Unbiased Gradient Estimation for Variational Auto-Encoders using Coupled Markov Chains

Francisco J. R. Ruiz[1] **Michalis K. Titsias**[1] **Taylan Cemgil**[1] **Arnaud Doucet**[1]

[1] DeepMind

Abstract

The variational auto-encoder (VAE) is a deep latent variable model that has two neural networks in an autoencoder-like architecture; one of them parameterizes the model's likelihood. Fitting its parameters via maximum likelihood (ML) is challenging since the computation of the marginal likelihood involves an intractable integral over the latent space; thus the VAE is trained instead by maximizing a variational lower bound. Here, we develop a ML training scheme for VAEs by introducing unbiased estimators of the log-likelihood gradient. We obtain the estimators by augmenting the latent space with a set of importance samples, similarly to the importance weighted auto-encoder (IWAE), and then constructing a Markov chain Monte Carlo coupling procedure on this augmented space. We provide the conditions under which the estimators can be computed in finite time and with finite variance. We show experimentally that VAEs fitted with unbiased estimators exhibit better predictive performance.

1 INTRODUCTION

The variational auto-encoder (VAE) [Kingma and Welling, 2014] is a deep latent variable model that uses a joint distribution $p_\theta(x, z)$, parameterized by θ, over an observation x and the corresponding latent variable z. The marginal log-likelihood involves an integral over the latent space,

$$\mathcal{L}(\theta) := \log p_\theta(x) = \log \left(\int p_\theta(x, z) \mathrm{d}z \right). \qquad (1)$$

As for any other latent variable model, fitting the VAE requires finding the parameters θ that best describe the observations. One (intractable) way to find θ would be via maximum likelihood, for which the gradient of Eq. 1 is required. Using Fisher's identity, this gradient can be written as an expectation w.r.t. the posterior $p_\theta(z \,|\, x)$,

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{p_\theta(z \,|\, x)} \left[\nabla_\theta \log p_\theta(x, z) \right]. \qquad (2)$$

The gradient in Eq. 2 could be approximated unbiasedly if we had access to samples from $p_\theta(z \,|\, x)$; however, the posterior is intractable. Although we could use Markov chain Monte Carlo (MCMC) to sample approximately from it [Hoffman, 2017, Naesseth et al., 2020], this would provide a biased estimate, and the bias is difficult to quantify.

Instead, VAEs introduce an encoder $q_\phi(z \,|\, x)$ and learn the parameters θ by maximizing a variational evidence lower bound (ELBO) [Wainwright and Jordan, 2008, Blei et al., 2017], for which unbiased gradients are readily available [Kingma and Welling, 2014]. Burda et al. [2016] form such a bound using a set of importance samples using the encoder as proposal distribution, leading to the so-called importance weighted auto-encoder (IWAE). The standard ELBO in variational inference can be thought of as a particular instance of the IWAE bound with one importance sample, where the importance distribution is given by the encoder. However, it remains difficult to quantify the difference between the true log-likelihood and the corresponding bound.

We develop here unbiased gradient estimators of the log-likelihood for VAEs by exploiting the coupling estimators developed by Jacob et al. [2020b]. Coupling estimators allow us to obtain unbiased estimators of expectations w.r.t. an intractable target distribution by running two coupled MCMC chains for a finite number of iterations. This approach does not require that the MCMC chains converge to the target, so the unbiased estimator can be computed in a finite (but random) time. However, MCMC couplings are a generic methodology that is not readily applicable to VAEs, since it requires an MCMC kernel that mixes well and provides a suitable coupling mechanism, while at the same time yielding a low-variance estimator.

We address these issues by building coupling estimators based on the iterated sampling importance resampling (ISIR) algorithm [Andrieu et al., 2010]. ISIR is in turn based on

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

importance sampling, and thus it uses multiple samples to reduce the variance of the estimator, similarly to the IWAE. Like the IWAE, we simultaneously fit an encoder as the proposal distribution for the importance sampling algorithm. Unfortunately, the variance of ISIR is still not small enough for fitting VAEs. To that end, we develop an extension of ISIR, called dependent iterated sampling importance resampling (DISIR), that combines the use of dependent importance samples and reparameterization ideas. DISIR drastically reduces the running time and the variance of the resulting coupling estimator. We develop a DISIR-based coupling estimator to estimate the gradient of the VAE log-likelihood.

Contributions. Our contributions are as follows.

- We show that the unbiased gradient estimators based on ISIR are not practical for VAEs, since they require to run the Markov chains for a long time even for only moderately high-dimensional models.
- We develop DISIR, an extension of ISIR that reduces the running time and the variance of the estimators.
- We use DISIR to form unbiased gradient estimators for VAEs. The resulting estimator is widely applicable as it can be used wherever the IWAE bound is applicable.
- We prove that the unbiased gradient estimates have finite variance and can be computed in finite expected time under some regularity conditions.
- We demonstrate experimentally that VAEs fitted with the DISIR-based gradient estimators exhibit better predictive log-likelihood on binarized MNIST, fashion-MNIST, and CIFAR-10, when compared to IWAE.

2 BACKGROUND

Here, we review the IWAE bound [Burda et al., 2016] and show that it can be seen as a standard ELBO on an augmented model. Consider a model $p_\theta(x, z)$ of data x and latent variables $z \in \mathcal{Z}$, and a proposal distribution $q_\phi(z \,|\, x)$, where θ and ϕ denote the model and proposal parameters, respectively. Let the proposal satisfy the assumption below.

Assumption 1 (The proposal and posterior have the same support)**.** *For any $z \in \mathcal{Z}$, we have $q_\phi(z \,|\, x) > 0$ if and only if $p_\theta(x, z) > 0$, so that $0 < w_{\theta,\phi}(z) < \infty$, where the* importance weights *are*

$$w_{\theta,\phi}(z) := \frac{p_\theta(x, z)}{q_\phi(z \,|\, x)}. \tag{3}$$

The IWAE bound. The IWAE is a lower bound of the marginal log-likelihood in Eq. 1 formed with $K \geq 1$ importance samples $z_{1:K}$ from the proposal, i.e., $\mathcal{L}(\theta) \geq \mathcal{L}_{\text{IWAE}}(\theta, \phi)$, with

$$\mathcal{L}_{\text{IWAE}}(\theta, \phi) = \mathbb{E}_{q_\phi(z_{1:K} \,|\, x)}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K} w_{\theta,\phi}(z_k)\right)\right], \tag{4}$$

where $q_\phi(z_{1:K} \,|\, x) := \prod_{k=1}^{K} q_\phi(z_k \,|\, x)$. Eq. 4 monotonically increases with K, converging towards $\mathcal{L}(\theta)$ as $K \to \infty$. For the case $K = 1$, it recovers the standard ELBO,

$$\mathcal{L}_{\text{ELBO}}(\theta, \phi) = \mathbb{E}_{q_\phi(z \,|\, x)}\left[\log w_{\theta,\phi}(z)\right]. \tag{5}$$

The importance samples $z_{1:K}$ also provide an approximation of the posterior $p_\theta(z \,|\, x)$. Specifically, if we define the importance weights $w_{\theta,\phi}^{(k)} := w_{\theta,\phi}(z_k)$ and the normalized importance weights $\widetilde{w}_{\theta,\phi}^{(k)} \propto w_{\theta,\phi}^{(k)}$, with $\sum_{k=1}^{K} \widetilde{w}_{\theta,\phi}^{(k)} = 1$, then the approximation is

$$\hat{p}_\theta(z \,|\, x) = \sum_{k=1}^{K} \widetilde{w}_{\theta,\phi}^{(k)} \delta_{z_k}(z), \tag{6}$$

where $\delta_{z_k}(\cdot)$ is the delta Dirac measure located at z_k.

Fitting VAEs. VAEs parameterize the likelihood $p_\theta(x \,|\, z)$ using a distribution whose parameters are given by a neural network (decoder) that inputs the latent variable z. The distribution $q_\phi(z \,|\, x)$ is amortized [Gershman and Goodman, 2014], i.e., its parameters are computed by a neural network (encoder) that inputs the observation x. Fitting a VAE involves maximizing the bound (either Eq. 4 or Eq. 5) w.r.t. both θ and ϕ using stochastic optimization. For that, the VAE uses unbiased gradient estimators of the objective. To form such gradients, we typically assume that $q_\phi(z \,|\, x)$ is reparameterizable [Kingma and Welling, 2014, Rezende et al., 2014, Titsias and Lázaro-Gredilla, 2014]. For convenience, we additionally assume that we can reparameterize in terms of a Gaussian (but we can easily relax this latter assumption).

Assumption 2 (The variational distribution is reparameterizable in terms of a Gaussian)**.** *There exists a mapping $g_\phi(\xi, x)$ such that by sampling $\xi \sim q(\xi)$, where $q(\xi) = \mathcal{N}(\xi; 0, I)$, and setting $z = g_\phi(\xi, x)$, we obtain $z \sim q_\phi(z \,|\, x)$.*

The IWAE bound as a standard ELBO. The IWAE bound in Eq. 4 can be interpreted as a regular ELBO on an augmented latent space [Cremer et al., 2017, Domke and Sheldon, 2018], and we use this perspective in Sections 3 and 4. Indeed, consider the K importance samples $z_{1:K}$ and an indicator variable $\ell \in \{1, \ldots, K\}$. We next define a generative model with latent variables $(z_{1:K}, \ell)$, as well as a variational distribution, such that its ELBO recovers Eq. 4.

The augmented generative model posits that the indicator $\ell \sim \text{Cat}(\frac{1}{K}, \ldots, \frac{1}{K})$, where Cat denotes the categorical distribution. Given ℓ, each z_k is distributed according to $q_\phi(z \,|\, x)$, except the ℓ-th one, which follows the prior:

$$p_{\theta,\phi}(x, z_{1:K}, \ell) = \frac{1}{K}\, p_\theta(x, z_\ell) \prod_{k=1, k\neq\ell}^{K} q_\phi(z_k \,|\, x). \tag{7}$$

Under the corresponding augmented posterior distribution, $p_{\theta,\phi}(z_{1:K}, \ell \,|\, x) \propto p_{\theta,\phi}(x, z_{1:K}, \ell)$, the random variable z_ℓ

follows the posterior $p_\theta(z \,|\, x)$. We next define a variational distribution on the same augmented space,

$$q_{\theta,\phi}(z_{1:K}, \ell \,|\, x) = \mathrm{Cat}\Big(\ell \,|\, \widetilde{w}^{(1)}_{\theta,\phi}, \cdots, \widetilde{w}^{(K)}_{\theta,\phi}\Big) \prod_{k=1}^{K} q_\phi(z_k \,|\, x). \tag{8}$$

We recover the IWAE bound (Eq. 4) as the ELBO of the augmented model, i.e., as $\mathbb{E}_{q_{\theta,\phi}(z_{1:K}, \ell \,|\, x)}\left[\log p_{\theta,\phi}(x, z_{1:K}, \ell)\right]$.

Towards unbiased gradient estimation. The gradient w.r.t. θ of the IWAE bound in Eq. 4 can be interpreted as a (biased) approximation of $\nabla_\theta \mathcal{L}$ from Eq. 2. To see this, note that $\nabla_\theta \mathcal{L}_{\text{IWAE}} = \mathbb{E}_{\hat{p}_\theta(z \,|\, x)}\left[\nabla_\theta \log p_\theta(x, z)\right]$; i.e., $\nabla_\theta \mathcal{L}_{\text{IWAE}}$ can be seen as an approximation of $\nabla_\theta \mathcal{L}$ where we replace the posterior $p_\theta(z \,|\, x)$ with the approximation $\hat{p}_\theta(z \,|\, x)$ in Eq. 6.

To obtain an unbiased estimate, we need an alternative approximation $\hat{p}_\theta(z \,|\, x)$ of $p_\theta(z \,|\, x)$ satisfying $\mathbb{E}_{\hat{p}_\theta(z \,|\, x)}\left[\nabla_\theta \log p_\theta(x, z)\right] = \nabla_\theta \mathcal{L}$. One such example is the empirical measure of exact samples from $p_\theta(z \,|\, x)$. However, as mentioned earlier, it is typically impossible to obtain such samples, as a finite run with an MCMC kernel only provides biased estimates when the chain is initialized out of equilibrium. We can obtain an approximation $\hat{p}_\theta(z \,|\, x)$ that leads to unbiased estimation using two coupled MCMC chains [Jacob et al., 2020b]. In Sections 3 and 4 we build on this idea to develop a method for unbiased gradient estimation.

3 IMPORTANCE SAMPLING-BASED MCMC SCHEMES

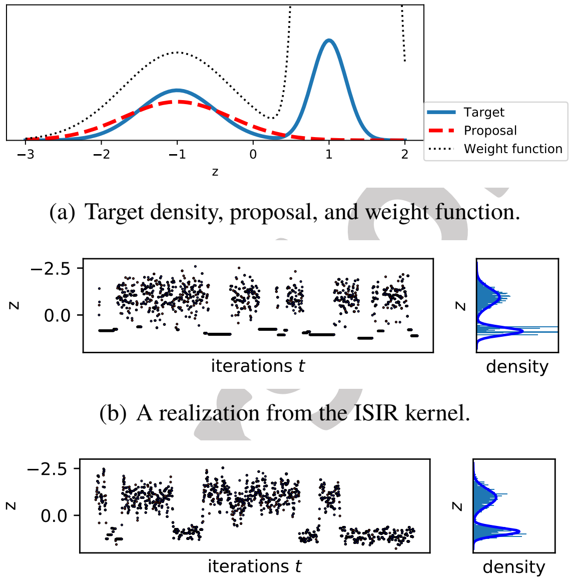
3.1 ITERATED SAMPLING IMPORTANCE RESAMPLING (ISIR)

Here we review ISIR [Andrieu et al., 2010], an MCMC scheme that samples from the augmented posterior $p_{\theta,\phi}(z_{1:K}, \ell \,|\, x)$ given by Eq. 7. Alternatively, ISIR can be interpreted as an algorithm that targets the posterior $p_\theta(z \,|\, x)$. Indeed, if $(z_{1:K}, \ell) \sim p_{\theta,\phi}(z_{1:K}, \ell \,|\, x)$ is a sample from the augmented posterior, then $z_\ell \sim p_\theta(z \,|\, x)$.

The ISIR transition kernel, $\mathcal{K}_{\text{ISIR}}(\cdot, \cdot \,|\, z_{1:K}, \ell)$, takes the current state $(z_{1:K}, \ell)$ and outputs a new state by following Algorithm 1. ISIR requires a proposal distribution; we use $q_\phi(z \,|\, x)$. It also requires to compute the importance weights $w_{\theta,\phi}(\cdot)$ in Eq. 3. Note that ISIR has a resampling step (Line 4) that is distinct from the resampling step of sequential Monte Carlo, which would involve resampling multiple times from the categorical before mutating the samples.

The kernel is invariant w.r.t. $p_{\theta,\phi}(z_{1:K}, \ell \,|\, x)$, as formalized in Proposition 5 in Appendix C.1, if Assumption 3 below is satisfied [Andrieu et al., 2010, 2018]. Assumption 3 holds when the proposal is at least as heavy-tailed as the target.

Assumption 3 (The importance weights are bounded). *There exists $w^{max}_{\theta,\phi} < \infty$ such that $w_{\theta,\phi}(z) \leq w^{max}_{\theta,\phi}$ $\forall z \in \mathcal{Z}$.*



(a) Target density, proposal, and weight function.

(b) A realization from the ISIR kernel.

(c) A realization from the DISIR kernel.

Figure 1: Qualitative comparison of ISIR and DISIR targetting a simple target distribution (a), and the realized chains by sampling from ISIR (b) and DISIR (c) transition kernels. Due to the poor proposal choice, the weight function significantly varies across the space z. ISIR has a low acceptance probability, especially in high-weight states. In contrast, DISIR is able to propose and accept local moves around high-weight regions and explores the target better.

3.2 DEPENDENT ITERATED SAMPLING IMPORTANCE RESAMPLING (DISIR)

For moderately high-dimensional z, ISIR can be inefficient. Indeed, if the importance weights become dominated by the weight of a single sample, then the corresponding Markov chain will typically get "stuck" for a large number of iterations. We illustrate this in Figure 1, which shows a one-dimensional illustrative comparison. In this toy experiment, we can observe that the ISIR chains gets stuck when the state $z^{(t)}$ corresponds to a high-weight region.

To mitigate this problem, here we develop DISIR, an extension of ISIR that uses dependent importance samples. Intuitively, this scheme proposes dependent samples $z^\star_{1:\ell_{\text{aux}}-1}, z^\star_{\ell_{\text{aux}}+1:K}$ that are close to the current sample $z^\star_{\ell_{\text{aux}}} = z_\ell$ with high probability, in the spirit of Shestopaloff and Neal [2018]. This modification increases the probability that the chain transitions to one of the new proposed values. As a result, in practice, the gradient estimators based on DISIR have smaller variance than the ones based on ISIR.

To make samples dependent, we use the reparameterization property (see Assumption 2) and introduce dependencies among the auxiliary variables $\xi_{1:K}$; this induces dependencies among the samples in the original latent space $z_{1:K}$.

Algorithm 1: ISIR kernel, $\mathcal{K}_{\text{ISIR}}(\cdot, \cdot \mid z_{1:K}, \ell)$

```
Input: Current state of the chain, $(z_{1:K}, \ell)$
Output: Next state of the chain
1 Sample $\ell_{\text{aux}} \sim \text{Cat}(\frac{1}{K}, \ldots, \frac{1}{K})$
2 Set $z^\star_{\ell_{\text{aux}}} = z_\ell$
3 Sample $z^\star_k \sim q_\phi(z \mid x)$ for $k \in \{1, \ldots, K\} \backslash \{\ell_{\text{aux}}\}$
4 Sample $\ell^\star \sim \text{Cat}(p_1, \ldots, p_K)$ with $p_k \propto w_{\theta,\phi}(z^\star_k)$
5 Return $(z^\star_{1:K}, \ell^\star)$
```

That is, rather than sampling $K-1$ importance samples $(\xi^\star_{1:\ell_{\text{aux}}-1}, \xi^\star_{\ell_{\text{aux}}+1:K})$ independently of $\xi^\star_{\ell_{\text{aux}}}$ (see Line 3 of Algorithm 1), we use two auxiliary Markov chains, each with transition kernel $p_\beta(\xi^\star \mid \xi)$. Specifically, given $\xi^\star_{\ell_{\text{aux}}}$, we sample $\xi^\star_k \sim p_\beta(\cdot \mid \xi^\star_{k-1})$ for $k > \ell_{\text{aux}}$ and $\xi^\star_k \sim p_\beta(\cdot \mid \xi^\star_{k+1})$ for $k < \ell_{\text{aux}}$. The kernel $p_\beta(\xi^\star \mid \xi)$ must have invariant density $q(\xi)$, so that marginally $\xi^\star_k \sim q(\xi)$ for all k if $\xi^\star_{\ell_{\text{aux}}} \sim q(\xi)$. This construction with two Markov chains ensures the validity of the scheme (see Appendix C.2 for details).

Under Assumption 2, we select a simple autoregressive normal kernel for $p_\beta(\xi^\star \mid \xi)$, i.e., we set

$$p_\beta(\xi^\star \mid \xi) = \mathcal{N}(\xi^\star; \beta\xi, (1-\beta^2)I). \tag{9}$$

Equivalently, $\xi^\star = \beta\xi + \sqrt{1-\beta^2}\xi^{\text{new}}$, where $\xi^{\text{new}} \sim \mathcal{N}(\xi; 0, I)$. The parameter β controls the strength of the correlation; we discuss its effect below.

Using the kernel in Eq. 9, we develop DISIR, which is described in Algorithm 2. DISIR replaces Line 3 of Algorithm 1 with an application of the auxiliary kernel $p_\beta(\xi^\star \mid \xi)$ (see Lines 4 and 5 of Algorithm 2).

We refer to the coefficient $\beta \in [0, 1)$ as the *correlation strength*. When $\beta = 0$, we have $\xi^\star_k = \xi^{\text{new}}_k \overset{\text{iid}}{\sim} q(\xi)$ and this approach is simply a reparameterized version of ISIR, which favours exploration of new regions of the space. When β approaches 1, all the proposed values $\xi^\star_k$ become closer to the current state ξ_ℓ (which may correspond to the sample whose importance weight currently dominates). This dependency among the samples $\xi^\star_{1:K}$ induces dependencies among $z^\star_{1:K}$, resulting in more uniform importance weights. Thus, we say that this approach favours exploitation.

As given in Algorithm 2, DISIR is an MCMC scheme that targets the augmented density provided below.

Proposition 1 (Invariant distribution of DISIR). *Let Assumptions 1 and 2 hold. For any $K \geq 2$ and any $\beta \in [0, 1)$, the DISIR transition kernel $\mathcal{K}_{DISIR}$ admits*

$$p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K}, \ell \mid x) = \frac{1}{K}\frac{w_{\theta,\phi}(g_\phi(\xi_\ell, x))q(\xi_\ell)}{p_\theta(x)} \times \prod_{k=1}^{\ell-1} p_\beta(\xi_k \mid \xi_{k+1}) \prod_{k=\ell+1}^{K} p_\beta(\xi_k \mid \xi_{k-1}) \tag{10}$$

as invariant distribution and is ergodic.

Algorithm 2: DISIR kernel, $\mathcal{K}_{\text{DISIR}}(\cdot, \cdot \mid \xi_{1:K}, \ell)$

```
Input: Current state of the chain $(\xi_{1:K}, \ell)$ and
       correlation strength $\beta$
Output: New state of the chain
1 Sample $\ell_{\text{aux}} \sim \text{Cat}(\frac{1}{K}, \ldots, \frac{1}{K})$
2 Set $\xi^\star_{\ell_{\text{aux}}} = \xi_\ell$
3 Sample $\xi^{\text{new}}_k \sim q(\xi)$ for $k \in \{1, \ldots, K\} \backslash \{\ell_{\text{aux}}\}$
4 Set $\xi^\star_k = \beta\xi^\star_{k-1} + \sqrt{1-\beta^2}\xi^{\text{new}}_k$ for
    $k = \ell_{\text{aux}} + 1, \ldots, K$
5 Set $\xi^\star_k = \beta\xi^\star_{k+1} + \sqrt{1-\beta^2}\xi^{\text{new}}_k$ for
    $k = \ell_{\text{aux}} - 1, \ldots, 1$
6 Set $z^\star_k = g_\phi(\xi^\star_k, x)$ for $k = 1, \ldots, K$
7 Sample $\ell^\star \sim \text{Cat}(p_1, \ldots, p_K)$ with $p_k \propto w_{\theta,\phi}(z^\star_k)$
8 Return $(\xi^\star_{1:K}, \ell^\star)$
```

The proofs of all propositions and lemmas are provided in Appendix C.

This target distribution has two desired properties. First, when the correlation strength $\beta \in [0, 1)$, the ℓ-th sample z_ℓ is distributed according to the posterior $p_\theta(z \mid x)$. That is, DISIR defines a Markov chain that targets Eq. 10, and $(z^{(t)}_{\ell^{(t)}})_{t \geq 0}$ is a Markov chain that converges to $p_\theta(z \mid x)$. Second, when $\beta = 0$, DISIR becomes identical to a reparameterized version of ISIR. That is, it simulates a Markov chain $(\xi^{(t)}_{1:K}, \ell^{(t)})_{t \geq 0}$ such that, setting each $z^{(t)}_k = g_\phi(\xi^{(t)}_k, x)$, the Markov chain $(z^{(t)}_{1:K}, \ell^{(t)})_{t \geq 0}$ obeys a law that is identical to the one simulated by ISIR. This is formalized below.

Lemma 1 (Distribution of DISIR samples). *Let Assumptions 1 and 2 hold. For any $\beta \in [0, 1)$, we have $z_\ell = g_\phi(\xi_\ell, x) \sim p_\theta(z \mid x)$ under $p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K}, \ell \mid x)$. Moreover, for $\beta = 0$, if $(\xi_{1:K}, \ell) \sim p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K}, \ell \mid x)$, then we have $(z_{1:K}, \ell) \sim p_{\theta,\phi}(z_{1:K}, \ell \mid x)$, where each $z_k = g_\phi(\xi_k, x)$.*

In practice, we interleave DISIR steps for which $\beta > 0$ with steps for which $\beta = 0$. Specifically, we define a composed kernel that consists of the consecutive application of two steps of Algorithm 2. The first step has $\beta = 0$ and favours exploration; the second step has $\beta > 0$ and favours exploitation. It is possible to interleave the two kernels since both can be reinterpreted as MCMC kernels targeting $p_\theta(z \mid x)$. We denote the composed kernel as $\mathcal{K}_{\text{ISIR-DISIR}}$.

Choice of the correlation strength. For the second step of the composed kernel, we wish to use a value β close to 1 to achieve exploitation, but not too close because then we will effectively have one importance sample repeated K times. We set β following a heuristic that is based on the effective sample size (ESS), defined as $\text{ESS} = (\sum_{k=1}^{K}(\widetilde{w}^{(k)}_{\theta,\phi})^2)^{-1}$. As β becomes closer to 1, the ESS becomes closer to K. We set the target ESS to $0.3K$, and we update β after each step of the kernel, so that the ESS becomes closer to the target value. In particular, we apply the update rule $\beta \leftarrow$

$\beta - 0.01(\mathrm{ESS} - 0.3K)$. We also constrain the resulting $\beta \in [10^{-6}, 1 - 10^{-6}]$ to avoid numerical issues.

Since we only modify β between iterations of the algorithm (and not while running the algorithm), the invariance of the Markov kernel w.r.t. to the target still holds. This also implies that the estimators that we derive in Section 4 are unbiased despite the adaptation of β.

3.3 DISIR ESTIMATES OF EXPECTATIONS

The DISIR samples $(z^{(t)}_{\ell^{(t)}})_{t\geq 0}$ are distributed asymptotically as $p_\theta(z\,|\,x)$ and, in the limit of infinite samples, we can estimate expectations under $p_\theta(z\,|\,x)$, as $\frac{1}{T+1}\sum_{t=0}^{T} h(z^{(t)}_{\ell^{(t)}}) \to \mathbb{E}_{p_\theta(z\,|\,x)}[h(z)]$ almost surely.

However, it may seem wasteful to generate $K-1$ proposals at each iteration of Algorithm 2 and then use only the ℓ-th importance sample to estimate an expectation. The proposition below shows that it is possible to use all the K importance samples: we can estimate an expectation $\mathbb{E}_{p_\theta(z\,|\,x)}[h(z)]$ with a weighted average of the importance samples.

Proposition 2 (The K importance samples can be used for estimating expectations)**.** *For any function $h : \mathcal{Z} \to \mathbb{R}$ such that $\mathbb{E}_{p_\theta(z\,|\,x)}[|h(z)|] < \infty$, we have the identity*

$$\mathbb{E}_{p_\theta(z\,|\,x)}[h(z)] = \mathbb{E}_{p^{\mathrm{DISIR}}_{\theta,\phi}(\xi_{1:K},\ell\,|\,x)}\left[\sum_{k=1}^{K} \widetilde{w}^{(k)}_{\theta,\phi} h(z_k)\right], \quad (11)$$

where $z_k = g_\phi(\xi_k, x)$ and the normalized importance weights are $\widetilde{w}^{(k)}_{\theta,\phi} \propto w_{\theta,\phi}(z_k)$ with $\sum_{k=1}^{K}\widetilde{w}^{(k)}_{\theta,\phi} = 1$. Setting $h(z) = \nabla_\theta \log p_\theta(x,z)$, and given that the DISIR kernel is ergodic, it follows from Eq. 2 that

$$\frac{1}{T+1}\sum_{t=0}^{T}\left[\sum_{k=1}^{K} \widetilde{w}^{(k,t)}_{\theta,\phi} \nabla_\theta \log p_\theta(x, z^{(t)}_k)\right] \to \nabla_\theta \mathcal{L} \quad (12)$$

almost surely as $T \to \infty$ for any $K \geq 2$, where $z^{(t)}_k = g_\phi(\xi^{(t)}_k, x)$ and $\widetilde{w}^{(k,t)}_{\theta,\phi} \propto w_{\theta,\phi}(z^{(t)}_k)$.

3.4 CHOICE OF THE PROPOSAL

Both ISIR and DISIR require a proposal distribution to sample the states $z^\star_k$, for which we use $q_\phi(z\,|\,x)$ (DISIR additionally requires the proposal to be reparameterizable). Here we discuss how to set the parameters ϕ of the proposal.

Like for IWAE, in our case $q_\phi(z\,|\,x)$ is a proposal distribution rather than a variational posterior. We fit ϕ via stochastic optimization of the IWAE bound in Eq. 4. To estimate the gradient w.r.t. ϕ, we use the doubly reparameterized estimator [Tucker et al., 2019], which addresses some issues of the estimator of Burda et al. [2016] for large values of K [Rainforth et al., 2019].

Alternatively, we could fit the encoder using the forward Kullback-Leibler (KL) divergence. This would imply to maximize $\mathbb{E}_{p_\theta(z\,|\,x)}[\log q_\phi(z\,|\,x)]$ w.r.t. ϕ, for which we can apply the unbiased estimators of Section 4 to estimate the expectation w.r.t. $p_\theta(z\,|\,x)$. We leave this for future work.

4 UNBIASED GRADIENT ESTIMATION WITH MCMC COUPLINGS

4.1 UNBIASED ESTIMATION WITH COUPLINGS

In this section, we review how to obtain an unbiased gradient estimator using two coupled Markov chains. We use the notation $u \in \mathcal{U}$ to refer to a generic random variable, keeping in mind that we will later set $u = [\xi_{1:K}, \ell]$, i.e., the latent variables in the augmented space.

Consider the estimation of the expectation

$$H := \mathbb{E}_{\pi(u)}\left[h(u)\right], \quad (13)$$

for some distribution $\pi(u)$ and function $h(u)$. As discussed in Section 2, a direct approximation $\hat{\pi}(u)$ of $\pi(u)$ via MCMC leads to a biased estimator.

We can obtain an unbiased estimator based on two coupled MCMC chains, each with invariant distribution $\pi(\cdot)$ [Glynn and Rhee, 2014, Jacob et al., 2020b]. The two chains have the same marginals at any time instant t, but they evolve according to a joint transition kernel $\mathcal{K}_{\mathrm{C}}$. Let $\mathcal{K}(\cdot\,|\,u)$ be the marginal transition kernel of each chain, and let $\mathcal{K}_{\mathrm{C}}(\cdot,\cdot\,|\,u,\bar{u})$ be a joint kernel that takes the state of both chains (denoted u and $\bar{u}$) and produces the new state of both chains.[1]

We next review the unbiased estimator of Vanetti and Doucet [2020] which reduces the variance of the estimator of Jacob et al. [2020b]. The main idea is to consider a lag $L \geq 1$ and jointly sample the states of both chains $(u^{(t)}, \bar{u}^{(t-L)})$ conditioned on their previous states, i.e., $(u^{(t-1)}, \bar{u}^{(t-L-1)})$. We then use (a finite number of) the samples from each chain to obtain the unbiased estimator (Appendix A shows how to derive it). Practically, we initialize the first Markov chain from some (arbitrary) initial distribution $\pi_0(\cdot)$, i.e., $u^{(0)} \sim \pi_0(u)$. We then sample this Markov chain using the marginal kernel $\mathcal{K}$, i.e., $u^{(t)} \sim \mathcal{K}(u\,|\,u^{(t-1)})$ for $t = 1, \ldots, L$. After L steps, we draw the initial state of the second Markov chain $\bar{u}^{(0)}$ (potentially conditionally upon $u^{(L-1)}, u^{(L)}$), such that marginally $\bar{u}^{(0)} \sim \pi_0(u)$. Afterwards, for $t > L$, we draw both states jointly as $(u^{(t)}, \bar{u}^{(t-L)}) \sim \mathcal{K}_{\mathrm{C}}(u, \bar{u}\,|\,u^{(t-1)}, \bar{u}^{(t-L-1)})$.

The joint kernel $\mathcal{K}_{\mathrm{C}}$ is chosen such that, after some time, both chains produce the same exact realizations of the random variables, i.e., $u^{(t)} = \bar{u}^{(t-L)}$ for $t \geq \tau$. Here, τ is the *meeting time*, defined as the first time instant in which both

[1] The joint kernel is such that $\mathcal{K}_{\mathrm{C}}(A, \mathcal{U}\,|\,u,\bar{u}) = \mathcal{K}(A\,|\,u)$ and $\mathcal{K}_{\mathrm{C}}(\mathcal{U}, A\,|\,u,\bar{u}) = \mathcal{K}(A\,|\,\bar{u})$ for any measurable set A.

chains meet, $\tau = \inf\{t \geq L : u^{(t)} = \bar{u}^{(t-L)}\}$ (it could be infinite, but we design the joint kernel so that τ is a random variable of finite expected value).

Based on this coupling procedure, the unbiased estimator of Eq. 13 by Vanetti and Doucet [2020] is (see Appendix A)

$$\hat{H} = \frac{1}{L}\left(\sum_{t=t_0}^{t_0+L-1} h(u^{(t)}) + \sum_{t=t_0+L}^{\tau-1}\left(h(u^{(t)}) - h(\bar{u}^{(t-L)})\right)\right), \quad (14)$$

where t_0 is a constant that plays the role of the burn-in period, although it is not a burn-in period in the usual sense, since we do not require the Markov chains to converge. Indeed, the estimator in Eq. 14 requires us to run the coupled Markov chains until they meet each other. Given that we design the joint kernel such that the meeting time τ is finite, this implies that we obtain the unbiased estimator in finite time.

We can also think of this unbiased coupling procedure as providing an empirical approximation[2] $\hat{\pi}(\cdot)$ of $\pi(\cdot)$,

$$\hat{\pi}(u) = \frac{1}{L}\left(\sum_{t=t_0}^{t_0+L-1} \delta_{u^{(t)}}(u) + \sum_{t=t_0+L}^{\tau-1}\left(\delta_{u^{(t)}}(u) - \delta_{\bar{u}^{(t-L)}}(u)\right)\right). \quad (15)$$

We next provide sufficient conditions that ensure that the estimator in Eq. 14 can be computed in expected finite time and has finite variance. These conditions are similar as for the original estimator of [Jacob et al., 2020b, Middleton et al., 2020] but the proof is slightly different.

Proposition 3 (The unbiased estimator can be computed in finite time and has finite variance)**.** *Assume the following conditions hold:*

a. (Convergence of the Markov chain.) Each of the two chains marginally starts from a distribution π_0, evolves according to a transition kernel $\mathcal{K}$ and is such that $\mathbb{E}\left[h(u^{(t)})\right] \rightarrow \mathbb{E}_{\pi(u)}\left[h(u)\right]$ as $t \rightarrow \infty$.

b. (Finite high-order moment.) There exist $\eta > 0$ and $D < \infty$ such that $\mathbb{E}\left[\left|h(u^{(t)})\right|^{2+\eta}\right] \leq D \quad \forall t \geq 0$.

c. (Distribution of the meeting time.) There exists an almost surely finite meeting time $\tau = \inf\{t \geq L : u^{(t)} = \bar{u}^{(t-L)}\}$ such that $\mathbb{P}(\tau > t) \leq Ct^{-\kappa}$ for some $C < \infty$ and $\kappa > 2(2\eta^{-1} + 1)$, where η appears in Condition (b).

d. (The chains stay together after meeting.) We have $u^{(t)} = \bar{u}^{(t-L)}$ for all $t \geq \tau$.

Then, Eq. 14 is an unbiased estimator of $\mathbb{E}_{\pi(u)}[h(u)]$ that can be computed in finite expected time and has finite variance.

Algorithm 3: C-DISIR kernel for two coupled chains, $\mathcal{K}_{\text{C-DISIR}}((\cdot,\cdot),(\cdot,\cdot) \mid (\xi_{1:K},\ell),(\bar{\xi}_{1:K},\bar{\ell}))$

```
Input: Current state of both chains, $(\xi_{1:K},\ell)$ and
       $(\bar{\xi}_{1:K},\bar{\ell})$, and correlation strength $\beta$
Output: New state of both chains
1 Sample $\ell_{\text{aux}} \sim \text{Cat}(\frac{1}{K},\ldots,\frac{1}{K})$
2 Set $\xi^\star_{\ell_{\text{aux}}} = \xi_\ell$ and $\bar{\xi}^\star_{\ell_{\text{aux}}} = \bar{\xi}_{\bar{\ell}}$
3 Sample $\xi_k^{\text{new}} \sim q(\xi)$ for $k \in \{1,\ldots,K\}\setminus\{\ell_{\text{aux}}\}$
4 Set $\xi^\star_k = \beta\xi^\star_{k-1} + \sqrt{1-\beta^2}\xi_k^{\text{new}}$ and
   $\bar{\xi}^\star_k = \beta\bar{\xi}^\star_{k-1} + \sqrt{1-\beta^2}\xi_k^{\text{new}}$ for $k = \ell_{\text{aux}}+1,\ldots,K$
5 Set $\xi^\star_k = \beta\xi^\star_{k+1} + \sqrt{1-\beta^2}\xi_k^{\text{new}}$ and
   $\bar{\xi}^\star_k = \beta\bar{\xi}^\star_{k+1} + \sqrt{1-\beta^2}\xi_k^{\text{new}}$ for $k = \ell_{\text{aux}}-1,\ldots,1$
6 Set $z_k^\star = g_\phi(\xi_k^\star, x)$ and $\bar{z}_k^\star = g_\phi(\bar{\xi}_k^\star, x)$ for
   $k = 1,\ldots,K$
7 Sample
   $\ell^\star, \bar{\ell}^\star \sim \mathcal{K}_{\text{C-Cat}}(\ell,\bar{\ell} \mid (w_{\theta,\phi}(z_1^\star),\ldots,w_{\theta,\phi}(z_K^\star)),$
   $(w_{\theta,\phi}(\bar{z}_1^\star),\ldots,w_{\theta,\phi}(\bar{z}_K^\star)))$ from the maximal
   coupling kernel (Algorithm 5 in Appendix B)
8 Return $((\xi^\star_{1:K},\ell^\star),(\bar{\xi}^\star_{1:K},\bar{\ell}^\star))$
```

Conditions (a) and (d) can be satisfied by careful design of the joint kernel $\mathcal{K}_\text{C}$. Condition (b) is a mild integrability condition. Condition (c) can be satisfied if the marginal kernel $\mathcal{K}$ is (only) polynomially ergodic and some additional mild irreducibility and aperiodicity conditions on the joint kernel $\mathcal{K}_\text{C}$ hold [Middleton et al., 2020].

4.2 COUPLING DISIR

Here we describe the main algorithm of this paper: an estimator based on two coupled Markov chains, each evolving according to the DISIR transition kernel from Section 3.2. That is, we build a joint kernel $\mathcal{K}_\text{C}$ for unbiased gradient estimation. We denote the joint kernel as $\mathcal{K}_{\text{C-DISIR}}((\cdot,\cdot),(\cdot,\cdot) \mid (\xi_{1:K},\ell),(\bar{\xi}_{1:K},\bar{\ell}))$. It inputs the current state of both Markov chains, $(\xi_{1:K},\ell)$ and $(\bar{\xi}_{1:K},\bar{\ell})$, and returns their new states.

The coupled DISIR kernel (C-DISIR) is given in Algorithm 3. It resembles the DISIR kernel of Algorithm 2, and in fact, as required, it behaves as $\mathcal{K}_{\text{DISIR}}(\cdot,\cdot \mid \xi_{1:K},\ell)$ marginally if we ignore one of the two Markov chains. Thus, Algorithm 3 guarantees that the marginal stationary distribution of each chain is $p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K},\ell \mid x)$ (Eq. 10).

The indicators $(\ell^\star,\bar{\ell}^\star)$ are sampled jointly from a kernel $\mathcal{K}_{\text{C-Cat}}$ (Line 7 of Algorithm 3), which is given in Appendix B. This corresponds to the maximal coupling kernel[3] for two categorical distributions [Lindvall, 2002].

When the correlation strength $\beta = 0$, i.e., when DISIR is equivalent to ISIR, coupling may occur; that is, Algorithm 3 may return the same state for both chains. To see this, note that when $\beta = 0$, Algorithm 3 shares the same values of the noise values generating the importance samples for both

[2] Eq. 15 is a signed measure, i.e., we can have $\mathbb{E}_{\hat{\pi}(u)}[h(u)] < 0$ even for a positive function $h(\cdot)$.

[3] A coupling procedure is maximal if it maximizes the probability that both chains meet.

Algorithm 4: Unbiased estimation with C-ISIR-DISIR

```
Input: The constant $t_0$ and the lag $L$
Output: An unbiased estimator of $\nabla_\theta \mathcal{L}$
1  Initialize $\xi_k \sim q(\xi)$ and $\bar{\xi}_k \sim q(\xi)$ for $k = 1, \dots, K$
2  Initialize $\ell \sim \mathrm{Cat}(\frac{1}{K}, \dots, \frac{1}{K})$ and $\bar{\ell} \sim \mathrm{Cat}(\frac{1}{K}, \dots, \frac{1}{K})$
3  for $t = 1, \dots, L$ do
4      Sample
         $(\xi_{1:K}^{(t)}, \ell^{(t)}) \sim \mathcal{K}_{\text{ISIR-DISIR}}(\cdot, \cdot \mid \xi_{1:K}^{(t-1)}, \ell^{(t-1)})$ (two
         steps of Algorithm 2, with $\beta = 0$ then $\beta > 0$)
5  end
6  Set the iteration $t = L$
7  while $t < t_0 + L - 1$ or the two chains have not met do
8      Sample $((\xi_{1:K}^{(t+1)}, \ell^{(t+1)}), (\bar{\xi}_{1:K}^{(t-L+1)}, \bar{\ell}^{(t-L+1)})) \sim$
         $\mathcal{K}_{\text{C-ISIR-DISIR}}((\cdot, \cdot), (\cdot, \cdot) \mid (\xi_{1:K}^{(t)}, \ell^{(t)}), (\bar{\xi}_{1:K}^{(t-L)}, \bar{\ell}^{(t-L)}))$
         (two steps of Algorithm 3, with $\beta = 0$ then $\beta > 0$)
9      Increase $t \leftarrow t + 1$
10 end
11 Return the estimator from Eq. 14 using the function
     $h(\xi_{1:K}, \ell) := \sum_{k=1}^{K} \widetilde{w}_{\theta,\phi}^{(k)} \nabla_\theta \log p_\theta(x, z_k)$
```

chains, i.e., $\xi_k^\star = \bar{\xi}_k^\star$ for $k \neq \ell_{\text{aux}}$, while the ℓ_{aux}-th importance sample is set to the current state of each chain, i.e., $\xi_{\ell_{\text{aux}}}^\star = \xi_\ell$ and $\bar{\xi}_{\ell_{\text{aux}}}^\star = \bar{\xi}_{\bar{\ell}}$. Thus, if the indicators sampled in Line 7 take the same value (i.e., $\ell^\star = \bar{\ell}^\star$) and this value is different from ℓ_{aux}, then both chains meet. After meeting, for any future iteration of the joint kernel, the states of both chains are guaranteed to be identical to each other. On the contrary, when the correlation strength $\beta \neq 0$, coupling cannot occur. However, for any $\beta \in [0, 1)$, Algorithm 3 guarantees that the chains remain equal to each other once they have previously met.

We use a composed kernel that consists of the consecutive application of two steps of Algorithm 3. The first step has $\beta = 0$; in this step the chains may meet each other. The second step has $\beta > 0$, which favours exploitation. As discussed in Section 3.2, it is valid to combine these two kernels. We denote this composed joint kernel as $\mathcal{K}_{\text{C-ISIR-DISIR}}$.

Unbiased gradient estimation with C-ISIR-DISIR. Algorithm 4 describes the procedure that provides an unbiased estimator of $\nabla_\theta \mathcal{L}$. It samples two Markov chains, $(\xi_{1:K}^{(t)}, \ell^{(t)})$ and $(\bar{\xi}_{1:K}^{(t)}, \bar{\ell}^{(t)})$, by inducing a coupling between the state of the first chain at time t and the state of the second chain at time $t - L$, where $L \geq 1$ is the lag. After both chains meet, it returns the unbiased gradient estimator using Eq. 14 for the function $h(\xi_{1:K}, \ell) := \sum_{k=1}^{K} \widetilde{w}_{\theta,\phi}^{(k)} \nabla_\theta \log p_\theta(x, z_k)$ (applying Proposition 2), where $z_k = g_\phi(\xi_k, x)$ and the normalized importance weights are $\widetilde{w}_{\theta,\phi}^{(k)} \propto w_{\theta,\phi}(z_k)$. Algorithm 4 provides a practical unbiased estimator, as we show next.

Proposition 4. *Let Assumptions 1 to 3 hold and condition (b) of Proposition 3 be satisfied. For any $K \geq 2$, Algorithm 4 returns an unbiased estimator of $\nabla_\theta \mathcal{L}$ of finite variance that can be computed in finite expected time. Additionally, $\mathbb{E}[\tau]$ can be upper bounded by a quantity decreasing with K.*

5 RELATED WORK

Our estimator builds on previous work discussed in the former sections. We now review other related works.

ISIR, as well as other particle MCMC algorithms, has been previously used for smoothing in state-space models [Andrieu et al., 2010] and for (biased) estimation of $\nabla_\theta \mathcal{L}$ [Naesseth et al., 2020]. Coupled variants of these algorithms have also been previously developed for unbiased smoothing [Jacob et al., 2020a, Middleton et al., 2019]. Indeed, Algorithm 3 for $\beta = 0$ has been used by Jacob et al. [2020a] (without reparameterization). However, we found experimentally that the unbiased estimators based on coupled ISIR suffer from high variance for moderately high dimensions, making them impractical for VAEs. The estimators based on coupled DISIR with $\beta \approx 1$ address this issue.

An unbiased estimator based on a coupled Gibbs sampler has also been presented for restricted Boltzmann machines [Qiu et al., 2019], but this method is not applicable for VAEs. An alternative unbiased gradient estimator for VAEs, based on Russian roulette ideas, was developed by Luo et al. [2020]. However, this estimator suffers from high variance (potentially infinite), and requires additional variance reduction methods such as gradient clipping, which defeats the purpose of unbiased gradient estimation. In our experiments, we use RMSProp and no gradient clipping is needed.

Dieng and Paisley [2019] maximize the marginal log-likelihood of the data using an expectation maximization scheme that gives a consistent (but not unbiased) estimator.

Finally, note that the coupling estimators approximate the gradient w.r.t. θ, and are orthogonal to the methods that improve the expressiveness of the encoder $q_\phi(z \mid x)$, such as semi-implicit methods [Yin and Zhou, 2018, Titsias and Ruiz, 2019] or normalizing flows [Rezende and Mohamed, 2015, Kingma et al., 2016, Papamakarios et al., 2017, Tomczak and Welling, 2016, 2017, Dinh et al., 2017], to name a few. These methods could be used together with the coupling estimators to obtain a more flexible proposal distribution, which could improve the mixing of the Markov chains.

6 EXPERIMENTS

In Section 6.1, we study the bias and variance of different estimators in an experiment where we have access to the exact gradient $\nabla_\theta \mathcal{L}$. In Section 6.2, we study the predictive performance of VAEs trained with coupled DISIR and show that models fitted with unbiased estimators outperform those fitted via ELBO or IWAE maximization. We implement all the estimators in JAX [Babuschkin et al., 2020].

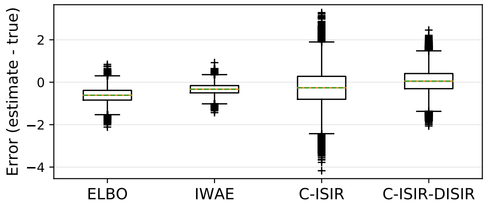


Figure 2: Boxplot representation of the error of different estimators for the gradient w.r.t. one of the intercepts of the PPCA model. The estimators based on variational bounds (ELBO and IWAE) are biased. Among the two unbiased estimators based on couplings, the one based on Algorithm 4 (C-ISIR-DISIR) exhibits lower variance.

6.1 Probabilistic principal component analysis

We first consider probabilistic principal component analysis (PPCA), as for this model we have access to the exact gradient $\nabla_\theta \mathcal{L}$. The model is $p_\theta(x, z) = \mathcal{N}(z; 0, I)\mathcal{N}(x; \theta_0 + \theta_1^\top z, 0.1I)$, where $z \in \mathbb{R}^{100}$. We randomly set the model parameters θ and fit a variational distribution $q_\phi(z \,|\, x)$ by maximizing the IWAE bound w.r.t. ϕ with $K = 100$ importance samples on binarized MNIST [Salakhutdinov and Murray, 2008]. The distribution $q_\phi(z \,|\, x)$ is a fully factorized Gaussian whose parameters depend linearly on x.

We obtain the exact gradient, $\nabla_\theta \mathcal{L} = \sum_{n=1}^{N} \nabla_\theta \log p_\theta(x_n)$, for a batch of $N = 100$ datapoints, and we compare it against four gradient estimators. Two estimators are the gradients of the ELBO and IWAE bounds ($\hat{\nabla}_\theta \mathcal{L}_{\text{ELBO}}$ and $\hat{\nabla}_\theta \mathcal{L}_{\text{IWAE}}$). The third one is the unbiased estimator obtained with coupled ISIR, i.e., a variant of Algorithm 4 where we replace the $\mathcal{K}_{\text{C-ISIR-DISIR}}$ kernel with $\mathcal{K}_{\text{C-ISIR}}$ (which is equivalent to $\mathcal{K}_{\text{C-DISIR}}$ with correlation strength $\beta = 0$). The fourth estimator is based on Algorithm 4. For all the estimators, we use the same (fixed) distribution $q_\phi(z \,|\, x)$. For the coupling estimators, we set $t_0 = 1$ and lag $L = 10$.

We obtain 50,000 samples from each estimator, and compute the (signed) error $\hat{\nabla}_\theta \mathcal{L} - \nabla_\theta \mathcal{L}$ for each sample. We show in Figure 2 the boxplot representation of the error for a randomly chosen component of the gradient w.r.t. the intercept term. (In Appendix D.1, we show a randomly chosen weight term in Figure 5, and the average over components in Figure 6.) As expected, the estimators of the ELBO and IWAE gradients are biased. The boxplots for C-ISIR and C-ISIR-DISIR are consistent with the unbiasedness of the estimators, and the one based on C-ISIR-DISIR has smaller variance. This property is key for fitting more complex models such as VAEs.

6.2 Variational auto-encoder

Now we apply the coupling estimators to fit VAEs and compare the predictive performance to the maximization of the ELBO and IWAE objectives. (We also implemented the method of Luo et al. [2020], but we found it led to unstable optimization despite using gradient clipping.) We provide further details on the experimental setup in Appendix D.2.

Binarized MNIST. We first fit a VAE on the statically binarized MNIST dataset. We use $K = 10$ importance samples and explore the dimensionality $D = \{20, 100, 300\}$ of $z \in \mathbb{R}^D$. We use RMSProp [Tieleman and Hinton, 2012] in the stochastic optimization procedure. For the coupling estimators, we set $t_0 = 1$ step and the lag $L = 10$.

Following Wu et al. [2017], we estimate the predictive log-likelihood using annealed importance sampling (AIS) [Neal, 2001]. Specifically, we use 16 independent AIS chains, with 10,000 intermediate annealing distributions, and a transition operator consisting of one Hamiltonian Monte Carlo (HMC) trajectory with 10 leapfrog steps and adaptive acceptance rate tuned to 0.65. For CIFAR-10, we use 4 AIS chains with 7,500 intermediate distributions and 5 HMC leapfrog steps. As this procedure is computationally intensive, we only evaluate the train log-likelihood on the current data batch, but we evaluate the log-likelihood on the entire test set at the end of the optimization.

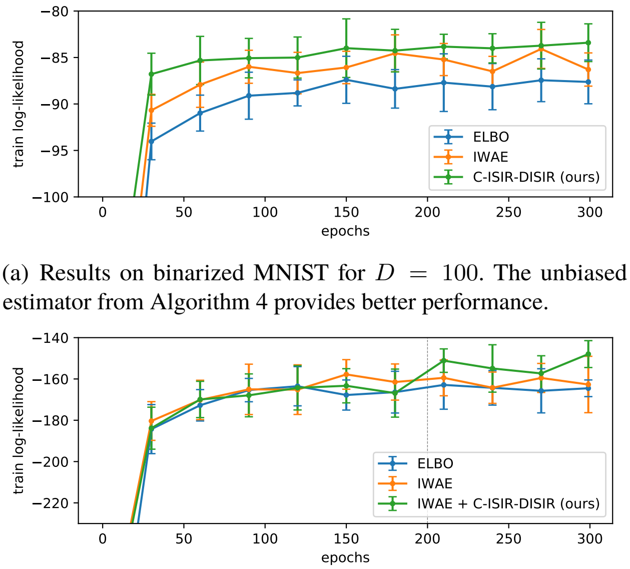
Figure 3(a) shows the evolution of the train log-likelihood; the error bars correspond to the standard deviation of 10 independent runs. (Figure 7 in Appendix D.3 shows similar plots for varying D.) Table 1(a) shows the test log-likelihood after 300 epochs. The VAE models fitted with the unbiased estimator of Algorithm 4 have better predictive performance.

The $\mathcal{K}_{\text{C-ISIR-DISIR}}$ kernel in Algorithm 4 is key for obtaining this improved performance. As a comparison, replacing it with $\mathcal{K}_{\text{C-ISIR}}$ leads to a test log-likelihood value of -90.70 ± 0.08 for $D = 20$, i.e., it is worse than using the standard ELBO (and the gap with the ELBO gets larger for increasing dimensionality D). Moreover, $\mathcal{K}_{\text{C-ISIR-DISIR}}$ alleviates the computational complexity of $\mathcal{K}_{\text{C-ISIR}}$, as measured by the number of MCMC iterations it requires. Figure 4 compares the histograms of the meeting time τ for both kernels; C-ISIR-DISIR requires significantly fewer iterations. (Figure 8 in Appendix D.4 shows that the histograms behave similarly across different values of D.)

The improved performance over IWAE comes at the expense of computational complexity. The cost of Algorithm 4 is roughly 10 times the cost of computing $\hat{\nabla}_\theta \mathcal{L}_{\text{IWAE}}$.

Fashion-MNIST and CIFAR-10. The estimator based on Algorithm 4 leads to improved models but it is also computationally more expensive. We now study the effect of switching to Algorithm 4 after fitting a VAE using the IWAE objective. That is, we first fit the VAE using the IWAE objective for 200 epochs, and then refine the result with the unbiased estimator based on $\mathcal{K}_{\text{C-ISIR-DISIR}}$. We use two datasets, fashion-MNIST [Xiao et al., 2017] and CIFAR-10 [Krizhevsky, 2009], and set $D = 100$.

Figure 3(b) shows the train log-likelihood during optimiza-



(a) Results on binarized MNIST for $D = 100$. The unbiased estimator from Algorithm 4 provides better performance.

(b) Results on fashion-MNIST. After switching from the IWAE to the unbiased estimator at epoch 200, the performance improves.

Figure 3: Train log-likelihood for a VAE.

Table 1: Test log-likelihood for the VAE. The unbiased estimators obtained via the coupled ISIR-DISIR kernel produce models with better predictive performance.

(a) Binarized MNIST.

	dimensionality of z		
	20	100	300
ELBO	-90.05 ± 0.21	-89.96 ± 0.14	-90.63 ± 0.12
IWAE	-88.06 ± 0.08	-88.07 ± 0.06	-89.05 ± 0.08
C-ISIR-DISIR	$\mathbf{-87.29 \pm 0.08}$	$\mathbf{-86.75 \pm 0.10}$	$\mathbf{-88.10 \pm 0.08}$

(b) Fashion-MNIST and CIFAR-10.

	Fashion-MNIST	CIFAR-10
ELBO	-173.36 ± 0.40	-152.06 ± 0.30
IWAE	-170.50 ± 0.30	-149.72 ± 0.39
IWAE + C-ISIR-DISIR	$\mathbf{-168.19 \pm 0.32}$	$\mathbf{-148.40 \pm 0.27}$

tion for fashion-MNIST. After switching from the IWAE objective to the unbiased estimator of Algorithm 4, it improves. Additionally, Table 1(b) shows the test log-likelihood on both fashion-MNIST and CIFAR-10 after 300 epochs. We can conclude that switching to an unbiased gradient estimator boosts the predictive performance of the VAE.

7 DISCUSSION

We have developed a practical algorithm to obtain unbiased estimators of the gradient of the log-likelihood for intractable models, and we have shown empirically that VAEs fitted with unbiased estimators exhibit better predictive performance. Compared to ELBO or IWAE gradients, the main limitation of this approach is its higher computational cost and the fact that the running time is random. While one could obtain more accurate estimators simply by increasing the number of samples of the IWAE bound, this would significantly increase the memory requirement, making it unpractical for datasets like CIFAR-10, and it would also remain biased for any (finite) number of samples.

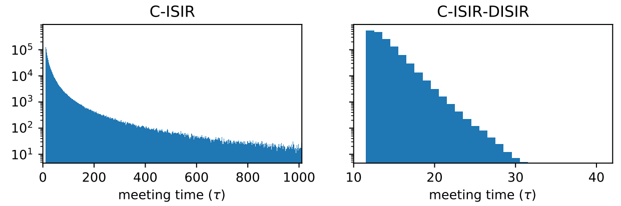


Figure 4: Histogram of the meeting time for a VAE fitted on binarized MNIST with $D = 300$. The histogram corresponding to C-ISIR has significantly heavier tails, which results in higher computational complexity of the overall estimator. Moreover, C-ISIR occasionally (1%) reaches the maximum allowed number of MCMC iterations (hard-coded at around 1,000), which induces a small bias in the estimator. C-ISIR-DISIR does not suffer from this issue.

The topic of coupling estimators is currently an active research field. We expect future work will improve the practical applicability of such estimators using methods like, e.g., control variates [Craiu and Meng, 2020].

Acknowledgements

We thank Andriy Mnih for his insightful comments and useful help. We also thank the anonymous reviewers and area chair for their feedback.

References

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

Christophe Andrieu, Anthony Lee, and Matti Vihola. Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24(2):842–872, 2018.

Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Claudio Fantacci, Jonathan Godwin, Chris Jones, Tom Hennigan, Matteo Hessel, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Lena Martens, Vladimir Mikulik, Tamara Norman, John Quan, George Papamakarios, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Wojciech

Stokowiec, and Fabio Viola. The DeepMind JAX Ecosystem, 2020. URL http://github.com/deepmind.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

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

Radu V. Craiu and Xiao-Li Meng. Double happiness: Enhancing the coupled gains of L-lag coupling via control variates. In *arXiv:2008.12662*, 2020.

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

Adji B. Dieng and John Paisley. Reweighted expectation maximization. In *arXiv:1906.05850*, 2019.

L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

Justin Domke and Daniel R. Sheldon. Importance weighting and variational inference. In *Advances in Neural Information Processing Systems*, 2018.

Samuel J. Gershman and Noah D. Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society*, 2014.

Peter W. Glynn and Chang-Han Rhee. Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014.

Matthew D. Hoffman. Learning deep latent Gaussian models with Markov chain Monte Carlo. In *International Conference on Machine Learning*, 2017.

Pierre E. Jacob, Fredrik Lindsten, and Thomas B. Schön. Smoothing with couplings of conditional particle filters. *Journal of the American Statistical Association*, 115(530): 721–729, 2020a.

Pierre E. Jacob, John O'Leary, and Yves F. Atchadé. Unbiased Markov chain Monte Carlo with couplings (with discussion). *Journal of the Royal Statistical Society Series B*, 82(3):543–600, 2020b.

D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, 2016.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Torgny Lindvall. *Lectures on the Coupling Method*. Courier Corporation, 2002.

Yucen Luo, Alex Beatson, Mohammad Norouzi, Jun Zhu, David Duvenaud, Ryan P. Adams, and Ricky T. Q. Chen. SUMO: Unbiased estimation of log marginal probability for latent variable models. In *International Conference on Learning Representations*, 2020.

Lawrence Middleton, George Deligiannidis, Arnaud Doucet, and Pierre E. Jacob. Unbiased smoothing using particle independent Metropolis–Hastings. In *Artificial Intelligence and Statistics*, 2019.

Lawrence Middleton, George Deligiannidis, Arnaud Doucet, and Pierre E. Jacob. Unbiased Markov chain Monte Carlo for intractable target distributions. *Electronic Journal of Statistics*, 14(2):2842–2891, 2020.

Christian A. Naesseth, Fredrik Lindsten, and David M. Blei. Markovian score climbing: Variational inference with KL(p || q). *Advances in Neural Information Processing Systems*, 2020.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

G. Papamakarios, I. Murray, and T. Pavlakou. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, 2017.

Yixuan Qiu, Lingsong Zhang, and Xiao Wang. Unbiased contrastive divergence algorithm for training energy-based latent variable models. In *International Conference on Learning Representations*, 2019.

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 *International Conference on Machine Learning*, 2019.

D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 2015.

Danilo J. Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks. In *International Conference on Machine Learning*, 2008.

Alexander Y. Shestopaloff and Radford M. Neal. Sampling latent states for high-dimensional non-linear state space models with the embedded HMM method. *Bayesian Analysis*, 13(3):797–822, 2018.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSPROP: Divide the gradient by a running average of its recent magnitude. Coursera: Neural Networks for Machine Learning, 4, 2012.

M. K. Titsias and F. J. R. Ruiz. Unbiased implicit variational inference. In *Artificial Intelligence and Statistics*, 2019.

Michalis K. Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*, 2014.

J. M. Tomczak and M. Welling. Improving variational autoencoders using convex combination linear inverse autoregressive flow. In *arXiv:1706.02326*, 2016.

J. M. Tomczak and M. Welling. Improving variational autoencoders using Householder flow. In *arXiv:1611.09630*, 2017.

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

Paul Vanetti and Arnaud Doucet. Discussion of “Unbiased Markov chain Monte Carlo with couplings” by Jacob et al. *Journal of the Royal Statistical Society Series B*, 82 (3):592–593, 2020.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2): 1–305, January 2008.

Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger Grosse. On the quantitative analysis of decoder-based generative models. In *International Conference on Learning Representations*, 2017.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. In *arXiv:1708.07747*, 2017.

M. Yin and M. Zhou. Semi-implicit variational inference. In *International Conference on Machine Learning*, 2018.