On importance-weighted autoencoders

Anonymous authors
Paper under double-blind review

Abstract

The *importance weighted autoencoder (IWAE)* (Burda et al., 2016) is a popular variational-inference method which achieves a tighter evidence bound (and hence a lower bias) than standard variational autoencoders by optimising a *multi-sample objective,* i.e. an objective that is expressible as an integral over $K > 1$ Monte Carlo samples. Unfortunately, IWAE crucially relies on the availability of reparametrisations and even if these exist, the multi-sample objective leads to inference-network gradients which break down as K is increased (Rainforth et al., 2018). This breakdown can only be circumvented by removing high-variance score-function terms, either by heuristically ignoring them (which yields the *'sticking-the-landing' IWAE (IWAE-STL)* gradient from Roeder et al. (2017)) or through an identity from Tucker et al. (2019) (which yields the *'doubly-reparametrised' IWAE (IWAE-DREG)* gradient). In this work, we argue that directly optimising the proposal distribution in importance sampling as in the *reweighted wake-sleep (RWS)* algorithm from Bornschein & Bengio (2015) is preferable to optimising IWAE-type multi-sample objectives. To formalise this argument, we introduce an adaptive-importance sampling framework termed *adaptive importance sampling for learning (AISLE)* which slightly generalises the RWS algorithm. We then show that AISLE admits IWAE-STL and IWAE-DREG (i.e. the IWAE-gradients which avoid breakdown) as special cases.

1 Introduction

1.1 Problem statement

Let x be some observation and let z be some latent variable taking values in some space Z. These are modeled via the *generative model* $p_\theta(z, x) = p_\theta(z)p_\theta(x|z)$ which gives rise to the marginal likelihood $p_\theta(x) = \int_{\mathsf{Z}} p_\theta(z, x)\, \mathrm{d}z$ of the model parameters θ. The latter may also be viewed as the evidence for the model parametrised by a particular value of θ. In this work, we analyse algorithms for *variational inference,* i.e. algorithms which aim to

1. learn the generative model, i.e. find a value $\theta^\star$ which is approximately equal to the *maximum-likelihood estimate (MLE)* $\theta^{\mathrm{ML}} := \arg\max_\theta p_\theta(x)$;
2. construct a tractable *variational approximation* $q_{\phi,x}(z)$ of $p_\theta(z|x) = p_\theta(z, x)/p_\theta(x)$, i.e. find the value $\phi^\star$ such that $q_{\phi^\star,x}(z)$ is as close as possible to $p_\theta(z|x)$ in some suitable sense.

A few comments about this setting are in order. Firstly, as is common in the literature, we restrict our presentation to a single latent representation–observation pair (z, x) to avoid notational clutter – the extension to multiple independent observations is straightforward. Secondly, we assume that no parameters are shared between the generative model $p_\theta(z, x)$ and the variational approximation $q_{\phi,x}(z)$. This is common in neural-network applications but could be relaxed. Thirdly, our setting is general enough to cover amortised inference which is why we often refer to ϕ as the parameters of an *inference network.*

In recent years, two classes of stochastic-gradient ascent algorithms for optimising (θ, ϕ) – which employ $K \geq 1$ Monte Carlo samples ('particles') to reduce errors – have been proposed.

- **IWAE.** The *importance weighted autoencoder (IWAE)* (Burda et al., 2016) optimises a joint objective for θ and ϕ (which is 'biased' for θ though optimising ϕ or increasing K decreases this bias) whose gradients are unbiasedly approximated via the Monte Carlo method. Unfortunately, as this *multi-sample* objective is expressible as an integral on a K-dimensional space, the signal-to-noise ratio of the IWAE ϕ-gradient vanishes as K grows (Rainforth et al., 2018). Two modified IWAE ϕ-gradients avoid this breakdown by removing high-variance 'score-function' terms:
  - **IWAE-STL.** The *'sticking-the-landing' IWAE (IWAE-STL)* ϕ-gradient (Roeder et al., 2017) heuristically drops the problematic score-function terms from the IWAE ϕ-gradient. This induces bias for the IWAE objective.
  - **IWAE-DREG.** The *'doubly-reparametrised' IWAE (IWAE-DREG)* ϕ-gradient (Tucker et al., 2019) unbiasedly removes the problematic score-function terms from the IWAE ϕ-gradient using a formal identity.
- **RWS.** The *reweighted wake-sleep (RWS)* algorithm (Bornschein & Bengio, 2015) optimises two separate but 'unbiased' objectives for θ and ϕ. Its gradients are approximated by self-normalised importance sampling with K particles which induces bias (though again, optimising ϕ or increasing K decreases this bias). RWS can be viewed as an adaptive importance-sampling approach which iteratively improves its proposal distribution while simultaneously optimising θ via stochastic approximation. Crucially, RWS is *not* a multi-sample objective approach and hence does not require continuous reparametrisations nor do its ϕ-gradients suffer from the breakdown highlighted in Rainforth et al. (2018).

Of these two methods, the IWAE is the most popular and Tucker et al. (2019) demonstrated empirically that RWS can break down, conjecturing that this is due to the fact that RWS does not optimise a joint objective (for θ and ϕ). Meanwhile, the IWAE-STL gradient performed consistently well despite lacking a firm theoretical footing. Yet, IWAE suffers from the above-mentioned ϕ-gradient breakdown and exhibited inferior empirical performance to RWS in some scenarios (Le et al., 2019). Thus, it is not clear whether the multi-sample objective approach of IWAE or the adaptive importance-sampling approach of RWS is preferable.

In this work, we argue that the adaptive importance-sampling paradigm of RWS is preferable to the multi-sample objective paradigm of IWAEs. This is because (a) the multi-sample objective crucially requires reparametrisations and, even if these are available, leads to the ϕ-gradient breakdown, (b) modifications of the IWAE ϕ-gradient which avoid this breakdown (i.e. IWAE-STL and IWAE-DREG) can be justified in a more principled manner by taking an RWS-type adaptive importance-sampling view.

To formalise these arguments, we slightly generalise the RWS algorithm to obtain a generic adaptive importance-sampling framework for variational inference which we term *adaptive importance sampling for learning (AISLE)* for ease of reference. We then show that AISLE admits not only RWS but also the IWAE-DREG and IWAE-STL gradients as special cases.

1.2 Contributions

Importance sampling as well as the IWAE and RWS algorithms are reviewed in Section 2. Novel material is presented in Section 3, where we we introduce the AISLE-framework:

- In Subsection 3.3, we show that AISLE admits RWS as a special case. In addition, we prove that the IWAE-STL gradient is in turn recovered as a special case of RWS (and hence of AISLE) via a principled and novel application of the 'double-reparametrisation' identity from Tucker et al. (2019). This indicates that the breakdown of RWS observed in Tucker et al. (2019) may not be due to its lack of a joint objective as previously conjectured (because IWAE-STL avoided this breakdown). Our work also provides a theoretical foundation for IWAE-STL which was hitherto only heuristically justified as a biased IWAE gradient.
- In Subsection 3.4, we prove that AISLE also admits the IWAE-DREG gradient as a special case. Our derivation also makes it clear that the learning rate should be scaled as $\mathcal{O}(K)$ for the IWAE ϕ-gradient (and its modified version IWAE-DREG)

unless the gradients are normalised as implicitly done by popular optimisers such as ADAM (Kingma & Ba, 2015). In contrast, the scaling of the learning rate for AISLE is independent of K.

- In the supplementary materials, we provide some insight into the impact of the self-normalisation bias on some of the importance-sampling based gradient approximations (Appendix A) and empirically compare all algorithms discussed in this work (Appendix B).

We stress that the point of our work is not to derive new algorithms nor to establish which of the various special cases of AISLE is preferable. Indeed, while we compare all algorithms discussed in this work empirically on Gaussian models in the in the supplementary materials available with this paper, we refer the reader to Tucker et al. (2019); Le et al. (2019) for a extensive empirical comparisons of all the algorithms discussed in this work. Instead, the main message of our work is that the AISLE-type adaptive importance-sampling paradigm is preferable to the IWAE-type multi-sample objective paradigm because the former allows us to derive all the above-mentioned variants of IWAE – as well as further algorithms which do not require reparametrisations – in a principled manner (the only exception is the standard IWAE reparametrisation ϕ-gradient but this variant suffers from the breakdown highlighted in Rainforth et al. (2018) and was therefore consistently outperformed by the other variants in the simulations shown in Appendix B and in Tucker et al. (2019), for $K > 1$).

1.3 Notation

We assume that all (probability) measures p used in this work are absolutely continuous w.r.t. some suitable dominating measure $\mathrm{d}z$ and with some abuse of notation, we use the same symbol for the measure and the density, i.e. we write $p(\mathrm{d}z) = p(z)\mathrm{d}z$. With this convention, we employ the shorthand $p(f) \coloneqq \int_{\mathsf{Z}} f(z)p(z)\,\mathrm{d}z$ for the integral of some p-integrable test function f; thus, $p(f) = \mathbb{E}_{z\sim p}[f(z)]$ if p is a probability measure. Furthermore, $q^{\otimes K}(z^{1:K}) \coloneqq \prod_{k=1}^K q(z^k)$. We also let 0 denote vectors or matrices of 0s of some appropriate size which will be clear from the context and we let 1 be the function that takes value 1 everywhere on its domain. To keep the notation concise, we hereafter suppress dependence on the observation x, i.e. we write $q_\phi(z) \coloneqq q_{\phi,x}(z)$ as well as

$$\pi_\theta(z) \coloneqq p_\theta(z|x) = \frac{p_\theta(z,x)}{p_\theta(x)} = \frac{\gamma_\theta(z)}{\mathcal{Z}_\theta},$$

where $\gamma_\theta(z) \coloneqq p_\theta(z,x)$ and where $\mathcal{Z}_\theta \coloneqq p_\theta(x) = \int_{\mathsf{Z}} \gamma_\theta(z)\,\mathrm{d}z = \gamma_\theta(1)$.

2 Background

2.1 Importance sampling

Basic idea. We hereafter write $\psi \coloneqq (\theta, \phi)$ and assume that the support of q_ϕ includes the support of π_θ so that the importance weight function $w_\psi(z) \coloneqq \gamma_\theta(z)/q_\phi(z)$ is well defined. For π_θ-integrable $f \colon \mathsf{Z} \to \mathbb{R}$, we can unbiasedly approximate integrals of the form

$$\gamma_\theta(f) \coloneqq \int_{\mathsf{Z}} f(z)\gamma_\theta(z)\,\mathrm{d}z = \int_{\mathsf{Z}} f(z)w_\psi(z)q_\phi(z)\,\mathrm{d}z = q_\phi(fw_\psi), \tag{1}$$

via *importance sampling* using a set of K particles, $\mathbf{z} \coloneqq (z^1, \dots, z^K) \sim q_\phi^{\otimes K}$, which are independent and identically distributed (IID) according to q_ϕ, as

$$\hat{\gamma}_\theta\langle\phi, \mathbf{z}\rangle(f) \coloneqq \frac{1}{K}\sum_{k=1}^K w_\psi(z^k)f(z^k).$$

Here, the notation $\langle\phi, \mathbf{z}\rangle$ stresses the dependence of the estimator on ϕ and $\mathbf{z}$. Note that this is simply an application of the vanilla Monte Carlo method to the expectation from the r.h.s. of (1). Hereafter, we use the convention that $\mathbb{E} = \mathbb{E}_{\mathbf{z}\sim q_\phi^{\otimes K}}$ and $\mathrm{var}_{\mathbf{z}\sim q_\phi^{\otimes K}}$ denote expectation and variance w.r.t. $\mathbf{z} = (z^1, \dots, z^K) \sim q_\phi^{\otimes K}$.

Self-normalised importance sampling. Approximating integrals of the form

$$\pi_\theta(f) := \int_{\mathsf{Z}} f(z)\pi_\theta(z)\,\mathrm{d}z = \frac{\gamma_\theta(f)}{\gamma_\theta(1)},$$

is slightly more complicated because the marginal likelihood $\mathcal{Z}_\theta = \gamma_\theta(1) = p_\theta(x)$ is intractable. Plugging in importance-sampling approximations for both the numerator and denominator leads to the following *self-normalised* importance sampling estimate:

$$\hat{\pi}_\theta\langle\phi, \mathbf{z}\rangle(f) := \frac{\hat{\gamma}_\theta\langle\phi, \mathbf{z}\rangle(f)}{\hat{\gamma}_\theta\langle\phi, \mathbf{z}\rangle(1)} = \sum_{k=1}^{K} \frac{w_\psi(z^k)}{\sum_{l=1}^{K} w_\psi(z^l)} f(z^k).$$

Properties. Proposition 1 summarises some well-known properties of importance-sampling approximations (see, e.g., Geweke, 1989) used throughout this work.

Proposition 1. *Let* $f\colon \mathsf{Z} \to \mathbb{R}$ *be* π_θ*-integrable and* $\mathbf{z} \sim q_\phi^{\otimes K}$*. Then if* $\sup w_\psi < \infty$*,*

1. $\mathbb{E}[\hat{\gamma}_\theta\langle\phi, \mathbf{z}\rangle(f)] = \gamma_\theta(f)$*, for any* $K \in \mathbb{N}$*,*
2. $\mathbb{E}[\hat{\pi}_\theta\langle\phi, \mathbf{z}\rangle(f)] = \pi_\theta(f) + \mathcal{O}(K^{-1})$ *and* $\operatorname{var}[\hat{\pi}_\theta\langle\phi, \mathbf{z}\rangle(f)] = \mathcal{O}(K^{-1})$*,*
3. $\hat{\gamma}_\theta\langle\phi, \mathbf{z}\rangle(f) \to \gamma_\theta(f)$ *and* $\hat{\pi}_\theta\langle\phi, \mathbf{z}\rangle(f) \to \pi_\theta(f)$*, almost surely, as* $K \to \infty$*.*

Proof. Part 1 is immediate; Part 2 is proved, e.g. in Liu (2001, p. 35); Part 3 is a direct consequence of the strong law of large numbers. □

Part 1 of Proposition 1 shows that (non self-normalised) importance-sampling approximations $\hat{\gamma}_\theta\langle\phi, \mathbf{z}\rangle(f)$ are unbiased. In particular,

$$\widehat{\mathcal{Z}}_\theta\langle\phi, \mathbf{z}\rangle := \hat{\gamma}_\theta\langle\phi, \mathbf{z}\rangle(1) = \frac{1}{K}\sum_{k=1}^{K} w_\psi(z^k),$$

is an unbiased estimate of the normalising constant $\mathcal{Z}_\theta = \gamma_\theta(1) = p_\theta(x)$. In contrast, the self-normalised importance-sampling approximation $\hat{\pi}_\theta\langle\phi, \mathbf{z}\rangle(f)$ is typically biased. However, Part 3 shows that it is still consistent and Part 2 ensures that the bias decays quickly in K.

2.2 Importance weighted autoencoder (IWAE)

Objective. The *importance weighted autoencoder (IWAE)*, introduced by Burda et al. (2016), seeks to find a value $\theta^\star$ of the generative-model parameters θ which maximises a lower bound $\mathcal{L}_\psi^K$ on the log-marginal likelihood ('evidence') which depends on the inference-network parameters ϕ and the number of samples, $K \geq 1$,

$$\begin{aligned} \psi^\star := (\theta^\star, \phi^\star) &:= \arg\max\nolimits_\psi \mathcal{L}_\psi^K, \\ \mathcal{L}_\psi^K &:= \mathbb{E}\big[\log \widehat{\mathcal{Z}}_\theta\langle\phi, \mathbf{z}\rangle\big]. \end{aligned} \tag{2}$$

For any finite K, optimisation of the inference-network parameters ϕ tightens the evidence bound. Burda et al. (2016) prove the following properties. Firstly, $\mathcal{L}_\psi^K \leq \log \mathcal{Z}_\theta$ follows from Jensen's inequality and Part 1 of Proposition 1. Secondly, again by Jensen's inequality, $\mathcal{L}_\psi^K \leq \mathcal{L}_\psi^{K+1}$. These inequalities are strict unless $\pi_\theta = q_\phi$. Finally, Part 3 of Proposition 1 (along with the dominated convergence theorem) shows that for any ϕ, $\mathcal{L}_\psi^K \uparrow \log \mathcal{Z}_\theta$ as $K \to \infty$. If $K = 1$, the IWAE reduces to the variational autoencoder (VAE) from Kingma & Welling (2014). However, for $K > 1$, as pointed out in Cremer et al. (2017); Domke & Sheldon (2018), the IWAE also constitutes another VAE on an extended space based on an auxiliary-variable construction developed in Andrieu & Roberts (2009); Andrieu et al. (2010); Lee (2011) (see, e.g. Finke, 2015, for a review).

Standard reparametrisation gradient. The gradient of the IWAE objective from (2) $\nabla_\psi \mathcal{L}_\psi^K = \mathbb{E}\big[\nabla_\psi \log \widehat{\mathcal{Z}}_\theta\langle\phi, \mathbf{z}\rangle + G_\psi(\mathbf{z})\big]$, with $G_\psi(\mathbf{z}) := \log \widehat{\mathcal{Z}}_\theta\langle\phi, \mathbf{z}\rangle \sum_{k=1}^{K} \nabla_\psi \log q_\phi(z^k)$, is typically intractable. However, it could be approximated unbiasedly via a vanilla Monte Carlo approximation using a single sample point $\mathbf{z} = (z^1, \dots, z^K) \sim q_\phi^{\otimes K}$. Unfortunately, the term $G_\psi(\mathbf{z})$ typically has such a large variance that the Monte Carlo approximation becomes impractically noisy (Paisley et al., 2012). To remove this high-variance term, the well known *reparametrisation trick* (Kingma & Welling, 2014) is usually employed. It requires that the following assumption holds.

(R1) *There exists a distribution q on some space E and a diffeomorphism $h_\phi \colon \mathsf{E} \to \mathsf{Z}$ such that $e \sim q \Leftrightarrow h_\phi(e) \sim q_\phi$.*

Under **R1**, the gradient can alternatively be expressed as

$$\begin{aligned}\nabla_\psi \mathcal{L}_\psi^K &= \mathbb{E}_{e^1,\dots,e^K \overset{\text{IID}}{\sim} q}\big[\nabla_\psi \log \widehat{\mathcal{Z}}_\theta\langle \phi, \{h_\phi(e^k)\}_{k=1}^K\rangle\big] \\ &= \mathbb{E}_{e^1,\dots,e^K \overset{\text{IID}}{\sim} q}\Bigg[\sum_{k=1}^K \frac{w_\psi(h_\phi(e^k))}{\sum_{l=1}^K w_\psi(h_\phi(e^l))} \nabla_\psi \log w_\psi(h_\phi(e^k))\Bigg] \\ &= \mathbb{E}\Bigg[\sum_{k=1}^K \frac{w_\psi(z^k)}{\sum_{l=1}^K w_\psi(z^l)} \begin{pmatrix} \nabla_\theta \log \gamma_\theta(z^k) \\ \blacktriangledown_\psi(z^k) - \nabla_\phi \log q_\phi(z^k)\end{pmatrix}\Bigg], \end{aligned} \tag{3}$$

with

$$\blacktriangledown_\psi(z) \coloneqq \nabla_\phi[\log \circ\, w_{\psi'} \circ h_\phi]|_{\psi'=\psi}(h_\phi^{-1}(z)).$$

IWAE then uses a vanilla Monte Carlo estimate of (3) (using a single sample point $\mathbf{z} \sim q_\phi^{\otimes K}$):

$$\begin{bmatrix} \widehat{\nabla}_\theta^{\text{IWAE}}\langle \phi, \mathbf{z}\rangle \\ \widehat{\nabla}_\phi^{\text{IWAE}}\langle \theta, \mathbf{z}\rangle \end{bmatrix} \coloneqq \sum_{k=1}^K \frac{w_\psi(z^k)}{\sum_{l=1}^K w_\psi(z^l)} \begin{bmatrix} \nabla_\theta \log \gamma_\theta(z^k) \\ \blacktriangledown_\psi(z^k) - \nabla_\phi \log q_\phi(z^k)\end{bmatrix}. \tag{4}$$

ϕ-gradient issues. Before proceeding, we state the following lemma, proved in Tucker et al. (2019, Section 8.1), which generalises of the well-known identity $q_\phi(\nabla_\phi \log q_\phi) = 0$.

Lemma 1 (Tucker et al. (2019)). *Under **R1**, for suitably integrable $f_\psi \colon \mathsf{Z} \to \mathbb{R}$:*

$$q_\phi(f_\psi \nabla_\phi \log q_\phi) = q_\phi(\nabla_\phi[f_{\psi'} \circ h_\phi]|_{\psi'=\psi} \circ h_\phi^{-1}). \qquad \square$$

We now exclusively focus on the ϕ-portion of the IWAE gradient, $\widehat{\nabla}_\phi^{\text{IWAE}}\langle \theta, \mathbf{z}\rangle$.

Remark 1 (drawbacks of the IWAE ϕ-gradient). *The gradient $\widehat{\nabla}_\phi^{\text{IWAE}}\langle \theta, \mathbf{z}\rangle$ has three drawbacks. The last two of these are attributable to the 'score-function' terms $\nabla_\phi \log q_\phi(z)$ in the ϕ-gradient portion of* (4)*.*

- ***Reliance on reparametrisations.*** *A continuous reparametrisation à la **R1** is necessary to remove the high-variance term $G_\psi(\mathbf{z})$; this makes it difficult to use IWAE for models with e.g. discrete latent variables z (Le et al., 2019).*

- ***Vanishing signal-to-noise ratio.*** *The ϕ-gradient breaks down in the sense that its signal-to-noise ratio vanishes as $\mathbb{E}[\widehat{\nabla}_\phi^{\text{IWAE}}\langle \theta, \mathbf{z}\rangle] / \operatorname{var}[\widehat{\nabla}_\phi^{\text{IWAE}}\langle \theta, \mathbf{z}\rangle]^{1/2} = \mathcal{O}(K^{-1/2})$ (Rainforth et al., 2018). This follows from Part 2 of Proposition 1 since $\widehat{\nabla}_\phi^{\text{IWAE}}\langle \theta, \mathbf{z}\rangle$ constitutes a self-normalised importance-sampling approximation of $\pi_\theta(\blacktriangledown_\psi - \nabla_\phi \log q_\phi) = 0$ (the last identity follows from Lemma 1 with $f_\psi = w_\psi$).*

- ***Inability to achieve zero variance.*** *As pointed out in Roeder et al. (2017), $\operatorname{var}[\widehat{\nabla}_\phi^{\text{IWAE}}\langle \theta, \mathbf{z}\rangle]] > 0$ even in the ideal scenario that $q_\phi = \pi_\theta$ despite the fact that in this case, w_ψ is constant and hence $\operatorname{var}[\log \widehat{\mathcal{Z}}_\theta\langle \phi, \mathbf{z}\rangle] = 0$.*

Two modifications of $\widehat{\nabla}_\phi^{\text{IWAE}}\langle \theta, \mathbf{z}\rangle$ have been proposed which (under **R1**) avoid the score-function terms in (4) and hence (a) exhibit a stable signal-to-noise ratio as $K \to \infty$ and (b) can achieve zero variance if $q_\phi = \pi_\theta$ (because then $\blacktriangledown_\psi \equiv 0$ since w_ψ is constant).

- **IWAE-STL.** The *'sticking-the-landing' IWAE (IWAE-STL)* gradient proposed by Roeder et al. (2017) heuristically ignores the score function terms (this introduces bias relative to $\widehat{\nabla}_\phi^{\text{IWAE}}\langle \phi, \mathbf{z}\rangle$ whenever $K > 1$ as shown in Tucker et al. (2019)):

$$\widehat{\nabla}_\phi^{\text{IWAE-STL}}\langle \theta, \mathbf{z}\rangle \coloneqq \sum_{k=1}^K \frac{w_\psi(z^k)}{\sum_{l=1}^K w_\psi(z^l)} \blacktriangledown_\psi(z^k). \tag{5}$$

- **IWAE-DREG.** The *'doubly-reparametrised' IWAE (IWAE-DREG)* gradient proposed by Tucker et al. (2019) removes the score-function terms through Lemma 1 (i.e. this does not introduce bias relative to $\widehat{\nabla}_\phi^{\text{IWAE}}\langle \phi, \mathbf{z}\rangle$):

$$\widehat{\nabla}_\phi^{\text{IWAE-DREG}}\langle \theta, \mathbf{z}\rangle \coloneqq \sum_{k=1}^K \left(\frac{w_\psi(z^k)}{\sum_{l=1}^K w_\psi(z^l)}\right)^2 \blacktriangledown_\psi(z^k). \tag{6}$$

2.3 Reweighted wake-sleep (RWS)

The *reweighted wake-sleep (RWS)* algorithm was proposed in Bornschein & Bengio (2015).[1] Letting $\mathrm{KL}(p\|q) := \int_{\mathsf{Z}} \log(p(z)/q(z))q(z)\,\mathrm{d}z$ is the Kullback–Leibler (KL)-divergence from p to q, the RWS algorithm seeks to optimise $\psi = (\theta, \phi)$ as

$$\begin{aligned}\theta^\star &:= \theta^{\mathrm{ML}} = \arg\max\nolimits_\theta \log \mathcal{Z}_\theta,\\ \phi^\star &:= \arg\min\nolimits_\phi \mathrm{KL}(\pi_{\theta^\star}\|q_\phi).\end{aligned}$$

The θ- and ϕ-gradients

$$\begin{bmatrix}\nabla_\theta \log \mathcal{Z}_\theta\\ -\nabla_\phi \mathrm{KL}(\pi_\theta\|q_\phi)\end{bmatrix} = \pi_\theta\begin{pmatrix}\nabla_\theta \log\gamma_\theta\\ \nabla_\phi \log q_\phi\end{pmatrix}, \tag{7}$$

are usually intractable and therefore approximated by replacing π_θ by the self-normalised importance sampling approximation $\hat{\pi}_\theta\langle\phi,\mathbf{z}\rangle$ (note that this does not need **R1**):

$$\begin{bmatrix}\widehat{\nabla}_\theta^{\mathrm{RWS}}\langle\phi,\mathbf{z}\rangle\\ \widehat{\nabla}_\phi^{\mathrm{RWS}}\langle\theta,\mathbf{z}\rangle\end{bmatrix} := \sum_{k=1}^K \frac{w_\psi(z^k)}{\sum_{l=1}^K w_\psi(z^l)}\begin{bmatrix}\nabla_\theta\log\gamma_\theta(z^k)\\ \nabla_\phi \log q_\phi(z^k)\end{bmatrix}. \tag{8}$$

Since (8) relies on self-normalised importance sampling, it biased relative to (7). However, by Part 2 of Proposition 1 the bias of the θ-gradient $\widehat{\nabla}_\theta^{\mathrm{RWS}}\langle\phi,\mathbf{z}\rangle = \widehat{\nabla}_\theta^{\mathrm{IWAE}}\langle\phi,\mathbf{z}\rangle$ relative to $\nabla_\theta \log\mathcal{Z}_\theta$ decays as $\mathcal{O}(K^{-1})$. Appendix A discusses the impact of the bias on the ϕ-gradients.

The optimisation of θ and ϕ is carried out simultaneously. This is because (a) a better proposal q_ϕ reduces both bias and variance of (self-normalised) importance-sampling approximations and can therefore be leveraged for reducing the bias and variance of the θ-gradients and (b) this strategy reduces the computational cost because the same set of particles $\mathbf{z}$ and weights $\{w_\psi(z^k)\}_{k=1}^K$ is shared by both gradients. However, this simultaneous optimisation is often viewed as the main drawback of RWS because there is no joint objective (for both θ and ϕ).

RWS-DREG. Under **R1**, Tucker et al. (2019) proposed the following *'doubly-reparametrised' RWS (RWS-DREG)* gradient which is equal to $\widehat{\nabla}_\phi^{\mathrm{RWS}}\langle\theta,\mathbf{z}\rangle$ in expectation and is derived by applying Lemma 1 to the latter:

$$\widehat{\nabla}_\phi^{\mathrm{RWS\text{-}DREG}}\langle\theta,\mathbf{z}\rangle := \sum_{k=1}^K\left[\frac{w_\psi(z^k)}{\sum_{l=1}^K w_\psi(z^l)} - \left(\frac{w_\psi(z^k)}{\sum_{l=1}^K w_\psi(z^l)}\right)^2\right]\blacktriangledown_\psi(z^k). \tag{9}$$

3 AISLE: A unified adaptive importance-sampling framework

3.1 Objective

If θ is fixed, the RWS algorithm reduces to an adaptive importance-sampling scheme which optimises the proposal distribution by minimising the KL-divergence from the target distribution π_θ to the proposal q_ϕ (see, e.g., Douc et al., 2007; Cappé et al., 2008). If instead ϕ is fixed, the RWS algorithm reduces to a stochastic-approximation algorithm for estimating the MLE of the generative-model parameters θ. The advantage of optimising θ and ϕ simultaneously is that (a) Monte Carlo samples used to approximate the θ-gradient can be re-used to approximate the ϕ-gradient and (b) optimising ϕ typically reduces the error (both in terms of bias and variance) of the θ-gradient approximation.

However, adapting the proposal distribution q_ϕ in importance-sampling schemes need not necessarily be based on minimising the KL-divergence. Numerous other techniques exist in the literature (e.g. Geweke, 1989; Evans, 1991; Oh & Berger, 1992; Richard & Zhang, 2007; Cornebise et al., 2008) and may sometimes be preferable. Indeed, another popular approach with strong theoretical support is based on minimising the χ^2-divergence (see, e.g., Deniz Akyildiz & Míguez, 2019). Based on this insight, we slightly generalise the RWS-objective as

$$\begin{aligned}\theta^\star &:= \arg\max\nolimits_\theta \log\mathcal{Z}_\theta (= \theta^{\mathrm{ML}}),\\ \phi^\star &:= \arg\min\nolimits_\phi \mathrm{D_f}(\pi_{\theta^\star}\|q_\phi).\end{aligned} \tag{10}$$

[1] Following Tucker et al. (2019) (based on empirical results in Le et al. (2019)), we only use the 'wake-phase' ϕ-updates for RWS.

Here, $\mathrm{D_f}(p\|q) \coloneqq \int_{\mathsf{Z}} \mathrm{f}(p(z)/q(z))q(z)\,\mathrm{d}z$ is some f-divergence from p to q. We reiterate that alternative approaches for optimising ϕ (which do not minimise f-divergences) could be used. However, we state (10) for concreteness as it suffices for the remainder of this work; we call the resulting algorithm *adaptive importance sampling for learning (AISLE)*. We stress again that AISLE is *not* introduced with the aim or claim of proposing a new algorithms but to formalise the argument that the adaptive importance-sampling paradigm avoids the drawbacks from Remark 1 thus making it preferable to the multi-sample objective paradigm.

3.2 θ-gradient

Optimisation is again performed via a stochastic gradient-ascent. The intractable θ-gradient $\nabla_\theta \log \mathcal{Z}_\theta = \pi_\theta(\nabla_\theta \log \gamma_\theta)$ is approximated as in RWS, i.e. for $\mathbf{z} \sim q_\phi^{\otimes K}$:

$$\widehat{\nabla}_\theta^{\text{AISLE}}\langle \phi, \mathbf{z}\rangle \coloneqq \widehat{\nabla}_\theta^{\text{RWS}}\langle \phi, \mathbf{z}\rangle = \widehat{\nabla}_\theta^{\text{IWAE}}\langle \phi, \mathbf{z}\rangle.$$

The θ-gradient is thus the same for all algorithms discussed in this work although the IWAE-paradigm views it as an unbiased gradient for a biased objective while AISLE (and RWS) interpret it as a self-normalised importance-sampling (and hence biased) approximation of the gradient $\nabla_\theta \log \mathcal{Z}_\theta$ for the 'exact' objective.

3.3 ϕ-gradient special case I: RWS and IWAE-STL

The ϕ-gradients depend on the particular choice of f-divergence in (10). By construction, we recover RWS as a special case of AISLE if we define the f-divergence through $\mathrm{f}(y) \coloneqq y \log y$ because in this case $\mathrm{D_f}(p\|q) = \mathrm{KL}(p\|q)$ reduces to the KL-divergence. Our main contribution in this subsection is to show that a more principled application of the identity from Lemma 1 leads to the IWAE-STL gradient from (5).

To derive the AISLE ϕ-gradients for this divergence we note that

$$-\nabla_\phi \mathrm{KL}(\pi_\theta\|q_\phi) = \pi_\theta(\nabla_\phi \log q_\phi), \tag{11}$$

which, under **R1**, by Lemma 1 with $f_\psi = w_\psi$, can be written as

$$\pi_\theta(\nabla_\phi \log q_\phi) = q_\phi(w_\psi \nabla_\phi \log q_\phi)/\mathcal{Z}_\theta = q_\phi(w_\psi \blacktriangledown_\psi)/\mathcal{Z}_\theta = \pi_\theta(\blacktriangledown_\psi). \tag{12}$$

We then obtain practical approximations of these gradients by plugging in $\hat{\pi}_\theta\langle\phi, \mathbf{z}\rangle$ for π_θ.

- **AISLE-KL-NOREP/RWS.** Without relying on any reparametrisation, (11) yields the following gradient, which clearly equals $\widehat{\nabla}_\phi^{\text{RWS}}\langle\theta, \mathbf{z}\rangle$:
$$\widehat{\nabla}_\phi^{\text{AISLE-KL-NOREP}}\langle\theta, \mathbf{z}\rangle \coloneqq \sum_{k=1}^K \frac{w_\psi(z^k)}{\sum_{l=1}^K w_\psi(z^l)} \nabla_\phi \log q_\phi(z^k). \tag{13}$$
- **AISLE-KL.** Using the reparametrisation from **R1**, (12) yields the gradient:
$$\widehat{\nabla}_\phi^{\text{AISLE-KL}}\langle\theta, \mathbf{z}\rangle \coloneqq \sum_{k=1}^K \frac{w_\psi(z^k)}{\sum_{l=1}^K w_\psi(z^l)} \blacktriangledown_\psi(z^k). \tag{14}$$

We thus arrive at the following result which demonstrates that IWAE-STL can be derived in a principled manner from AISLE, i.e. without the need for a multi-sample objective.

Proposition 2. *For any* $(\theta, \phi, \mathbf{z})$, $\widehat{\nabla}_\phi^{\text{AISLE-KL}}\langle\theta, \mathbf{z}\rangle = \widehat{\nabla}_\phi^{\text{IWAE-STL}}\langle\theta, \mathbf{z}\rangle$. □

Proposition 2 thus provides a theoretical basis for IWAE-STL which was previously viewed as an alternative gradient for IWAE for which it is biased and only heuristically justified. Furthermore, the fact that IWAE-STL exhibited good empirical performance in Tucker et al. (2019) even in an example in which RWS broke down, suggests that this breakdown may not be due to RWS' lack of optimising a joint objective as previously conjectured.

Finally, recall that Tucker et al. (2019) obtained an alternative 'doubly-reparametrised' RWS ϕ-gradient $\widehat{\nabla}_\phi^{\text{RWS-DREG}}\langle\theta, \mathbf{z}\rangle$ given in (9) by *first* replacing the exact (but intractable) ϕ-gradient from (11) by the self-normalised importance-sampling approximation $\widehat{\nabla}_\phi^{\text{RWS}}\langle\theta, \mathbf{z}\rangle$ and *then* applying the identity from Lemma 1. Note that this may result in a variance reduction but does not change the bias of the gradient estimator. In contrast, AISLE-KL is derived by *first* applying Lemma 1 to the exact (RWS) ϕ-gradient and *then* approximating the resulting expression. This can potentially reduce both bias and variance.

3.4 ϕ-GRADIENT SPECIAL CASE II: IWAE-DREG

We now demonstrate that the IWAE-DREG gradient can be recovered as a special case of AISLE (up to a proportionality constant). To establish this relationship, we take $\mathrm{f}(y) \coloneqq (y-1)^2$ so that $\mathrm{D_f}(p\|q) = \chi^2(p\|q) \coloneqq \int_{\mathsf{Z}}([p(z)/q(z)] - 1)^2 q(z)\,\mathrm{d}z = \int_{\mathsf{Z}}[p(z)/q(z)]p(z)\,\mathrm{d}z - 1$, is the χ^2-divergence. Minimising this divergence is natural in importance sampling since $\chi^2(\pi_\theta\|q_\phi) = \mathrm{var}_{z\sim q_\phi}[w_\psi/\mathcal{Z}_\theta]$ is the variance of the importance weights.

To derive the AISLE ϕ-gradients for this divergence we note that

$$-\nabla_\phi\,\chi^2(\pi_\theta\|q_\phi) = -\pi_\theta(\nabla_\phi w_\psi)/\mathcal{Z}_\theta = \pi_\theta(w_\psi \nabla_\phi \log q_\phi)/\mathcal{Z}_\theta, \tag{15}$$

which, under **R1**, by Lemma 1 with $f_\psi = w_\psi^2$, can be written as

$$\begin{aligned}\pi_\theta(w_\psi \nabla_\phi \log q_\phi)/\mathcal{Z}_\theta &= q_\phi(w_\psi^2 \nabla_\phi \log q_\phi)/\mathcal{Z}_\theta^2\\ &= q_\phi(w_\psi^2 \nabla_\phi[\log\circ\, w_{\psi'}^2 \circ h_\phi]|_{\psi'=\psi} \circ h_\phi^{-1})/\mathcal{Z}_\theta^2 = \pi_\theta(2 w_\psi \blacktriangledown_\psi)/\mathcal{Z}_\theta.\end{aligned} \tag{16}$$

Again plugging in $\hat{\pi}_\theta\langle\phi,\mathbf{z}\rangle$ for π_θ and $\widehat{\mathcal{Z}}_\theta\langle\phi,\mathbf{z}\rangle$ for $\mathcal{Z}_\theta$ yields the following approximations.

- **AISLE-χ^2-NOREP.** Without relying on any reparametrisation, (15) yields the following gradient which is also proportional to the 'score gradient' from Dieng et al. (2017, Appendix G):
$$\widehat{\nabla}_\phi^{\text{AISLE-}\chi^2\text{-NOREP}}\langle\theta,\mathbf{z}\rangle \coloneqq K\sum_{k=1}^K \left(\frac{w_\psi(z^k)}{\sum_{l=1}^K w_\psi(z^l)}\right)^2 \nabla_\phi \log q_\phi(z^k). \tag{17}$$
- **AISLE-χ^2.** Using the reparametrisation from **R1**, (16) yields the gradient:
$$\widehat{\nabla}_\phi^{\text{AISLE-}\chi^2}\langle\theta,\mathbf{z}\rangle \coloneqq 2K\sum_{k=1}^K \left(\frac{w_\psi(z^k)}{\sum_{l=1}^K w_\psi(z^l)}\right)^2 \blacktriangledown_\psi(z^k). \tag{18}$$

We thus arrive at the following result which demonstrates that IWAE-DREG can be derived (up to the proportionality factor $2K$) in a principled manner from AISLE, i.e. without the need for a multi-sample objective.

Proposition 3. *For any* $(\theta,\phi,\mathbf{z})$, $\widehat{\nabla}_\phi^{\text{AISLE-}\chi^2}\langle\theta,\mathbf{z}\rangle = 2K\widehat{\nabla}_\phi^{\text{IWAE-DREG}}\langle\theta,\mathbf{z}\rangle$. □

Note that if the implementation normalises the gradients, e.g. as effectively done by ADAM (Kingma & Ba, 2015), the constant factor cancels out and AISLE-χ^2 becomes equivalent to IWAE-DREG. Otherwise (e.g. in plain stochastic gradient-ascent) Proposition 3 shows that the learning rate needs to be scaled as $\mathcal{O}(K)$ for the IWAE or IWAE-DREG ϕ-gradients.

4 Conclusion

We have shown that the adaptive-importance sampling paradigm of the *reweighted wake-sleep (RWS)* (Bornschein & Bengio, 2015) is preferable to the multi-sample objective paradigm of *importance weighted autoencoders (IWAEs)* (Burda et al., 2016) because the former achieves all the goals of the latter whilst avoiding its drawbacks. To formalise this argument, we have introduced a simple, unified adaptive-importance-sampling framework termed *adaptive importance sampling for learning (AISLE)* (which slightly generalises the RWS algorithm) and have proved that AISLE allows us to derive the *'sticking-the-landing' IWAE (IWAE-STL)* gradient from Roeder et al. (2017) and the *'doubly-reparametrised' IWAE (IWAE-DREG)* gradient from Tucker et al. (2019) as special cases.

We hope that this work highlights the potential for further improving variational techniques by drawing upon the vast body of research on (adaptive) importance sampling in the computational statistics literature. Conversely, the methodological connections established in this work may also serve to emphasise the utility of the reparametrisation trick from Kingma & Welling (2014); Tucker et al. (2019) to computational statisticians.

In a companion article, we are extending the present work to the *variational sequential Monte Carlo* methods from Maddison et al. (2017); Le et al. (2018); Naesseth et al. (2018) and to the *tensor Monte Carlo* approach from Aitchison (2018).

References

Laurence Aitchison. Tensor Monte Carlo: particle methods for the GPU era. *arXiv e-prints*, art. arXiv:1806.08593, Jun 2018.

Christophe Andrieu and Gareth O Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010. With discussion.

Robert Bamler, Cheng Zhang, Manfred Opper, and Stephan Mandt. Perturbative black box variational inference. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5079–5088, 2017.

Jörg Bornschein and Yoshua Bengio. Reweighted wake-sleep. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *4th International Conference on Learning Representations (ICLR)*, 2016.

Olivier Cappé, Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18 (4):447–459, 2008.

Julien Cornebise, Éric Moulines, and Jimmy Olsson. Adaptive methods for sequential importance sampling with application to state space models. *Statistics and Computing*, 18 (4):461–480, 2008.

Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting importance-weighted autoencoders. In *5th International Conference on Learning Representations (ICLR)*, 2017.

Ömer Deniz Akyildiz and Joaquín Míguez. Convergence rates for optimised adaptive importance samplers. *arXiv e-prints*, art. arXiv:1903.12044, 2019.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ upper bound minimization. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2732–2741, 2017.

Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4475–4484, 2018.

Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P Robert. Convergence of adaptive mixtures of importance sampling schemes. *The Annals of Statistics*, 35(1): 420–448, 2007.

Michael Evans. Adaptive importance sampling and chaining. *Statistical Numerical Integration, Contemporary Mathematics*, 115:137–143, 1991.

Axel Finke. *On extended state-space constructions for Monte Carlo methods*. PhD thesis, Department of Statistics, University of Warwick, UK, 2015.

John Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339, 1989.

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

Diederik P Kingma and Jimmy Lei Ba. ADAM: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014.

Augustine Kong, Jun S. Liu, and Wing Hung Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, 1994.

Tuan Anh Le, Maximilian Igl, Tom Rainforth, Tom Jin, and Frank Wood. Auto-encoding sequential Monte Carlo. In *6th International Conference on Learning Representations (ICLR)*, 2018.

Tuan Anh Le, Adam R Kosiorek, N Siddharth, Yee Whye Teh, and Frank Wood. Revisiting reweighted wake-sleep for models with stochastic control flow. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

Anthony Lee. *On auxiliary variables and many-core architectures in computational statistics*. PhD thesis, Department of Statistics, University of Oxford, UK, 2011.

Jun S Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6(2):113–119, 1996.

Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer, 2001.

Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Whye Teh. Filtering variational objectives. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6573–6583, 2017.

Christian A Naesseth, Scott W Linderman, Rajesh Ranganath, and David M Blei. Variational sequential Monte Carlo. In *21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

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

John Paisley, David Blei, and Michael Jordan. Variational Bayesian inference with stochastic search. In *29th International Conference on Machine Learning (ICML)*, 2012.

Tom Rainforth, Adam R Kosiorek, Tuan Anh Le, Chris J Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In *Bayesian Deep Learning (NeurIPS 2018 workshop)*, 2018.

Jean-Franąois Richard and Wei Zhang. Efficient high-dimensional importance sampling. *Journal of Econometrics*, 141(2):1385–1411, 2007.

Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6925–6934, 2017.

George Tucker, Dieterich Lawson, Shixiang Gu, and Chris J Maddison. Doubly reparameterized gradient estimators for Monte Carlo objectives. In *7th International Conference on Learning Representations (ICLR)*, 2019.

Ming Xu, Matias Quiroz, Robert Kohn, and Scott A Sisson. Variance reduction properties of the reparameterization trick. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2711–2720, 2019.

A On the rôle of the self-normalisation bias within RWS/AISLE

A.1 The self-normalisation bias

Within the self-normalised importance-sampling approximation, the number of particles, K, interpolates between two extremes:

- As $K \uparrow \infty$, $\hat{\pi}_\theta\langle\phi, \mathbf{z}\rangle(f)$ becomes an increasingly accurate approximation of $\pi_\theta(f)$.
- For $K = 1$, however, $\hat{\pi}_\theta\langle\phi, \mathbf{z}\rangle(f) = f(z^1)$ reduces to a vanilla Monte Carlo approximation of $q_\phi(f)$ (because the single self-normalised importance weight is always equal to 1).

This leads to the following insight about the estimators $\widehat{\nabla}_\phi^{\text{AISLE-KL}}\langle\theta, \mathbf{z}\rangle$ and $\widehat{\nabla}_\phi^{\text{AISLE-}\chi^2}\langle\theta, \mathbf{z}\rangle$.

- As $K \uparrow \infty$, these two estimators become increasingly accurate approximations of the *'inclusive'*-divergence gradients $-\nabla_\phi \operatorname{KL}(\pi_\theta \| q_\phi) = \pi_\theta(\blacktriangledown_\phi)$ and $-\nabla_\phi \chi^2(\pi_\theta \| q_\phi) = 2\pi_\theta([w_\psi/\mathcal{Z}_\theta]\blacktriangledown_\phi)$, respectively.
- For $K = 1$, however, these two estimators reduce to vanilla Monte Carlo approximations of the *'exclusive'*-divergence gradients $-\nabla_\phi \operatorname{KL}(q_\phi \| \pi_\theta) = q_\phi(\blacktriangledown_\phi)$ and $-2\nabla_\phi \operatorname{KL}(q_\phi \| \pi_\theta) = 2q_\phi(\blacktriangledown_\phi)$, respectively.

This is similar to the standard IWAE ϕ-gradient which also represents a vanilla Monte Carlo approximation of $-\nabla_\phi \operatorname{KL}(q_\phi \| \pi_\theta)$ if $K = 1$ as IWAE reduces to a VAE in this case.

Characterising the small-K self-normalisation bias of the reparametrisation-free AISLE ϕ gradients, AISLE-KL-NOREP and AISLE-χ^2-NOREP, is more difficult because if $K = 1$, they constitute vanilla Monte Carlo approximations of $q_\phi(\nabla_\phi \log q_\phi) = 0$. Nonetheless, Le et al. (2019, Figure 5) lends some support to the hypothesis that the small-K self-normalisation bias of these gradients also favours a minimisation of the exclusive KL-divergence.

A.2 Inclusive vs exclusive KL-divergence minimisation

Recall that the main motivation for use of IWAEs (instead of VAEs) was the idea that we could use self-normalised importance-sampling approximations with $K > 1$ particles to reduce the bias of the θ-gradient relative to $\nabla_\theta \log \mathcal{Z}_\theta$. The error of such (self-normalised) importance-sampling approximations can be controlled by ensuring that q_ϕ is close to π_θ (in some suitable sense) in any part of the space Z in which π_θ has positive probability mass. For instance, it is well known that the error will be small if the 'inclusive' KL-divergence $\operatorname{KL}(\pi_\theta \| q_\phi)$ is small as this implies well-behaved importance weights. In contrast, a small 'exclusive' KL-divergence $\operatorname{KL}(q_\phi \| \pi_\theta)$ is not sufficient for well-behaved importance weights because the latter only ensures that q_ϕ is close to π_θ in those parts of the space Z in which q_ϕ has positive probability mass.

Let $\mathcal{Q} := \{q_\phi\}$ (which is indexed by ϕ) be the family of proposal distributions/the variational family. Then we can distinguish two scenarios.

1. **Sufficiently expressive $\mathcal{Q}$.** For the moment, assume that the family $\mathcal{Q}$ is flexible ('expressive') enough in the sense that it contains a distribution $q_{\phi^\star}$ which is (at least approximately) equal to π_θ and that our optimiser can reach the value $\phi^\star$ of ϕ. In this case, minimising the exclusive KL-divergence can still yield well-behaved importance weights because in this case, $\phi^\star := \arg\min_\phi \operatorname{KL}(\pi_\theta \| q_\phi)$ is (at least approximately) equal to $\arg\min_\phi \operatorname{KL}(q_\phi \| \pi_\theta)$.
2. **Insufficiently expressive $\mathcal{Q}$.** In general, the family $\mathcal{Q}$ is not flexible enough in the sense that all of its members are 'far away' from π_θ, e.g. if the D components $z_1, \dots, z_D$ of $z = z_{1:D}$ are highly correlated under π_θ whilst $q_\phi(z) = \prod_{d=1}^D q_{\phi,d}(z_d)$ is fully factorised. In this case, minimising the exclusive KL-divergence could lead to poorly-behaved importance weights and we should optimise $\phi^\star := \arg\min_\phi \operatorname{KL}(\pi_\theta \| q_\phi)$ as discussed above.

Remark 2. *In Scenario 1 above, i.e. for a sufficiently flexible $\mathcal{Q}$, using a gradient-descent algorithm which seeks to minimise the* exclusive *divergence can sometimes be preferable to a gradient-descent algorithm which seeks to minimise the* inclusive *divergence. This is because both find (approximately) the same optimum but the latter may exhibit faster convergence in some applications. In such scenarios, the discussion in Subsection A.1 indicates that a smaller number of particles, K, could then be preferable for some of the ϕ-gradients because (a) the $\mathcal{O}(K^{-1})$ self-normalisation bias outweighs the $\mathcal{O}(K^{-1/2})$ standard deviation and (b) the direction of this bias may favour faster convergence.*

Unfortunately, simply setting $K = 1$ for the approximation of the ϕ-gradients[2] is not necessarily optimal because

- even in the somewhat idealised scenario 1 above and even if the direction of the self-normalisation bias encourages faster convergence, increasing K is still desirable to reduce the variance of the gradient approximations;
- not using the information contained in *all* K particles and weights (which have already been sampled/calculated to approximate the θ-gradient) seems wasteful;
- if $K = 1$, the reparametrisation-free AISLEϕ gradients, AISLE-KL-NOREP and AISLE-χ^2-NOREP are simply vanilla Monte Carlo estimates of 0 and the RWS-DREG ϕ-gradient is then equal to 0.

A.3 Regularisation

We propose here to 'regularise' the importance weights. That is, letting

$$ESS(w_{1:K}) \coloneqq \left[\sum_{k=1}^{K}\left(\frac{w_k}{\sum_{l=1}^{K} w_l}\right)^2\right]^{-1} \in [1, K],$$

be the *effective sample size* (Kong et al., 1994; Liu, 1996), we propose to replace the weights $w_{\psi,x}(z^k) =: w_k$ in any of the gradients discussed in this work by $w_k^{\alpha^\star}$, where $\alpha^\star \coloneqq \sup\{\alpha \in [0,1] : ESS(w_{1:K}^{\alpha}) \geq \eta K\}$ can be inexpensively found with bisection; the tuning parameter $\eta \in (0, 1]$ governs the amount of regularisation.

This weight-regularisation strategy typically reduces the variance but increases the bias of (self-normalised) importance-sampling approximations $\hat{\pi}_\theta\langle\phi, \mathbf{z}\rangle(f)$ relative to $\pi_\theta(f)$. When applied to the ϕ-gradients, it may be interpreted in two different ways.

1. Within Scenario 1, we can view such a regularisation strategy as a means of interpolating between K-sample vanilla Monte Carlo approximations of gradients of the exclusive KL-divergence ($\alpha^\star = 0$) and self-normalised importance-sampling approximations of inclusive-divergence type gradients ($\alpha^\star = 1$).
2. Within Scenario 2, we can view such a regularisation strategy as a means reducing the overall approximation error of importance sampling through a more favourable bias–variance trade-off (Ionides, 2008). Thus, in this scenario, we interpret the regularisation as a way of attaining ϕ-gradients whose overall error (relative to the intractable inclusive-divergence gradient) is reduced.

We note that the second interpretation applies to the θ-gradient in either scenario. Furthermore, such weight regularisation also circumvents the signal-to-noise ratio breakdown in the standard IWAE ϕ-gradient. For an alternative weight-regularisation strategy used in the context of variational inference, see Bamler et al. (2017).

Finally, a more principled approach would be to regularise the problem rather than the approximation. That is, we could instead regularise the generative model, i.e. replace $\pi_{\theta,x}(z)$ by a regularised distribution, e.g. by a distribution proportional to $\pi_{\vartheta,x}(z)^\alpha q_{\varphi,x}(z)^{1-\alpha}$ for $\alpha \in (0, 1]$ (and also replacing θ by (ϑ, φ)). We are currently investigating such ideas.

B Empirical illustration

B.1 Algorithms

In these supplementary materials, we illustrate the different ϕ-gradient estimators (recall that all algorithms discussed in this work share the same θ-gradient estimator). Specifically, we compare the following approximations.

[2] Within the IWAE-paradigm, using different numbers of particles for the θ and ϕ-gradients was recently proposed in Rainforth et al. (2018); Le et al. (2018) who termed this approach *'alternating evidence lower bounds'*, albeit their aim was to circumvent the signal-to-noise ratio breakdown of the IWAE ϕ-gradient which is distinct from the phenomenon discussed here.

- **AISLE-KL-NOREP.** The gradient for AISLE based on the KL-divergence without any further reparametrisation from (13) i.e. this coincides with the standard RWS-gradient from (8). This gradient does not require **R1** but does not achieve zero variance even if $q_\phi = \pi_\theta$.
- **AISLE-KL.** The gradient for AISLE based on the KL-divergence after reparametrising and exploiting the identity from Lemma 1; it is given by (14) and coincides with the IWAE-STL-gradient from (5).
- **AISLE-χ^2-NOREP.** The gradient for AISLE based on the χ^2-divergence without any reparametrisation given in (17). This gradient again does not require **R1** but does not achieve zero variance even if $q_\phi = \pi_\theta$.
- **AISLE-χ^2.** The gradient for AISLE based on the χ^2-divergence after reparametrising and exploiting the identity from Lemma 1; it is given by (18) and is alsow proportional to IWAE-DREG from Tucker et al. (2019) which was stated in (6). When normalising the gradients (as, e.g. implicitly done by optimisers such as ADAM Kingma & Ba, 2015) the proportionality constant cancels out so that both these gradient approximations lead to computationally the same algorithm.
- **IWAE.** The gradient for IWAE employing the reparametrisation trick from Kingma & Welling (2014). Its sampling approximation is given in (4). Recall that this is the ϕ-gradient whose signal-to-noise ratio degenerates with K as pointed out in Rainforth et al. (2018) (and which also cannot achieve zero variance even if $q_\phi = \pi_\theta$).
- **IWAE-DREG.** The 'doubly-reparametrised' IWAE gradient from (6) which was proposed in Tucker et al. (2019). It is proportional to AISLE-χ^2.
- **RWS-DREG.** The 'doubly-reparametrised' RWS ϕ-gradient from (9) which was proposed in Tucker et al. (2019) who derived it by applying the identity from Lemma 1 to the RWS ϕ-gradient.

B.2 Model

Generative model. We have N D-dimensional observations $x^{(1)}, \dots, x^{(N)} \in \mathbb{R}^D$ and N D-dimensional latent variables $z^{(1)}, \dots, z^{(N)} \in \mathbb{R}^D$. Unless otherwise stated, any vector $y \in \mathbb{R}^D$ is to be viewed as a $D \times 1$ column vector.

Hereafter, wherever necessary, we add an additional subscript to make the dependence on the observations explicit. The joint law (the 'generative model'), parametrised by θ, of the observations and latent variables then factorises as

$$\prod_{n=1}^{N} p_\theta(z^{(n)}) p_\theta(x^{(n)}|z^{(n)}) = \prod_{n=1}^{N} \gamma_{\theta,x^{(n)}}(z^{(n)}).$$

We model each latent variable–observation pair (z, x) as

$$\begin{aligned} p_\theta(z) &\coloneqq \mathrm{N}(z; \mu, \Sigma), \\ p_\theta(x|z) &\coloneqq \mathrm{N}(x; z; \mathrm{I}), \end{aligned}$$

where $\theta \coloneqq \mu = \mu_{1:D} \in \mathbb{R}^D$, where $\Sigma \coloneqq (\sigma_{d,d'})_{(d,d') \in \{1,\dots,D\}} \in \mathbb{R}^{D \times D}$ is assumed to be known and where I denotes the $D \times D$-identity matrix. For any θ,

$$\mathcal{Z}_{\theta,x} = p_\theta(x) = \mathrm{N}(x; \mu, \mathrm{I} + \Sigma), \tag{19}$$

$$\pi_{\theta,x}(z) = p_\theta(z|x) = \mathrm{N}(z; \nu_{\theta,x}, P), \tag{20}$$

with $P \coloneqq (\Sigma^{-1} + \mathrm{I})^{-1}$ and $\nu_{\theta,x} \coloneqq P(\Sigma^{-1}\mu + x)$. In particular, (19) implies that $\theta^{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x^{(n)}$.

Proposal/variational approximation. We take the proposal distributions as a fully-factored Gaussian:

$$q_{\phi,x}(z) \coloneqq \mathrm{N}(z; Ax + b, C), \tag{21}$$

where $A = (a_{d,d'})_{(d,d') \in \{1,\dots,D\}^2} \in \mathbb{R}^{D \times D}$, $b = b_{1:D} \in \mathbb{R}^D$ and, for $c_{1:D} =: c \in \mathbb{R}^D$, $C \coloneqq \operatorname{diag}(\mathrm{e}^{2c_1}, \dots, \mathrm{e}^{2c_D})$. The parameters to optimise are thus

$$\phi \coloneqq (a_1^{\mathrm{T}}, \dots, a_D^{\mathrm{T}}, b^{\mathrm{T}}, c^{\mathrm{T}}),$$

where $a_d \coloneqq [a_{d,1}, a_{d,2}, \dots, a_{d,D}]^{\mathrm{T}} \in \mathbb{R}^{D \times 1}$ denotes the column vector formed by the elements in the dth row of A. Furthermore, for the reparametrisation trick, we take $q(e) \coloneqq \mathrm{N}(e; 0, \mathrm{I})$, where $0 \in \mathbb{R}^D$ is a vector whose elements are all 0, so that

$$h_{\phi,x}(e) \coloneqq Ax + b + C^{1/2} e,$$

which means that $h_{\phi,x}^{-1}(z) = C^{-1/2}(z - Ax - b)$.

Note that the mean of the proposal in (21) coincides with the mean of the posterior in (20) if $A = P$ and $b = P\varSigma^{-1}\mu$.

This model is similar to the one used as a benchmark in Rainforth et al. (2018, Section 4) and also in Tucker et al. (2019, Section 6.1) who specified both the generative model and the variational approximation to be isotropic Gaussians. Specifically, their setting can be recovered by taking $\varSigma \coloneqq \mathrm{I}$ and fixing $c_d = \log(2/3)/2$ so that $C = \frac{2}{3}\mathrm{I}$ throughout. Here, in order to investigate a slightly more realistic scenario, we also allow for the components of the latent vectors z to be *correlated/dependent* under the generative model. However, as the variational approximation remains restricted to being fully factored, it may fail to fully capture the uncertainty about the latent variables.

Gradient calculations. We end this subsection by stating the expressions needed to calculate the gradients in the Gaussian example presented above. Throughout, we use the *denominator-layout* notation for vector and matrix calculus and sometimes write $e = e_{1:D} = h_{\phi,x}^{-1}(z)$ to simplify the notation. Thus,

$$\begin{aligned} \nabla_\theta \log \gamma_{\theta,x}(z) &= \varSigma^{-1}(z - \mu) \in \mathbb{R}^D, \\ \nabla_z \log \gamma_{\theta,x}(z) &= \varSigma^{-1}(\mu - z) + x - z \in \mathbb{R}^D, \qquad (22) \\ \nabla_z \log q_{\phi,x}(z) &= -C^{-1}(z - Ax - b) \\ &= -C^{-1/2} e \in \mathbb{R}^D. \qquad (23) \end{aligned}$$

Let $a_d \coloneqq [a_{d,1}, a_{d,2}, \dots, a_{d,D}]^{\mathrm{T}} \in \mathbb{R}^{D \times 1}$ denote the column vector formed by the elements in the dth row of A. Then, letting $\odot$ denote elementwise multiplication and using the convention that addition or subtraction of the scalar 1 is to be done elementwise,

$$\begin{aligned} \nabla_{a_d} \log q_{\phi,x}(z) &= C^{-1}(z_d - a_d^{\mathrm{T}} x - b_d) x \\ &= C^{-1/2} e_d x \in \mathbb{R}^D, \quad d \in \{1, \dots, D\}, \\ \nabla_b \log q_{\phi,x}(z) &= C^{-1}(z - Ax - b) \\ &= C^{-1/2} e \in \mathbb{R}^D, \\ \nabla_c \log q_{\phi,x}(z) &= C^{-1/2}(z - Ax - b) \odot C^{-1/2}(z - Ax - b) - 1 \\ &= e \odot e - 1 \in \mathbb{R}^D, \end{aligned}$$

Furthermore, write $h_{\phi,x} = [h_{\phi,x,1}, \dots, h_{\phi,x,D}]^{\mathrm{T}}$, i.e.

$$h_{\phi,x,d}(e) = z_d = a_d^{\mathrm{T}} x + b_d + \exp(c_d) e_d,$$

and let $\iota^{(d)} = [0, \dots, 0, 1, 0, \dots, 0]^{\mathrm{T}} \in \mathbb{R}^D$ be the vector whose entries are all 0 except for the dth entry which is 1. Then, for $d \in \{1, \dots, D\}$,

$$\begin{aligned} [\nabla_{a_{d'}} h_{\phi,x,d}](e) &= \mathbf{1}\{d = d'\} x \in \mathbb{R}^D, \quad d' \in \{1, \dots, D\}, \qquad (24) \\ [\nabla_b h_{\phi,x,d}](e) &= \iota^{(d)} \in \mathbb{R}^D, \qquad (25) \\ [\nabla_c h_{\phi,x,d}](e) &= \exp(c_d) e_d \iota^{(d)} \in \mathbb{R}^D. \qquad (26) \end{aligned}$$

Again writing $e = h_{\phi,x}^{-1}(z)$ implies that

$$\nabla_\phi [\log \circ w_{\psi',x} \circ h_{\phi,x}]|_{\psi' = \psi}(e) = [\nabla_\phi h_{\phi,x,1}, \dots, \nabla_\phi h_{\phi,x,D}](e) \nabla_z \log w_{\psi,x}(z),$$

so that, letting $[\nabla_z \log w_{\psi,x}(z)]_d$ denote the dth element of the vector $\nabla_z \log w_{\psi,x}(z)$,

$$\begin{aligned}\nabla_{a_d}[\log \circ w_{\psi',x} \circ h_{\phi,x}]|_{\psi'=\psi}(e) &= [\nabla_z \log w_{\psi,x}(z)]_d x,\\ \nabla_b[\log \circ w_{\psi',x} \circ h_{\phi,x}]|_{\psi'=\psi}(e) &= \nabla_z \log w_{\psi,x}(z),\\ \nabla_c[\log \circ w_{\psi',x} \circ h_{\phi,x}]|_{\psi'=\psi}(e) &= e \odot C^{1/2}\nabla_z \log w_{\psi,x}(z).\end{aligned}$$

From this, since

$$\nabla_\phi[\log \circ w_{\psi,x} \circ h_{\phi,x}](e) = \nabla_\phi[\log \circ w_{\psi',x} \circ h_{\phi,x}]|_{\psi'=\psi}(e) - \nabla_\phi \log q_{\phi,x}(z),$$

we have that

$$\begin{aligned}\nabla_{a_d}[\log \circ w_{\psi,x} \circ h_{\phi,x}](e) &= ([\nabla_z \log w_{\psi,x}(z)]_d - C^{-1/2}e_d)x,\\ \nabla_b[\log \circ w_{\psi,x} \circ h_{\phi,x}](e) &= \nabla_z \log w_{\psi,x}(z) - C^{-1/2}e,\\ \nabla_c[\log \circ w_{\psi,x} \circ h_{\phi,x}](e) &= e \odot C^{1/2}\nabla_z \log w_{\psi,x}(z) - e \odot e + 1.\end{aligned}$$

Impact of the reparametrisation. We end this subsection by briefly illustrating the impact of the reparametrisation trick combined with the identity from Tucker et al. (2019) which was given in Lemma 1. Recall that this approach yields ϕ-gradients that are expressible as integrals of path-derivative functions $\blacktriangledown_{\psi,x} := \nabla_\phi[\log \circ w_{\psi',x} \circ h_{\phi,x}]|_{\psi'=\psi} \circ h_{\phi,x}^{-1}$, Thus, if there exists a value ϕ such that $q_{\phi,x} = \pi_{\theta,x}$ then $w_{\psi,x} \propto \pi_{\theta,x}/q_{\phi,x} \equiv 1$ is constant so that we obtain zero-variance ϕ-gradients (see, e.g., Roeder et al., 2017, for a discussion on this).

For simplicity, assume that $\Sigma = \mathrm{I}$ and recall that we then have $q_{\phi^\star,x} = \pi_{\theta,x}$ if the values (A, b, C) implied by $\phi^\star$ are $(A^\star, b^\star, C^\star) = (\frac{1}{2}\mathrm{I}, \frac{1}{2}\mu, \frac{1}{2}\mathrm{I})$.

By (22) and (23), and with the usual convention $e = h_{\phi,x}^{-1}(z)$, we then have

$$\begin{aligned}\nabla_z \log w_{\psi,x}(z) &= (x+\mu) - 2z + C^{-1}(z - Ax - b)\\ &= 2[(A^\star x + b^\star) - (Ax + b) + C^{-1/2}(C^\star - C)e]. \end{aligned} \tag{27}$$

Note that the only source of randomness in this expression is the multivariate normal random variable e. Thus, by (24) and (25), for *any* values of A and b and *any* $K \geq 1$, the variance of the A- and b-gradient portion of AISLE-KL/IWAE-STL and AISLE-χ^2/IWAE-DREG goes to zero as $C \to C^\star = \frac{1}{2}\mathrm{I}$. In other words, in this model, these 'score-function free' ϕ-gradients achieve (near) zero variance for the parameters governing the proposal mean as soon as the variance-parameters fall within a neighbourhood of their optimal values. Furthermore, (26) combined with (27) shows that for *any* $K \geq 1$, the variance of the C-gradient portion also goes to zero as $(A, b, C) \to (A^\star, b^\star, C^\star)$. A more thorough analysis of the benefits of reparametrisation-trick gradients in Gaussian settings is carried out in Xu et al. (2019).

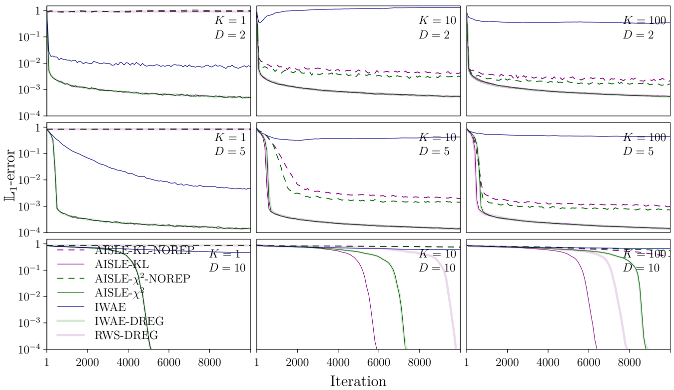
B.3 Simulations

Setup. We end this section by empirically comparing the algorithms from Subsection B.1. We run each of these algorithms for a varying number of particles, $K \in \{1, 10, 100\}$, and varying model dimensions, $D \in \{2, 5, 10\}$. Each of these configurations is repeated independently 250 times. Each time using a new synthetic data set consisting of $N = 25$ observations sampled from the generative model after generating a new 'true' prior mean vector as $\mu \sim \mathrm{N}(0, \mathrm{I})$. Since all the algorithms share the same θ-gradient, we focus only on the optimisation of ϕ and thus simply fix $\theta := \theta^{\mathrm{ML}}$ throughout. We show results for the following model settings.

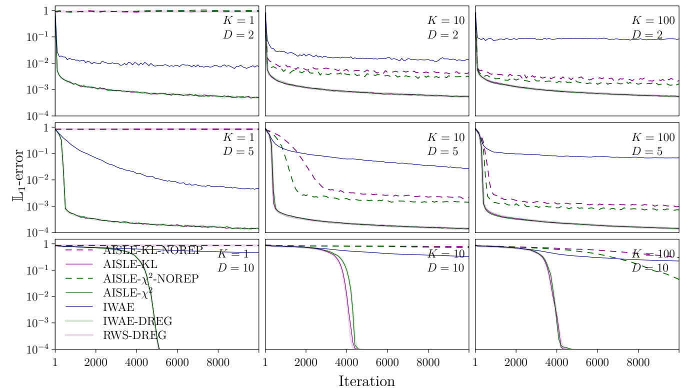
- **Figure 1.** The generative model is specified via $\Sigma = \mathrm{I}$. In this case, there exists a value $\phi^\star$ of ϕ such that $q_{\phi,x}(z) = \pi_{\theta,x}(z)$. Note that this corresponds to Scenario 1 in Subsection A.2.
- **Figure 2.** The generative model is specified via $\Sigma = (0.95^{|d-d'|+1})_{(d,d') \in \{1,\dots,D\}^2}$. Note that in this case, the fully-factored variational approximation cannot fully mimic the dependence structure of the latent variables under the generative model. That is, in this case, $q_{\phi,x}(z) \neq \pi_{\theta,x}(z)$ for any values of ϕ. Note that this corresponds to Scenario 2 in Subsection A.2.

To initialise the gradient-ascent algorithm, we draw each component of the initial values ϕ_0 of ϕ IID according to a standard normal distribution. We use both plain stochastic gradient-ascent with the gradients normalised to have unit $\mathbb{L}_1$-norm (Figures 1a, 1b, 2a, 2b) and ADAM (Kingma & Ba, 2015) with default parameter values (Figures 1c, 1d, 2c, 2d). In each case, we also show results for the 'regularised importance weights' strategy from Subsection A.3 with tuning parameter $\eta = 0.8$ (Figures 1b, 1d, 2b, 2d). The total number of iterations is $10,000$; in each case, the learning-rate parameters at the ith step are $i^{-1/2}$.

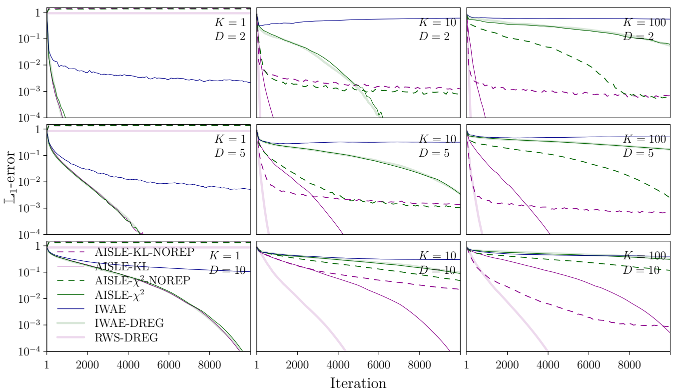
We also ran the algorithms in each of the above-mentioned scenarios with fixed values of c_d, e.g. as in Rainforth et al. (2018); Tucker et al. (2019). However, we omit the results as this did not significantly change the relative performance of the different algorithms. For the same reason, we omit results related to the optimisation of A and C.



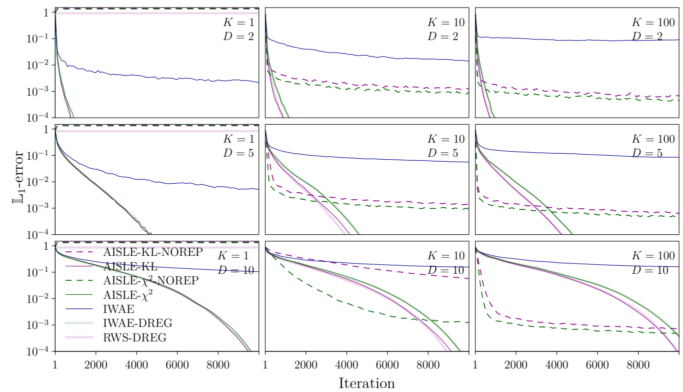
a. Gradient ascent with standard weights.



b. Gradient ascent with regularised weights.

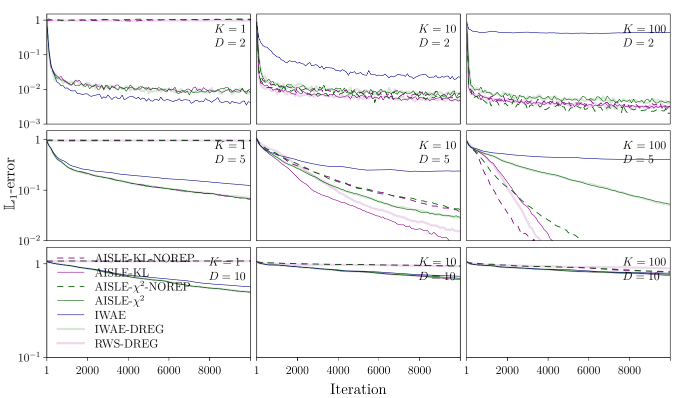


c. ADAM with standard weights.

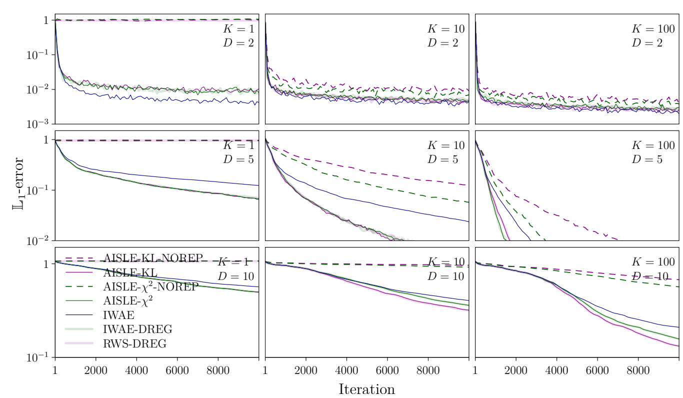


d. ADAM with regularised weights.

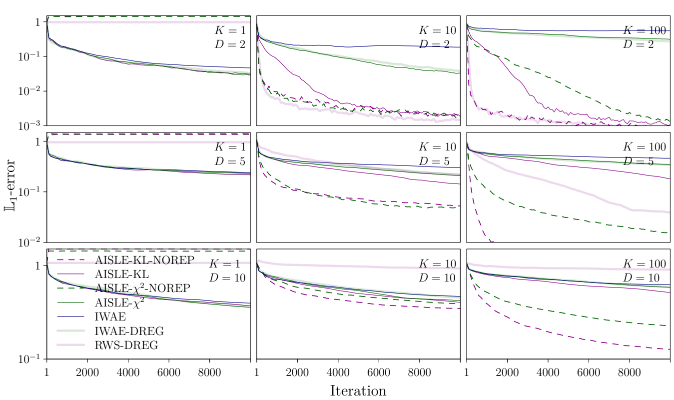
Figure 1. Average $\mathbb{L}_1$-error of the estimates of the parameters $b = b_{1:D}$ governing the mean of the Gaussian variational family. The average is taken over the D components of b and the figure displays the median error at each iteration over 100 independent runs of each algorithm, each using a different data set consisting of 25 observations sampled from the model. Note the logarithmic scaling on the second axis. Here, the covariance matrix $\Sigma = \mathrm{I}$ *is diagonal.*



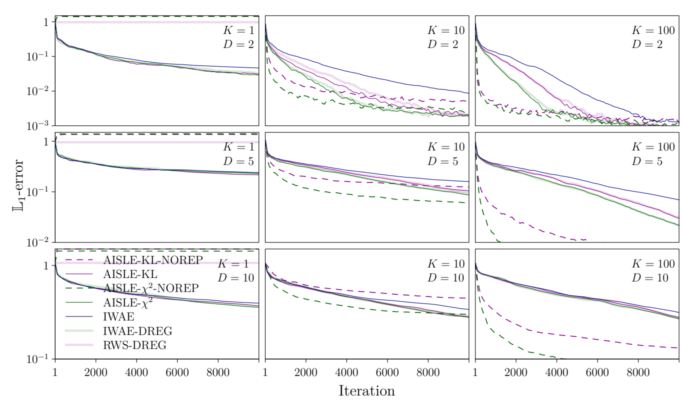
a. Gradient ascent with standard weights.



b. Gradient ascent with regularised weights.



c. ADAM with standard weights.



d. ADAM with regularised weights.

Figure 2. The same setting as in Figure 1 except that here, the covariance matrix $\Sigma = (0.95^{|d-e|+1})_{(d,e)\in\{1,\ldots,D\}^2}$ *is not a diagonal matrix.* Again, note the logarithmic scaling on the second axis.

Summary of results. Below, we outline what we believe to be the main takeaways from these simulation results for this particular model. However, further theoretical analysis is required to determine whether these hold in more general scenarios.

1. The KL-divergence based AISLE algorithms typically performed somewhat better than their χ^2-divergence based AISLE counterparts, i.e. AISLE-KL-NOREP outperformed AISLE-χ^2-NOREP while AISLE-KL outperformed AISLE-χ^2. We conjecture that this is due to the fact that the χ^2-divergence based variants square the (self-normalised) importance weights which increases the variance of the ϕ-gradients.
2. The performance of the ϕ-gradients AISLE-KL-NOREP and AISLE-χ^2-NOREP (which do not use/need any reparametrisation) typically benefited strongly from moderate (relative to the dimension of the latent variables) increases in the number of particles. When ADAM was used (and for larger K), these gradients outperformed the 'score-function free' ϕ-gradients AISLE-KL/IWAE-STL, AISLE-χ^2/ IWAE-DREG in the scenario shown in Figure 2c. We conjecture that this is due to the fact that the variational family does not include the target distribution in this scenario, i.e. $q_\phi \neq \pi_\theta$ for any ϕ, and as a result, the main advantage of the 'score-function free' gradients – i.e. the fact that they can potentially achieve zero variance – cannot be realised.
3. As expected, the performance of the standard IWAE ϕ-gradient consistently became worse with increasing K (see Figures 1a, 1c, 2a and 2c). This can be attributed to the fact that the signal-to-noise ratio of this gradient vanishes as $\mathcal{O}(K^{-1/2})$ as this gradient constitutes a self-normalised importance-sampling approximation of an integral which is equal to zero (see Rainforth et al. (2018) and also Subsection 2.2).
4. More surprisingly, the 'score-function free' ϕ-gradients AISLE-KL/IWAE-STL, AISLE-χ^2/IWAE-DREG (as well as the AISLE-χ^2-NOREP gradient in Figure 1c) did not appear to improve with increasing K. Indeed, their performance sometimes became worse. We note that this *cannot* be explained by the signal-to-noise ratio decay (which Rainforth et al. (2018) highlighted for the standard IWAE ϕ-gradient) because the 'score-function free' ϕ-gradients do not constitute self-normalised importance-sampling approximations of integrals which are equal to zero. Instead, we conjecture that as discussed in Remark 2 in this model, the $\mathcal{O}(K^{-1})$ self-normalisation bias of these gradients happens to be beneficial and outweighs the $\mathcal{O}(K^{-1/2})$ standard-deviation decrease obtained from increasing K. To counteract this issue, we also regularised the weights in each of these estimators as discussed in Subsection A.3, i.e. we replaced $w_{\psi,x}(z^k)/\sum_{l=1}^K w_{\psi,x}(z^l)$ by $w_{\psi,x}^{\alpha^\star}(z^k)/\sum_{l=1}^K w_{\psi,x}^{\alpha^\star}(z^l)$, where $\alpha^\star$ was determined as explained in Subsection A.3 with $\eta = 0.8$. Figures 1b, 1d. 2b and 2d show that this regularisation strategy appears to be especially beneficial to the χ^2-divergence based AISLE gradients.
5. The 'doubly-reparametrised' RWS-gradient RWS-DREG from Tucker et al. (2019) and given in (9) performed well for a moderate to large number of particles $K > 1$ in settings in which the oscillation of the importance-weight function, w_ψ, is relatively small (or at least if it becomes small as ϕ is optimised). However, this requires that $q_{\phi,x}$ can be made very close to $\pi_{\theta,x}$ for an appropriate choice of ϕ which is typically only possible in low-dimensional settings and if the variational family is sufficiently expressive, i.e. in the scenario from Figure 1. Otherwise, e.g. in dimension $D = 10$ in the scenario from Figures 2, the performance of RWS-DREG was worse than that of any AISLE variants and also worse than the standard IWAE reparametrisation-trick gradient. We conjecture that this is because the variance of the weights is so large that typically one of the self-normalised weights $w_{\psi,x}(z^k)/\sum_{l=1}^K w_{\psi,x}(z^l)$ is numerically equal to 1 while all the others are numerically equal to 0. Note that whenever this happens, the RWS-DREG gradient reduces to a vector of 0s. Again, Figures 2b and 2d show that the regularisation strategy from Subsection A.3 alleviates this problem (though Figures 1c and 1d make it clear that RWS-DREG does not necessarily benefit from this kind of regularisation under all circumstances).