < \infty$.

***Remark 13*** Let us rewrite $D_\pi^j$ in Assumption 3 in terms of the weight function, namely

$$D_\pi^j = \sup_{\theta\in\Theta} \mathbb{E}_{q_\theta}\left[\left| w_\theta^2(X)\frac{\partial \log q_\theta}{\partial \theta_j}(X)\right|^2\right].$$

When $q_\theta(x)$ belongs to the exponential family, we readily obtain

$$D_\pi^j = \sup_{\theta\in\Theta} \mathbb{E}_{q_\theta}\left[w_\theta^4(X)\left(\frac{\partial A(\theta)}{\partial \theta_i} - T_i(X)\right)^2\right],$$

where $T_i(X)$ is the $i$-th sufficient statistic for $q_\theta(x)$. Let us construct a bounding function for the weights of the form

$$K(\theta) := \sup_{x\in\mathsf{X}} w_\theta(x).$$

If we choose the compact space $\Theta$ in such a way that $K(\theta)$ is bounded, then we readily have

$$\begin{aligned} D_\pi^j &\le \sup_{\theta\in\Theta} K^4(\theta)\mathbb{E}_{q_\theta}\left[\left(\frac{\partial A(\theta)}{\partial \theta_i} - T_i(X)\right)^2\right] \\ &\le \|K\|_\infty^4 \mathrm{Var}(T_i(X)), \end{aligned}$$

where we have used the fact that $\frac{\partial^m A(\theta)}{\partial \theta_i} = \mathbb{E}_{q_\theta}\left[T_i^m(X)\right]$. Therefore, if the weights remain bounded in $\Theta$, a sufficient condition for Assumption 3 to hold is that the sufficient statistics of the proposal distribution all have finite variances, i.e. $\max_{i\in\{1,\dots,d_\theta\}} T_i(X) < \infty$.
There are alternative conditions that, when satisfied, lead to Assumption 3 holding true. As an example, in Appendix A.4, we provide an alternative sufficient condition in terms of the function $\rho_2(\theta) := \mathbb{E}[w_\theta^4(X)]$.

Now, we show that when $g_t$ is an iid Monte Carlo estimator of $\nabla\rho$, we have the following finite-sample bound for the MSE.

**Lemma 4** *If Assumption 3 holds, the following inequality holds,*

$$\mathbb{E}[\|g_t - \nabla\rho(\theta_{t-1})\|_2^2] \le \frac{d_\theta c_\rho D_\pi}{N},$$

*where $d_\theta$ is the parameter dimension and $c_\rho$, $D_\pi < \infty$ are constant w.r.t. $N$.*

***Proof*** See Appendix A.5. □

In order to obtain convergence rates for the estimator $(\varphi, \pi_{\bar\theta_t}^N)$, we first recall a classical result for the SGD [see, e.g. Bubeck et al. (2015)].

**Lemma 5** *Let Assumptions 2 and 3 hold, apply recursion (3.4) and let $(g_t)_{t\ge0}$ be the stochastic gradient estimates in Lemma 4. If we choose the step-size sequence $\gamma_k = \alpha/\sqrt{k}$, $1 \le k \le t$, for any $\alpha > 0$, then*

$$\mathbb{E}[\rho(\bar\theta_t) - \rho^\star] \le \frac{\mathbb{E}\|\theta_0 - \theta^\star\|_2^2}{2\alpha\sqrt{t}} + \frac{\alpha d_\theta c_\rho D_\pi}{\sqrt{t}N} + \frac{\alpha G_\rho^2}{\sqrt{t}}, \tag{4.3}$$

*where $\bar\theta_t = \frac{1}{t}\sum_{k=0}^{t-1}\theta_k$.*

***Proof*** See Appendix A.6 for a self-contained proof. □

We can now state the first core result of the paper, which is the convergence rate for the AIS algorithm using a SGD adaptation of the parameter vectors $\theta_t$.

**Theorem 5** *Let Assumptions 2 and 3 hold, let the sequence $(\theta_t)_{t\ge1}$ be computed using (3.4) and construct the averaged iterates $\bar\theta_t = \frac{1}{t}\sum_{k=0}^{t-1}\theta_k$. Then, the sequence of proposal distributions $(q_{\bar\theta_t})_{t\ge1}$ satisfies the inequality*

$$\mathbb{E}\left[\left((\varphi,\pi) - (\varphi,\pi_{\bar\theta_t}^N)\right)^2\right] \le \frac{c_1}{\sqrt{t}N} + \frac{c_2}{\sqrt{t}N^2} + \frac{c_3}{\sqrt{t}N} + \frac{c_4}{N} \tag{4.4}$$

*for $t \ge 1$ and any $\varphi \in B(\mathsf{X})$, where*

$$\begin{aligned} c_1 &= \frac{c_\varphi \mathbb{E}\|\theta_0 - \theta^\star\|_2^2}{2\alpha}, \\ c_2 &= c_\varphi c_\rho \alpha d_\theta D_\pi, \\ c_3 &= c_\varphi \alpha G_\rho^2, \\ c_4 &= c_\varphi \rho^\star, \end{aligned}$$

*and $c_\varphi = 4\|\varphi\|_\infty^2$ are finite constants independent of $t$ and $N$.*

***Proof*** See Appendix A.7. □

***Remark 14*** Note that the expectation on the left-hand side of (4.4) is taken w.r.t. the distribution of the measure-valued random variable $\pi_{\bar\theta_t}^N$. □

Theorem 5 can be interpreted similarly to Theorem 4. One can see that the overall rate of the MSE bound is $\mathcal{O}\left(1/\sqrt{t}N + 1/N\right)$. This means that as $t \to \infty$, we are only left with a rate that is optimal for the AIS for a given parametric proposal family. In particular, again, $\rho^\star$ is related to the minimal $\chi^2$-divergence between the target $\pi$ and the parametric proposal $q_\theta$. When the proposal and the target are from the same family, we are back to the case $\rho^\star = 1$, thus the adaptation leads to the optimal Monte Carlo rate $\mathcal{O}(1/\sqrt{N})$ for the $L_2$ error within this setting as well.

#### 4.2.2 Self-normalised estimators

We have noted that it is possible to obtain an unbiased estimate of $\nabla\rho(\theta)$ when the normalised target $\pi(x)$ can be evaluated. However, if we can only evaluate the unnormalised density $\Pi(x)$ instead of $\pi(x)$ and use the self-normalized IS estimator, the estimate of $\nabla\rho(\theta)$ is no longer unbiased. We refer to Sec. 5 of Tadić and Doucet (2017) for stochastic optimisation with biased gradients for adaptive Monte Carlo, where the discussion revolves around minimizing the Kullback–Leibler divergence rather than the $\chi^2$-divergence. The results presented in Tadić and Doucet (2017), however, are asymptotic, while herein we are interested in finite-time bounds. Due to the structure of the AIS scheme, it is possible to avoid working with biased gradient estimators. In particular, we can implement the proposed AIS schemes using unbiased estimators of $\nabla R(\theta)$ instead of biased estimators of $\nabla\rho(\theta)$. Since optimizing the unnormalised function $R(\theta)$ leads to the same minima as optimizing the normalised function $\rho(\theta)$, we can simply use $\nabla R(\theta)$ for the adaptation in the self-normalised case.

Similar to the argument in Sect. 4.2.1, we first start the assumption below, which is the obvious counterpart of Assumption 3.

**Assumption 4** The unnormalized target $\Pi(x)$ and the proposal densities $q_\theta(x)$ satisfy the inequalities

$$\sup_{\theta\in\Theta}\mathbb{E}_{q_\theta}\left[\left|\frac{\Pi^2(X)}{q_\theta^2(X)}\frac{\partial\log q_\theta}{\partial\theta_j}(X)\right|^2\right] =: D_\Pi^j < \infty$$

for $j = 1,\dots,d_\theta$. We denote $D_\Pi := \frac{1}{d_\theta}\sum_{j=1}^{d_\theta} D_\Pi^j < \infty$.

Remark 13 holds directly for Assumption 4 as long as $Z_\pi < \infty$. Next, we prove an MSE bound for the stochastic gradients $(\tilde{g}_t)_{t\geq 0}$ employed in recursion (3.5), i.e. the unbiased stochastic estimates of $\nabla R(\theta)$.

**Lemma 6** *If Assumptions 2 and 4 hold, the inequality*

$$\mathbb{E}[\|\tilde{g}_t - \nabla R(\theta_{t-1})\|_2^2] \leq \frac{d_\theta c_R D_\Pi}{N},$$

*is satisfied, where $c_R$, $D_\Pi < \infty$ are constants w.r.t. of $N$.*

***Proof*** The proof is identical to the proof of Lemma 4. □

We can now obtain explicit rates for the convergence of the unnormalized function $R(\bar{\theta}_t)$, evaluated at the averaged iterates $\bar{\theta}_t$.

**Lemma 7** *If Assumptions 2 and 4 hold and the sequence $(\theta_t)_{t\geq 1}$ is computed via recursion (3.5), with step-sizes $\gamma_k = \beta/\sqrt{k}$ for $1 \leq k \leq t$ and $\beta > 0$, we obtain the inequality*

$$\mathbb{E}[R(\bar{\theta}_t) - R^\star] \leq \frac{\mathbb{E}\|\theta_0 - \theta^\star\|_2^2}{2\beta\sqrt{t}} + \frac{\beta d_\theta c_R D_\Pi}{\sqrt{t}N} + \frac{\beta G_R^2}{\sqrt{t}} \tag{4.5}$$

*where $c_R$, $D_\Pi < \infty$ are constants w.r.t. $t$ and $N$. Relationship 4.5 implies that*

$$\mathbb{E}[\rho(\bar{\theta}_t) - \rho^\star] \leq \frac{\mathbb{E}\|\theta_0 - \theta^\star\|_2^2}{2\beta Z_\pi^2\sqrt{t}} + \frac{\beta d_\theta c_R D_\Pi}{Z_\pi^2\sqrt{t}N} + \frac{\beta G_R^2}{Z_\pi^2\sqrt{t}}. \tag{4.6}$$

***Proof*** The proof of the rate in (4.5) is identical to the proof of Lemma 5. The rate in (4.6) follows by observing that $\rho(\theta) = R(\theta)/Z_\pi^2$ for every $\theta \in \Theta$. □

Finally, using Lemma 7, we can state our main result: an explicit error rate for the MSE of Algorithm 3 as a function of the number of iterations $t$ and the number of samples $N$.

**Theorem 6** *Let Assumptions 2 and 4 hold and let the sequence $(\theta_t)_{t\geq 1}$ be computed via recursion (3.5), with step-sizes $\gamma_k = \beta/\sqrt{k}$ for $1 \leq k \leq t$ and $\beta > 0$. We have the following inequality for the sequence of proposal distributions $(q_{\bar{\theta}_t})_{t\geq 1}$,*

$$\mathbb{E}\left[\left((\varphi,\pi) - (\varphi,\pi^N_{\bar{\theta}_t})\right)^2\right] \leq \frac{C_1}{\sqrt{t}N} + \frac{C_2}{\sqrt{t}N^2} + \frac{C_3}{\sqrt{t}N} + \frac{C_4}{N}, \tag{4.7}$$

*where*

$$C_1 = \frac{c_\varphi\mathbb{E}\|\theta_0 - \theta^\star\|_2^2}{2\beta Z_\pi^2},$$
$$C_2 = \frac{c_\varphi\beta c_R d_\theta D_\Pi}{Z_\pi^2},$$
$$C_3 = \frac{c_\varphi\beta G_R^2}{Z_\pi^2},$$
$$C_4 = c_\varphi\rho^\star,$$

*and $c_\varphi = 4\|\varphi\|_\infty^2$ are finite constants independent of $t$ and $N$.*

***Proof*** The proof follows from Lemma 7 and mimicking the exact same steps as in the proof of Theorem 5. □

***Remark 15*** Theorem 6, as in Remark 10, provides relevant insights regarding the performance of the stochastic gradient OAIS algorithm. In particular, for a general target $\pi$, we obtain

$$\lim_{t\to\infty} \left\| (\varphi,\pi) - (\varphi,\pi^N_{\bar{\theta}_t}) \right\|_2 = \mathcal{O}\left(\sqrt{\frac{\rho^\star}{N}}\right).$$

This result shows that the $L_2$ error is asymptotically optimal. As in previous cases, if the target $\pi$ is in the exponential family, then the asymptotic convergence rate is $\mathcal{O}(1/\sqrt{N})$ as $t \to \infty$. □

***Remark 16*** Theorem 6 also yields a practical heuristic to tune the step-size and the number of particles together. Assume that $0 < \beta < 1$ and let $N = 1/\beta$ (which we assume to be an integer without loss of generality). In this case, the rate (4.7) simplifies into

$$\mathbb{E}\left[\left((\varphi,\pi) - (\varphi,\pi^N_{\bar{\theta}_t})\right)^2\right] \leq \frac{c_\varphi \mathbb{E}\|\theta_0 - \theta^\star\|_2^2}{2Z_\pi^2\sqrt{t}} + \frac{c_\varphi \beta^3 c_R d_\theta D_\Pi}{Z_\pi^2\sqrt{t}} + \frac{c_\varphi \beta^2 G_R^2}{Z_\pi^2\sqrt{t}} + c_\varphi \rho^\star \beta$$

Now, if we let $t = \mathcal{O}(1/\beta^2)$, we readily obtain

$$\mathbb{E}\left[\left((\varphi,\pi) - (\varphi,\pi^N_{\bar{\theta}_t})\right)^2\right] \leq \mathcal{O}(\beta).$$

Therefore, one can control the error using the step-size of the optimisation scheme provided that other parameters of the algorithm are chosen accordingly. The same argument also holds for Theorem 5. □

***Remark 17*** It is not straightforward to compare the rates in inequality (4.7) (for the unnormalized target $\Pi(x)$) and inequality (4.4) (for the normalized target $\pi(x)$). Even if (4.7) may "look better" by a constant factor compared to the rate in (4.4), this is usually not the case. Indeed, the variance of the errors in the unnormalised gradient estimators is typically higher and this reflects on the variance of the moment estimators. Another way to look at this issue is to realise that, very often, $Z_\pi << 1$, which makes the bound in (4.7) much greater than the bound in (4.4).

Finally, we can adapt Theorem 2 to our case, providing a convergence rate of the bias of the importance sampler given by Algorithm 3.

**Theorem 7** *Under the setting of Theorem 6, we have*

$$\left|\mathbb{E}\left[(\varphi,\pi^N_{\bar{\theta}_t})\right] - (\varphi,\pi)\right| \leq \frac{3C_1}{\sqrt{t}N} + \frac{3C_2}{\sqrt{t}N^2} + \frac{3C_3}{\sqrt{t}N} + \frac{3C_4}{N}, \tag{4.8}$$

*where $C_1, C_2, C_3, C_4$ are finite constants given in Theorem 6 and independent of $t$ and $N$.*

***Proof*** The proof follows from Theorem 2 and mimicking the same proof technique used to prove Theorem 6. □

### 4.3 Convergence rate with vanilla SGD

The arguments of Section 4.2 can be carried over to the analysis of Algorithm 2, where the proposal functions $q_{\theta_t}(x)$ are constructed using the iterates $\theta_t$ rather than the averages $\bar{\theta}_t$. Unfortunately, achieving the optimal $\mathcal{O}(1/\sqrt{t})$ rate for the vanilla SGD is difficult in general. The best available rate without significant restrictions on the step-size is given by Shamir and Zhang (2013). In particular, we can adapt (Shamir and Zhang 2013, Theorem 2) to our setting in order to state the following lemma.

**Lemma 8** *Apply recursion (3.5) for the computation of the iterates $(\theta_t)_{t\geq 1}$, choose the step-sizes $\gamma_k = \beta/\sqrt{k}$ for $1 \leq k \leq t$, where $\beta > 0$, and let Assumptions 2 and 4 hold. Then, we have the inequality*

$$\mathbb{E}[R(\theta_t) - R^\star] \leq \left(\frac{D^2}{\beta\sqrt{t}} + \frac{\beta d_\theta c_R D_\Pi}{\sqrt{t}N} + \frac{\beta G_R^2}{\sqrt{t}}\right)(2 + \log t), \tag{4.9}$$

*where $D := \sup_{\theta,\theta'\in\Theta} \|\theta - \theta'\| < \infty$. This in turn implies that*

$$\mathbb{E}[\rho(\theta_t) - \rho^\star] \leq \left(\frac{D^2}{\beta\sqrt{t}} + \frac{\beta d_\theta c_R D_\Pi}{\sqrt{t}N} + \frac{\beta G_R^2}{\sqrt{t}}\right)\frac{(2 + \log t)}{Z_\pi^2}. \tag{4.10}$$

***Proof*** It is straightforward to prove this result using (Shamir and Zhang 2013, Theorem 2) and the proof of Lemma 5. □

The obtained rate is, in general, $\mathcal{O}\left(\frac{\log t}{\sqrt{t}}\right)$. This is known to be suboptimal, and it can be improved to the information-theoretical optimal $\mathcal{O}(1/\sqrt{t})$ rate by choosing a specific step-size scheduling, see, e.g. Jain et al. (2019). However, in this case, the scheduling of $(\gamma_t)_{t\geq 1}$ depends directly on the total number of iterates to be generated, in such a way that the error $\mathcal{O}(1/\sqrt{t})$ is guaranteed only for the *last* iterate, at the final time $t$.

We can extend Lemma 8 to obtain the following result.

**Theorem 8** *Apply recursion (3.5) for the computation of the iterates $(\theta_t)_{t\geq 1}$, choose the step-sizes $\gamma_k = \beta/\sqrt{k}$ for $1 \leq k \leq t$, where $\beta > 0$, and let Assumptions 2 and 4 hold. If we construct the sequence of proposal distributions $(q_{\theta_t})_{t\geq 1}$ be the sequence of proposal distributions, we obtain the following MSE bounds*

$$\mathbb{E}\left[\left((\varphi,\pi)-(\varphi,\pi_{\theta_t}^N)\right)^2\right] \leq \left(\frac{C_1}{\sqrt{t}N}+\frac{C_2}{\sqrt{t}N^2}+\frac{C_3}{\sqrt{t}N}\right)(2+\log t)+\frac{C_4}{N}, \tag{4.11}$$

*where*

$$C_1=\frac{c_\varphi D^2}{2\beta Z_\pi^2},$$
$$C_2=\frac{c_\varphi \beta c_R d_\theta D_\Pi}{Z_\pi^2},$$
$$C_3=\frac{c_\varphi \beta G_R^2}{Z_\pi^2},$$
$$C_4=c_\varphi \rho^\star,$$

*and* $c_\varphi = 4\|\varphi\|_\infty^2$ *are finite constants independent of* $t$ *and* $N$.

***Proof*** The proof follows from Lemma 8 with the exact same steps as in the proof of Theorem 5. □

Finally, it is also straightforward to adapt the bias result in Theorem 7 to this case, which results in the similar bound. We skip it for space reasons and also because it has the same form as in Theorem 7 with an extra $\log t$ factor.

## 5 Conclusions

We have presented and analysed *optimised* parametric adaptive importance samplers and provided non-asymptotic convergence bounds for the MSE of these samplers. Our results display the precise interplay between the number of iterations $t$ and the number of samples $N$. In particular, we have shown that the optimised samplers converge to an optimal proposal as $t \to \infty$, leading to an asymptotic rate of $\mathcal{O}(\sqrt{\rho^\star/N})$. This intuitively shows that the number of samples $N$ should be set in proportion to the minimum $\chi^2$-divergence between the target and the exponential family proposal, as we have shown that the adaptation (in the sense of minimising $\chi^2$-divergence or, equivalently, the variance of the weight function) cannot improve the error rate beyond $\mathcal{O}(\sqrt{\rho^\star/N})$. The error rates in this regime may be dominated by how close the target is to the exponential family.

Note that the algorithms we have analysed require constant computational load at each iteration and the computational load does not increase with $t$ as we do not reuse the samples in past iterations. Such schemes, however, can also be considered and analysed in the same manner. More specifically, in the present setup the computational cost of each iteration depends on the cost of evaluating $\Pi(x)$.

Our work opens up several other paths for research. One direction is to analyse the methods with more advanced optimisation algorithms. Another challenging direction is to consider more general proposals than the natural exponential family, which may lead to non-convex optimisation problems of adaptation. Analysing and providing guarantees for this general case would provide foundational insights for general adaptive importance sampling procedures. Also, as shown by Ryu (2016), similar theorems can also be proved for $\alpha$-divergences.

Another related piece of work arises from variational inference (Wainwright and Jordan 2008). In particular, Dieng et al. (2017) have recently considered performing variational inference by minimising the $\chi^2$-divergence, which is close to the setting in this paper. In particular, the variational approximation of the target distribution in the variational setting coincides with the proposal distribution we consider within the importance sampling context in this paper. This also implies that our results may be used to obtain finite-time guarantees for the expectations estimated using the variational approximations of target distributions.

Finally, the adaptation procedure can be modified to handle the non-convex case as well. In particular, the SGD step can be converted into a stochastic gradient Langevin dynamics (SGLD) step. The SGLD method can be used as a global optimiser when $\rho$ and $R$ are non-convex and a global convergence rate can be obtained using the standard SGLD results, see, e.g. Raginsky et al. (2017); Zhang et al. (2019). Global convergence results for other adaptation schemes such as stochastic gradient Hamiltonian Monte Carlo (SGHMC) can also be achieved using results from nonconvex optimisation literature, see, e.g. Akyildiz and Sabanis (2020).

# A Appendix

## A.1 Proof of Theorem 1

We first note the following inequalities,

$$\begin{aligned}
|(\varphi,\pi)-(\varphi,\pi_\theta^N)| &= \left|\frac{(\varphi W_\theta,q_\theta)}{(W_\theta,q_\theta)}-\frac{(\varphi W_\theta,q_\theta^N)}{(W_\theta,q_\theta^N)}\right| \\
&\le \frac{\left|(\varphi W_\theta,q_\theta)-(\varphi W_\theta,q_\theta^N)\right|}{|(W_\theta,q_\theta)|} \\
&\quad + |(\varphi W_\theta,q_\theta^N)|\left|\frac{1}{(W_\theta,q_\theta)}-\frac{1}{(W_\theta,q_\theta^N)}\right| \\
&= \frac{\left|(\varphi W_\theta,q_\theta)-(\varphi W_\theta,q_\theta^N)\right|}{|(W_\theta,q_\theta)|} \\
&\quad + \|\varphi\|_\infty|(W_\theta,q_\theta^N)|\left|\frac{(W_\theta,q_\theta^N)-(W_\theta,q_\theta)}{(W_\theta,q_\theta)(W_\theta,q_\theta^N)}\right| \\
&= \frac{\left|(\varphi W_\theta,q_\theta)-(\varphi W_\theta,q_\theta^N)\right|}{(W_\theta,q_\theta)} \\
&\quad + \frac{\|\varphi\|_\infty|(W_\theta,q_\theta^N)-(W_\theta,q_\theta)|}{(W_\theta,q_\theta)}.
\end{aligned}$$

We take squares of both sides and apply the inequality $(a+b)^2 \le 2(a^2+b^2)$ to further bound the rhs,

$$\begin{aligned}
|(\varphi,\pi)-(\varphi,\pi_\theta^N)|^2 &\le 2\frac{\left|(\varphi W_\theta,q_\theta)-(\varphi W_\theta,q_\theta^N)\right|^2}{(W_\theta,q_\theta)^2} \\
&\quad + 2\frac{\|\varphi\|_\infty^2|(W_\theta,q_\theta^N)-(W_\theta,q_\theta)|^2}{(W_\theta,q_\theta)^2}
\end{aligned}$$

We now take the expectation of both sides,

$$\begin{aligned}
\mathbb{E}\left[\left((\varphi,\pi)-(\varphi,\pi_\theta^N)\right)^2\right] &\le \frac{2\mathbb{E}\left[\left((\varphi W_\theta,q_\theta)-(\varphi W_\theta,q_\theta^N)\right)^2\right]}{(W_\theta,q_\theta)^2} \\
&\quad + \frac{2\|\varphi\|_\infty^2\mathbb{E}\left[\left((W_\theta,q_\theta^N)-(W_\theta,q_\theta)\right)^2\right]}{(W_\theta,q_\theta)^2}.
\end{aligned}$$

Note that, both terms in the right-hand side are perfect Monte Carlo estimates of the integrals. Bounding the MSE of these integrals yields

$$\begin{aligned}
\mathbb{E}\left[\left((\varphi,\pi)-(\varphi,\pi_\theta^N)\right)^2\right] &\le \frac{2}{N}\frac{(\varphi^2W_\theta^2,q_\theta)-(\varphi W_\theta,q_\theta)^2}{(W_\theta,q_\theta)^2} \\
&\quad + \frac{2\|\varphi\|_\infty^2}{N}\frac{(W_\theta^2,q_\theta)-(W_\theta,q_\theta)^2}{(W_\theta,q_\theta)^2}, \\
&\le \frac{2\|\varphi\|_\infty^2}{N}\frac{(W_\theta^2,q_\theta)}{(W_\theta,q_\theta)^2} \\
&\quad + \frac{2\|\varphi\|_\infty^2}{N}\frac{(W_\theta^2,q_\theta)-(W_\theta,q_\theta)^2}{(W_\theta,q_\theta)^2}.
\end{aligned}$$

Therefore, we can straightforwardly write,

$$\mathbb{E}\left[\left((\varphi,\pi)-(\varphi,\pi_\theta^N)\right)^2\right] \le \frac{4\|\varphi\|_\infty^2}{(W_\theta,q_\theta)^2}\frac{(W_\theta^2,q_\theta)}{N}.$$

Now, it remains to show the relation of the bound to $\chi^2$ divergence. Note that

$$\begin{aligned}
\frac{(W_\theta^2,q_\theta)}{(W_\theta,q_\theta)^2} &= \frac{\int\frac{\Pi^2(x)}{q_\theta^2(x)}q_\theta(x)\mathrm{d}x}{\left(\int\frac{\Pi(x)}{q_\theta(x)}q_\theta(x)\mathrm{d}x\right)^2} \\
&= \frac{Z^2\int\frac{\pi^2(x)}{q_\theta^2(x)}q_\theta(x)\mathrm{d}x}{Z^2\left(\int\pi\mathrm{d}x\right)^2} \\
&= \mathbb{E}_{q_\theta}\left[\frac{\pi^2(X)}{q_\theta^2(X)}\right] := \rho(\theta).
\end{aligned}$$

Note that $\rho$ is not exactly $\chi^2$ divergence, which is defined as $\rho-1$. Plugging everything into our bound, we have the result,

$$\mathbb{E}\left[\left((\varphi,\pi)-(\varphi,\pi_\theta^N)\right)^2\right] \le \frac{4\|\varphi\|_\infty^2\rho(\theta)}{N}.$$

□

## A.2 Proof of Lemma 1

We adapt this proof from Ryu and Boyd (2014) by following the same steps. We first show that $A(\theta)$ is convex by first showing that $\exp(A(\theta))$ is convex. Choose $0<\eta<1$ and using Hölder's inequality,

$$\begin{aligned}
\exp(A(\eta\theta_1+(1-\eta)\theta_2)) &= \int\exp((\eta\theta_1+(1-\eta)\theta_2)^\top T(x))h(x)\mathrm{d}x \\
&= \int\left(\exp(\theta_1^\top T(x))h(x)\right)^\eta\left(\exp(\theta_2^\top T(x))h(x)\right)^{1-\eta}\mathrm{d}x \\
&\le \left(\int\exp(\theta_1^\top T(x))h(x)\mathrm{d}x\right)^\eta\left(\int\exp(\theta_2^\top T(x))h(x)\mathrm{d}x\right)^{1-\eta}.
\end{aligned}$$

Taking log of both sides yields

$$A(\eta\theta_1+(1-\eta)\theta_2) \le \eta A(\theta_1)+(1-\eta)A(\theta_2),$$

which shows the convexity of $A(\theta)$. Note that $A(\theta)-\theta^\top T(x)$ is convex in $\theta$ since it is a sum of a convex and a linear function of $\theta$. Since exp is an increasing convex function and the composition of convex functions is convex, $M(\theta,x) := \exp(A(\theta)-\theta^\top T(x))$ is convex in $\theta$. Finally, we prove that $\rho(\theta)$ is convex. First let us write it as

$$\rho(\theta) = \int\frac{\pi^2(x)}{q_\theta(x)^2}q_\theta(x)\mathrm{d}x = \int\frac{\pi^2(x)}{h(x)}M(\theta,x)\mathrm{d}x.$$

Then, we have the following sequence of inequalities

$$\begin{aligned}
\rho(\eta\theta_1+(1-\eta)\theta_2) &= \int\frac{\pi^2(x)}{h(x)}M(\eta\theta_1+(1-\eta)\theta_2,x)\mathrm{d}x \\
&\le \int\frac{\pi^2(x)}{h(x)}(\eta M(\theta_1,x)+(1-\eta)M(\theta_2,x))\mathrm{d}x
\end{aligned}$$

$$
\begin{aligned}
&= \eta \int \frac{\pi^2(x)}{h(x)} M(\theta_1, x)\mathrm{d}x \\
&\quad + (1-\eta)\int \frac{\pi^2(x)}{h(x)} M(\theta_2, x)\mathrm{d}x \\
&= \eta\rho(\theta_1) + (1-\eta)\rho(\theta_2),
\end{aligned}
$$

which concludes the claim. □

### A.3 Proof of Theorem 4

First note that using Theorem 3, we have

$$
\begin{aligned}
\mathbb{E}\left[\left((\varphi,\pi) - (\varphi,\pi^N_{\theta_t})\right)^2\right] &\leq \frac{c_\varphi \rho(\theta_t)}{N}, \\
&= \frac{c_\varphi(\rho(\theta_t) - \rho^\star)}{N} + \frac{c_\varphi \rho^\star}{N}, \\
&\leq \frac{c_\varphi \|\theta_0 - \theta^\star\|^2}{2\gamma t N} + \frac{c_\varphi \rho^\star}{N},
\end{aligned}
$$

where the last inequality follows from Lemma 3. □

### A.4 A sufficient condition for Assumption 3 to hold

Recall that we have defined $\rho_2(\theta) = \mathbb{E}[w_\theta^4(X)] = \mathbb{E}\left[\frac{\pi^4(X)}{q_\theta^4(X)}\right]$ and $q_\theta(x) = \exp\left\{\left(\theta^\top T(x) - A(\theta)\right)h(x)\right\}$ whenever $q_\theta(x)$ belongs to the exponential family. We have the following result.

**Proposition 1** *Let the $\rho_2$ be Lipschitz with Lipschitz derivatives, let $q_\theta(x)$ belong to the exponential family and let $\Theta$ be compact. If the sufficient statistics $T(X)$ of the distribution $q_\theta$ all have finite variances, i.e.*

$$
\max_{i=1,\ldots,d_\theta} Var(T_i(X)) < \infty,
$$

*then Assumption 3 holds.*

***Proof*** Using the fact that

$$
\frac{\partial q_\theta(x)}{\partial \theta_i} = q_\theta(x)\frac{\partial \log q_\theta(x)}{\partial \theta_i}
$$

one can readily calculate the second order derivatives of $\rho_2(\theta)$. In particular,

$$
\begin{aligned}
\frac{\partial^2 \rho_2(\theta)}{\partial \theta_i^2} &= 9\mathbb{E}\left[w_\theta^4(X)\left(T_i(X) - \frac{\partial A(\theta)}{\partial \theta_i}\right)^2\right] \\
&\quad + 3\mathbb{E}\left[w_\theta^4(X)\right]\frac{\partial^2 A(\theta)}{\partial \theta_i^2} < \infty, \qquad \text{(A.1)}
\end{aligned}
$$

where the inequality holds because $\rho_2(\theta)$ has Lipschitz derivatives in $\Theta$. However, $\frac{\partial^2 A(\theta)}{\partial \theta_i^2} = \mathrm{Var}(T_i(X))$ and, by assumption, $\max_i \mathrm{Var}(T_i(X)) < \infty$. Moreover, $\mathbb{E}\left[w_\theta^4(X)\right] = \rho_2(\theta) < \infty$ because $\rho_2(\theta)$ is Lipschitz and the parameter space $\Theta$ is compact. Therefore, it follows that

$$
\mathbb{E}\left[w_\theta^4(X)\left(T_i(X) - \frac{\partial A(\theta)}{\partial \theta_i}\right)^2\right] < \infty
$$

and Assumption 3 holds. □

### A.5 Proof of Lemma 4

We first note that the exact gradient can be written as

$$
\begin{aligned}
\nabla_\theta \mathbb{E}_{q_\theta}\left[\frac{\pi^2(X)}{q_\theta^2(X)}\right] &= \nabla_\theta \int \frac{\pi^2(x)}{q_\theta(x)}\mathrm{d}x \\
&= -\int \frac{\pi^2(x)}{q_\theta^2(x)}\nabla_\theta \log q_\theta(x) q_\theta(x)\mathrm{d}x.
\end{aligned}
$$

Now, note that,

$$
\nabla_\theta \log q_\theta(x) = \begin{bmatrix} \frac{\partial \log q_\theta}{\partial \theta_1} \\ \frac{\partial \log q_\theta}{\partial \theta_2} \\ \vdots \\ \frac{\partial \log q_\theta}{\partial \theta_{d_\theta}} \end{bmatrix}.
$$

Given the samples $x_t^{(i)} \sim q_{\theta_{t-1}}$ for $i = 1, \ldots, N$ to estimate the gradient, we can write the mean-squared error $\mathbb{E}\|g_t - \nabla\rho(\theta_{t-1})\|_2^2$ as

$$
\begin{aligned}
&\mathbb{E}\left[\left\| \frac{1}{N}\sum_{i=1}^N \frac{\pi^2(x_t^{(i)})}{q_{\theta_{t-1}}^2(x_t^{(i)})}\nabla_\theta \log q_{\theta_{t-1}}(x_t^{(i)}) \right.\right. \\
&\quad \left.\left. - \int \frac{\pi^2(x)}{q_\theta^2(x)}\nabla_\theta \log q_{\theta_{t-1}}(x) q_{\theta_{t-1}}(x)\mathrm{d}x \right\|_2^2\right] \\
&= \sum_{j=1}^{d_\theta} \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \frac{\pi^2(x_t^{(i)})}{q_{\theta_{t-1}}^2(x_t^{(i)})}\frac{\partial \log q_{\theta_{t-1}}}{\partial \theta_j}(x_t^{(i)}) \right.\right. \\
&\quad \left.\left. - \int \frac{\pi^2(x)}{q_{\theta_{t-1}}^2(x)}\frac{\partial \log q_{\theta_{t-1}}}{\partial \theta_j}(x) q_{\theta_{t-1}}(x)\mathrm{d}x\right)^2\right].
\end{aligned}
$$

Now the expectation is a standard Monte Carlo error for the test function,

$$
\varphi_j(x) = \frac{\pi^2(x)}{q_\theta^2(x)}\frac{\partial \log q_\theta}{\partial \theta_j}(x).
$$

Assumption 3 together with Lemma A.1 in Crisan and Míguez (2014) yields

$$
\mathbb{E}[\|g_t - \nabla\rho(\theta)\|_2^2] \leq \frac{d_\theta c_\rho D_\pi}{N},
$$

where $c_\rho < \infty$ and $D_\pi = \max_{j\in\{1,\ldots,d_\theta\}} D_\pi^j < \infty$ are constants independent of $N$. □

### A.6 Proof of Lemma 5

Since projections reduce distances, we have,

$$\begin{aligned}\|\theta_k - \theta^\star\|_2^2 &\le \|\theta_{k-1} - \gamma_k g_k - \theta^\star\|_2^2 \\ &= \|\theta_{k-1} - \theta^\star\|_2^2 - 2\gamma_k g_k^\top(\theta_{k-1} - \theta^\star) + \gamma_k^2\|g_k\|_2^2.\end{aligned}$$

Let $\mathcal{F}_{k-1} = \sigma(\theta_0, \ldots, \theta_{k-1}, g_1, \ldots, g_{k-1})$ be the $\sigma$-algebra generated by random variables $\theta_0, \ldots, \theta_{k-1}, g_1, \ldots, g_{k-1}$ and take the conditional expectations with respect to $\mathcal{F}_{k-1}$

$$\begin{aligned}\mathbb{E}\left[\|\theta_k - \theta^\star\|_2^2|\mathcal{F}_{k-1}\right] &\le \|\theta_{k-1} - \theta^\star\|_2^2 - 2\gamma_k\nabla\rho(\theta_{k-1})^\top(\theta_{k-1} - \theta^\star) \\ &\quad + \gamma_k^2\mathbb{E}\left[\|g_k\|_2^2|\mathcal{F}_{k-1}\right].\end{aligned}$$

Next, using the convexity of $\rho$ yields

$$\begin{aligned}&\mathbb{E}\left[\|\theta_k - \theta^\star\|_2^2|\mathcal{F}_{k-1}\right] \\ &\le \|\theta_{k-1} - \theta^\star\|_2^2 - 2\gamma_k[\rho(\theta_{k-1}) - \rho(\theta^\star)] \\ &\quad + \gamma_k^2\mathbb{E}\left[\|g_k\|_2^2|\mathcal{F}_{k-1}\right].\end{aligned}$$

Finally, we take unconditional expectations of both sides,

$$\begin{aligned}\mathbb{E}\|\theta_k - \theta^\star\|_2^2 &\le \mathbb{E}\|\theta_{k-1} - \theta^\star\|_2^2 \\ &\quad - 2\gamma_k\mathbb{E}[(\rho(\theta_{k-1}) - \rho(\theta^\star)] + \gamma_k^2\mathbb{E}\|g_k\|_2^2.\end{aligned}$$

With rearranging, using $\mathbb{E}\|g_k - \nabla\rho(\theta_{k-1})\|_2^2 = \mathbb{E}\|g_k\|^2 - \|\nabla\rho(\theta_{k-1})\|^2$ and invoking Assumption 2, we arrive at

$$\begin{aligned}\mathbb{E}[\rho(\theta_{k-1}) - \rho(\theta^\star)] &\le \frac{\mathbb{E}\|\theta_{k-1} - \theta^\star\|_2^2 - \mathbb{E}\|\theta_k - \theta^\star\|_2^2}{2\gamma_k} \\ &\quad + \frac{\gamma_k(\sigma_\rho^2 + G_\rho^2)}{2}.\end{aligned}$$

where $\sigma_\rho^2 = d_\theta D_\pi / N$ as given in Lemma 4. Now summing both sides from $k = 1$ to $t$ and dividing both sides by $t$,

$$\begin{aligned}\mathbb{E}[\rho(\bar\theta_t) - \rho(\theta^\star)] &\le \frac{1}{t}\sum_{k=1}^{t}\mathbb{E}[\rho(\theta_{k-1}) - \rho(\theta^\star)] \le \frac{\mathbb{E}\|\theta_0 - \theta^\star\|_2^2}{2\gamma_t t} \\ &\quad + \sum_{k=1}^{t}\frac{\gamma_k(\sigma_\rho^2 + G_\rho^2)}{2t},\end{aligned}$$

since $\frac{1}{\gamma_k} \le \frac{1}{\gamma_t}$ for all $k \le t$. Substituting $\gamma_k = \alpha/\sqrt{k}$ and noting that

$$\sum_{k=1}^{t}\frac{1}{\sqrt{k}} \le \int_0^t \frac{1}{\sqrt{\tau}}\mathrm{d}\tau = 2\sqrt{t},$$

we arrive at

$$\mathbb{E}[\rho(\bar\theta_t) - \rho(\theta^\star)] \le \frac{\mathbb{E}\|\theta_0 - \theta^\star\|_2^2}{2\alpha\sqrt{t}} + \frac{\alpha(\sigma_\rho^2 + G_\rho^2)}{\sqrt{t}},$$

where $\bar\theta_t = \frac{1}{t}\sum_{k=0}^{t-1}\theta_k$. □

### A.7 Proof of Theorem 5

Let $\mathcal{F}_{t-1} = \sigma(\theta_0, \ldots, \theta_{t-1}, g_1, \ldots, g_{t-1})$ be the $\sigma$-algebra generated by the random variables $\theta_0, \ldots, \theta_{t-1}, g_1, \ldots, g_{t-1}$. Then

$$\begin{aligned}\mathbb{E}\left[\left((\varphi,\pi) - (\varphi,\pi_{\bar\theta_t}^N)\right)^2\middle|\mathcal{F}_{t-1}\right] &\le \frac{c_\varphi\rho(\bar\theta_t)}{N} \\ &= \frac{c_\varphi(\rho(\bar\theta_t) - \rho^\star)}{N} + \frac{c_\varphi\rho^\star}{N},\end{aligned}$$

where $\bar\theta_t = \frac{1}{t}\sum_{k=0}^{t-1}\theta_k$ is an $\mathcal{F}_{t-1}$-measurable random variable. Now if we take unconditional expectations of both sides,

$$\mathbb{E}\left[\left((\varphi,\pi) - (\varphi,\pi_{\bar\theta_t}^N)\right)^2\right] \le \frac{c_\varphi\mathbb{E}\left[(\rho(\bar\theta_t) - \rho^\star)\right]}{N} + \frac{c_\varphi\rho^\star}{N}.$$

The result follows from applying Lemma 5 for $\mathbb{E}\left[(\rho(\bar\theta_t) - \rho^\star)\right]$. □

## References


Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., Stuart, A.: Importance sampling: intrinsic dimension and computational cost. Stat. Sci. **32**(3), 405–431 (2017)

Akyildiz, ÖD., Sabanis, S.: Nonasymptotic analysis of Stochastic Gradient Hamiltonian Monte Carlo under local conditions for nonconvex optimization. (2020). arXiv preprint arXiv:2002.05465

Arouna, B.: Adaptative monte carlo method, a variance reduction technique. Monte Carlo Methods Appl. **10**(1), 1–24 (2004a)

Arouna, B.: Robbins-Monro algorithms and variance reduction in finance. J. Comput. Finance **7**(2), 35–62 (2004b)

Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. (2016). arXiv:1606.04838

Bubeck, S., et al.: Convex optimization: algorithms and complexity. Found. Trends® Mach. Learn. **8**(3–4), 231–357 (2015)

Bugallo, M.F., Martino, L., Corander, J.: Adaptive importance sampling in signal processing. Digit. Signal Proc. **47**, 36–49 (2015)

Bugallo, M.F., Elvira, V., Martino, L., Luengo, D., Miguez, J., Djuric, P.M.: Adaptive Importance Sampling: The past, the present, and the future. IEEE Signal Process. Mag. **34**(4), 60–79 (2017)

Cappé, O., Guillin, A., Marin, J.M., Robert, C.P.: Population Monte Carlo. J. Comput. Graph. Stat. **13**(4), 907–929 (2004)

Cappé, O., Douc, R., Guillin, A., Marin, J.M., Robert, C.P.: Adaptive importance sampling in general mixture classes. Stat. Comput. **18**(4), 447–459 (2008)

Chatterjee, S., Diaconis, P., et al.: The sample size required in importance sampling. Ann. Appl. Probab. **28**(2), 1099–1135 (2018)

Crisan, D., Míguez, J.: Particle-kernel estimation of the filter density in state-space models. Bernoulli **20**(4), 1879–1929 (2014)

Dieng, A.B., Tran, D., Ranganath, R., Paisley, J., Blei, D.: Variational inference via $\chi$-upper bound minimization. In: Advances in Neural Information Processing Systems, pp 2732–2741 (2017)

Douc, R., Guillin, A., Marin, J.M., Robert, C.P.: Convergence of adaptive mixtures of importance sampling schemes. Ann. Stat. **35**(1), 420–448 (2007)

Jain, P., Nagaraj, D., Netrapalli, P.: Making the Last Iterate of SGD Information Theoretically Optimal. In: Conference on Learning Theory, pp. 1752–1755 (2019)

Kappen, H.J., Ruiz, H.C.: Adaptive importance sampling for control and inference. J. Stat. Phys. **162**(5), 1244–1266 (2016)

Kawai, R.: Adaptive monte carlo variance reduction for lévy processes with two-time-scale stochastic approximation. Methodol. Comput. Appl. Probab. **10**(2), 199–223 (2008)

Kawai, R.: Acceleration on adaptive importance sampling with sample average approximation. SIAM J. Sci. Comput. **39**(4), A1586–A1615 (2017)

Kawai, R.: Optimizing adaptive importance sampling by stochastic approximation. SIAM J. Sci. Comput. **40**(4), A2774–A2800 (2018)

Lapeyre, B., Lelong, J.: A framework for adaptive monte carlo procedures. Monte Carlo Methods Appl. **17**(1), 77–98 (2011)

Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer, Berlin (2013)

Raginsky, M., Rakhlin, A., Telgarsky, M.: Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In: Conference on Learning Theory, pp. 1674–1703 (2017)

Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**, 400–407 (1951)

Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Wiley, New York (2004)

Ryu, E.K.: Convex optimization for Monte Carlo: Stochastic optimization for importance sampling. PhD thesis, Stanford University (2016)

Ryu, E.K., Boyd, S.P. Adaptive importance sampling via stochastic convex programming. (2014). arXiv:1412.4845

Sanz-Alonso, D.: Importance sampling and necessary sample size: an information theory approach. SIAM/ASA J. Uncertain. Quantif. **6**(2), 867–879 (2018)

Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Program. **162**(1–2), 83–112 (2017)

Shamir, O., Zhang, T.: Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In: International Conference on Machine Learning, pp 71–79 (2013)

Tadić, V.B., Doucet, A.: Asymptotic bias of stochastic gradient search. Ann. Appl. Probab. **27**(6), 3255–3304 (2017)

Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Found. Trends® Mach. Learn. **1**(1–2), 1–305 (2008)

Zhang, Y., Akyildiz, ÖD., Damoulas, T., Sabanis, S.: Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization. (2019). arXiv preprint arXiv:1910.02008