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

Francisco J. R. Ruiz **Michalis K. Titsias** **Taylan Cemgil** **Arnaud Doucet**
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 is challenging since the computation of the 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 maximum likelihood training scheme for VAEs by introducing unbiased gradient estimators of the log-likelihood. We obtain the unbiased 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 (MCMC) coupling procedure on this augmented space. We provide the conditions under which the estimators can be computed in finite time and have finite variance. We demonstrate experimentally that VAEs fitted with unbiased estimators exhibit better predictive performance on three image datasets.

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,

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

Fitting the VAE requires to find the parameters θ that best describe the observations. One (intractable) way to find θ is via maximum likelihood, for which the gradient of Eq. 1 is required. Using Fisher's identity, the gradient of the log-likelihood can be written as an expectation w.r.t. the posterior distribution $p_\theta(z \mid x) \propto p_\theta(x, z)$,

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

The gradient in Eq. 2 can be approximated with samples from the posterior $p_\theta(z \mid x)$; however, the posterior is intractable and we cannot obtain exact samples from it. In practice, we could use a few Markov chain Monte Carlo (MCMC) steps to sample approximately from the posterior, but this would provide a biased estimate of the gradient, and this bias is typically difficult to quantify.

Instead, the parameters of a VAE are usually learned 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). There are multiple ways to construct a bound on the log-marginal likelihood. For example, Burda et al. (2016) form a bound using a set of importance samples, leading to the so-called importance weighted auto-encoder (IWAE). The standard ELBO can be thought of as a particular instance of the IWAE with one importance sample. However, it remains difficult to quantify the difference between the true log-likelihood and the corresponding bound.

In this paper, we develop unbiased gradient estimators of the log-likelihood for VAEs by leveraging recent advances on the literature of unbiased estimation via MCMC (Glynn and Rhee, 2014; Jacob et al., 2020b). In particular, we rely on the coupling approach of Jacob et al. (2020b), who show that it is possible to obtain unbiased estimators of expectations w.r.t. an intractable target distribution by running two coupled MCMC chains. This approach does not require that the MCMC chains converge to their target distribution; thus the unbiased estimator can be computed in finite time.

We develop the unbiased estimators using two key ideas. First, we augment the latent space with a set of importance samples, similarly to the IWAE. Second, we use two coupled MCMC chains on the augmented space. The MCMC procedure can be thought of as a coupled version of the iterated sampling importance resampling (ISIR) algorithm of Andrieu et al. (2010). Moreover, to improve efficiency in high-dimensional settings, we develop an extension of ISIR, called dependent iterated sampling importance resampling

Preprint.

(DISIR), that combines the use of dependent importance samples (similarly to Shestopaloff et al., 2018) and reparameterization ideas with auto-regressive Gaussian processes (similarly to Deligiannidis et al., 2018). Under some regularity conditions, the unbiased gradient estimates can be computed in finite expected time and have finite variance.

The resulting approach is widely applicable: the estimator can be used wherever the IWAE bound is applicable; this includes VAEs but also other applications. We demonstrate experimentally that VAEs fitted with unbiased gradient estimates exhibit better predictive performance on three datasets (binarized MNIST, fashion-MNIST, and CIFAR-10).

2 BACKGROUND

In this section, we review the ELBO and the IWAE, and we show that the IWAE bound can be seen as a standard ELBO on an augmented model.

2.1 The Evidence Lower Bound and VAEs

The ELBO is a lower bound on the marginal log-likelihood used for variational inference (see, e.g., Blei et al., 2017). Given a model $p_\theta(x, z)$ of data x and latent variables $z \in \mathcal{Z}$, consider a variational distribution $q_\phi(z \mid x)$, parameterized by ϕ, that satisfies the following assumption.

Assumption 1 (The variational distribution covers the posterior)**.** *For any z such that $p_\theta(x, z) > 0$, we have $q_\phi(z \mid x) > 0$, so that $0 < w_{\theta,\phi}(z) < \infty$, where*

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

Then, the ELBO is the expectation under $q_\phi(z \mid x)$ of the log-ratio from Eq. 3, i.e.,

$$\mathcal{L}_{\text{ELBO}}(\theta, \phi) = \mathbb{E}_{q_\phi(z \mid x)} \left[\log w_{\theta,\phi}(z)\right] \leq \mathcal{L}(\theta). \tag{4}$$

In variational inference, the ELBO is optimized with respect to the parameters ϕ of the variational distribution $q_\phi(z \mid x)$; this corresponds to minimizing the Kullback-Leibler (KL) divergence $\text{KL}(q_\phi(z \mid x) \mid\mid p_\theta(z \mid x))$ from the variational approximation to the posterior. At the same time, the model parameters θ are also fitted by maximizing the ELBO.

For VAEs (Kingma and Welling, 2014), the likelihood $p_\theta(x \mid z)$ is parameterized by a distribution (e.g., Gaussian) whose parameters (e.g., mean and covariance) are given by the output of a neural network that takes the latent variable z as input. The model parameters θ are the parameters of this neural network, which is also known as the decoder network. In addition, the VAE uses amortized inference (Gershman and Goodman, 2014) by setting the variational distribution $q_\phi(z \mid x)$ as a density network that takes an observation x as input. In this case, the variational parameters ϕ correspond to the parameters of the encoder network. The VAE optimizes the ELBO in Eq. 4 w.r.t. both θ and ϕ using stochastic optimization, since unbiased estimates of the gradients can be easily obtained; e.g., $\hat{\nabla}_\theta \mathcal{L}_{\text{ELBO}} = \nabla_\theta \log p_\theta(x, z)$, with $z \sim q_\phi(z \mid x)$. The gradient estimator w.r.t. ϕ is more intricate to derive. In the rest of the paper, we rely on the reparameterization trick (Kingma and Welling, 2014; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014), which requires to make the following assumption on $q_\phi(z \mid x)$.

Assumption 2 (The variational distribution is reparameterizable)**.** *There exists a distribution $q(\xi)$ independent of ϕ and a mapping $g_\phi(\xi, x)$ such that by sampling $\xi \sim q(\xi)$ and setting $z = g_\phi(\xi, x)$, we obtain $z \sim q_\phi(z \mid x)$.*

In this case, we can obtain an unbiased estimator of the gradient $\nabla_\phi \mathcal{L}_{\text{ELBO}}$ using a sample $\xi \sim q(\xi)$,

$$\hat{\nabla}_\phi \mathcal{L}_{\text{ELBO}} = \nabla_z \log w_{\theta,\phi}(z)\big|_{z=g_\phi(\xi,x)} \nabla_\phi g_\phi(\xi, x). \tag{5}$$

2.2 The Importance Weighted Bound

Burda et al. (2016) derive a variational lower bound based on an importance sampling approximation of the log-evidence. They apply that bound to fit VAEs, and therefore their method is known as IWAE. Using $K \geq 1$ importance samples $z_{1:K}$, the IWAE bound is given by

$$\mathcal{L}_{\text{IWAE}}(\theta, \phi) = \mathbb{E}_{q_\phi(z_{1:K} \mid x)} \left[\log\left(\frac{1}{K}\sum_{k=1}^{K} w_{\theta,\phi}(z_k)\right)\right], \tag{6}$$

where $q_\phi(z_{1:K} \mid x) = \prod_{k=1}^{K} q_\phi(z_k \mid x)$. For $K = 1$, Eq. 6 recovers Eq. 4. The IWAE bound monotonically increases with K, converging towards $\mathcal{L}(\theta)$ as $K \to \infty$.

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

$$\hat{p}_\theta(z \mid x) = \sum_{k=1}^{K} \widetilde{w}^{(k)}_{\theta,\phi} \delta_{z_k}(z), \tag{7}$$

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

The IWAE bound as a standard ELBO. The importance weighted bound in Eq. 6 can be interpreted as a regular ELBO on an augmented latent space (Cremer et al., 2017; Domke and Sheldon, 2018). 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 standard ELBO coincides with Eq. 6.

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 importance sample z_k is distributed according to $q_\phi(z \mid x)$, except the ℓ-th importance sample,

which follows the model $p_\theta(x, z)$. This generative process can be expressed as a joint distribution,

$$p_{\theta,\phi}(x, z_{1:K}, \ell) = \frac{1}{K}\, p_\theta(x, z_\ell) \prod_{k=1, k\neq\ell}^{K} q_\phi(z_k \,|\, x). \quad (8)$$

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,

$$q_{\theta,\phi}(z_{1:K}, \ell \,|\, x) = \text{Cat}\left(\ell \,|\, \widetilde{w}_{\theta,\phi}^{(1)}, \cdots, \widetilde{w}_{\theta,\phi}^{(K)}\right) \prod_{k=1}^{K} q_\phi(z_k \,|\, x). \quad (9)$$

By substituting Eqs. 8 and 9 into Eq. 4, we obtain the ELBO for this augmented model, and it turns out it recovers the IWAE bound in Eq. 6. Thus, maximizing Eq. 6 w.r.t. ϕ corresponds to minimizing $\text{KL}(q_{\theta,\phi}(z_{1:K}, \ell \,|\, x) \,||\, p_{\theta,\phi}(z_{1:K}, \ell \,|\, x))$.

We use the augmented distribution in Sections 3 and 4 to derive MCMC algorithms that target $p_{\theta,\phi}(z_{1:K}, \ell \,|\, x)$, for which we use Eq. 9 as a proposal.

Towards unbiased gradient estimation. The gradient w.r.t. θ of the IWAE bound in Eq. 6 can be interpreted as a (biased) approximation to the gradient of the marginal log-likelihood $\nabla_\theta \mathcal{L}$ from Eq. 2. To see this, we first note that the gradient $\nabla_\theta \mathcal{L}_{\text{IWAE}}$ can be unbiasedly approximated using the approximate posterior from Eq. 7. Specifically,

$$\hat{\nabla}_\theta \mathcal{L}_{\text{IWAE}} = \mathbb{E}_{\hat{p}_\theta(z \,|\, x)} \left[\nabla_\theta \log p_\theta(x, z)\right]. \quad (10)$$

Eq. 10 emphasizes that $\nabla_\theta \mathcal{L}_{\text{IWAE}}$ can be seen as an approximation to $\nabla_\theta \mathcal{L}$ (see Eq. 2), where we replace $p_\theta(z \,|\, x)$ with $\hat{p}_\theta(z \,|\, x)$. In fact, as the number of importance samples $K \to \infty$, Eq. 10 converges to $\nabla_\theta \mathcal{L}$ under mild conditions.

However, for finite K, the expected value under $q_\phi(z_{1:K} \,|\, x)$ of Eq. 10 is a biased estimate of $\nabla_\theta \mathcal{L}$, as the equality

$$\mathbb{E}_{q_\phi(z_{1:K} \,|\, x)} \left[\mathbb{E}_{\hat{p}_\theta(z \,|\, x)} \left[h(z)\right]\right] = \int h(z) p_\theta(z \,|\, x) \mathrm{d}z \quad (11)$$

does *not* hold for all functions $h(\cdot)$.

If we had an alternative approximation $\hat{p}_\theta(z \,|\, x)$ of $p_\theta(z \,|\, x)$ satisfying Eq. 11 then, applied to the function $h(z) = \nabla_\theta \log p_\theta(x, z)$, $\int h(z) \hat{p}_\theta(z \,|\, x) \mathrm{d}z$ would be an unbiased estimate of $\nabla_\theta \mathcal{L}$.

One such example is the empirical measure of exact samples from $p_\theta(z \,|\, x)$. Although this approximation of the posterior would satisfy Eq. 11, it is typically impossible to obtain exact samples from $p_\theta(z \,|\, x)$. We can sample from the posterior by running an MCMC kernel, but in practice, a finite number of MCMC steps provides biased estimates when the chain is initialized out of equilibrium.

Fortunately, we can obtain an approximation $\hat{p}_\theta(z \,|\, x)$ that satisfies Eq. 11 by using two coupled MCMC chains (Jacob et al., 2020b). This method requires to run the two chains for a finite (albeit random) number of steps.

In Section 3 we discuss an importance sampling-based MCMC algorithm on the augmented space (Eq. 8), and we develop an improved MCMC algorithm that addresses a failure mode of this method. In Section 4 we describe the coupled version of these algorithms and provide the algorithm to obtain an unbiased gradient estimator.

3 IMPORTANCE SAMPLING-BASED MCMC SCHEMES

In Section 3.1, we review an MCMC scheme know as ISIR (Andrieu et al., 2010). In Section 3.2, we develop an extension called DISIR, which prevents a failure mode of ISIR that occurs when the latent space $\mathcal{Z}$ is moderately high dimensional. In Section 3.3, we prove that DISIR can be used to estimate expectations under the posterior $p_\theta(z \,|\, x)$, and in Section 3.4 we discuss the choice of the proposal distribution for the MCMC algorithms.

3.1 Iterated Sampling Importance Resampling (ISIR)

ISIR (Andrieu et al., 2010) is an MCMC algorithm that targets the posterior $p_\theta(z \,|\, x)$, but it can also be interpreted as an algorithm that targets the augmented posterior implied by Eq. 8. Here we take this latter view, keeping in mind that if $(z_{1:K}, \ell)$ is a sample from the augmented posterior $p_{\theta,\phi}(z_{1:K}, \ell \,|\, x)$, then z_ℓ is a sample from $p_\theta(z \,|\, x)$.

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 the procedure described in Algorithm 1. ISIR requires a proposal distribution; we use $q_\phi(z \,|\, x)$. It also requires to compute the importance weights $w_{\theta,\phi}(\cdot)$ defined in Eq. 3.

We next review some properties of the ISIR kernel; specifically, that it is invariant w.r.t. the augmented posterior and that there is an upper bound on the total variation distance; see Andrieu et al. (2010, Theorem 4) and Andrieu et al. (2018, Theorem 1). One of these results holds under Assumption 3 below, which in turn holds when the proposal distribution is at least as heavy-tailed as the target.

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

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

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

Proposition 1 (ISIR is invariant w.r.t. the posterior)**.** *Under Assumption 1, for any $K \geq 2$, the ISIR transition kernel $\mathcal{K}_{ISIR}$ is invariant w.r.t. $p_{\theta,\phi}(z_{1:K}, \ell \mid x)$ and the corresponding Markov chain is ergodic. Additionally, if Assumption 3 is satisfied, then for any initial value $(z^{(0)}_{1:K}, \ell^{(0)})$, the total variation distance w.r.t. the target is upper bounded by*

$$\left|\left|\mathcal{K}^T_{\text{ISIR}}(\cdot,\cdot \mid z^{(0)}_{1:K}, \ell^{(0)}) - p_{\theta,\phi}(\cdot,\cdot \mid x)\right|\right|_{TV} \leq \rho^T_K, \quad (12)$$

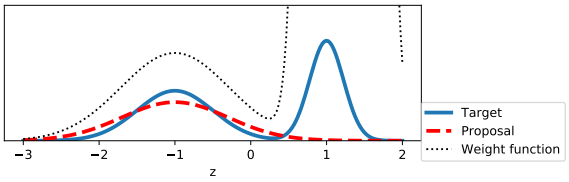
where $\rho_K := 1 - \frac{K-1}{2w^{max}_{\theta,\phi}/p_\theta(x)+K-2} < 1$, and the notation $\mathcal{K}^T_{\text{ISIR}}(\cdot,\cdot \mid z^{(0)}_{1:K}, \ell^{(0)})$ indicates the distribution of the state of the chain after T steps of the kernel initialized at $(z^{(0)}_{1:K}, \ell^{(0)})$.

Proof. See Appendix A.1. □

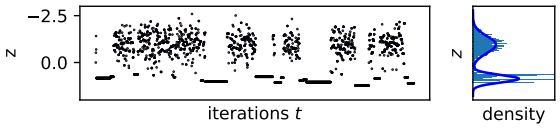
3.2 Dependent Iterated Sampling Importance Resampling (DISIR)

For moderately high-dimensional z, the ISIR algorithm 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; see Figure 1 for an illustrative example in which this effect becomes apparent. 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 (similarly to Shestopaloff et al., 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.

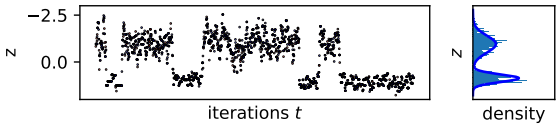
To make samples dependent, we use the reparameterization trick (see Assumption 2) and introduce dependencies among the auxiliary variables $\xi_{1:K}$; this induces dependencies among the samples in the original latent space $z_{1:K}$. That is, rather than sampling $K-1$ importance samples $(\xi^\star_{1:\ell_{\text{aux}}-1}, \xi^\star_{\ell_{\text{aux}}+1:K})$ independently of $\xi^\star_{\ell_{\text{aux}}}$ (as implied by 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 A.2 for details). We describe our choice for the kernel $p_\beta(\xi^\star \mid \xi)$ below. To simplify the presentation here, we make an additional assumption, but we emphasize that this methodology is more generally applicable.



(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.

Assumption 4 (The variational distribution can be reparameterized in terms of a Gaussian)**.** *The auxiliary distribution $q(\xi) = \mathcal{N}(\xi; 0, I)$ is a standard multivariate normal.*

Under Assumption 4, we select a simple autoregressive normal kernel for $p_\beta(\xi^\star \mid \xi)$, i.e., we set

$$p_\beta(\xi^\star \mid \xi) = \mathcal{N}(\xi^\star; \beta\xi, (1-\beta^2)I). \quad (13)$$

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. 13, we develop DISIR, which is described in Algorithm 2. The main difference with respect to ISIR is that 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 $\beta \in (-1, 1)$ as the *correlation strength*. When $\beta = 0$, there is no correlation between samples because the proposed samples are $\xi^\star_k = \xi^{\text{new}}_k \overset{\text{iid}}{\sim} q(\xi)$, and they do not depend on the current state ξ_ℓ. This approach is simply a reparameterized version of ISIR, and it favours exploration of new samples. On the other extreme, when β approaches 1, all the proposed values $\xi^\star_k$ become

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

Input: Current state of the chain $(\xi_{1:K}, \ell)$ and correlation strength β
Output: New state of the chain

Sample $\ell_{\text{aux}} \sim \text{Cat}(\frac{1}{K}, \dots, \frac{1}{K})$
Set $\xi^\star_{\ell_{\text{aux}}} = \xi_\ell$
Sample $\xi^{\text{new}}_k \sim q(\xi)$ for $k \in \{1, \dots, K\} \backslash \{\ell_{\text{aux}}\}$
Set $\xi^\star_k = \beta \xi^\star_{k-1} + \sqrt{1-\beta^2}\xi^{\text{new}}_k$ for $k = \ell_{\text{aux}} + 1, \dots, K$
Set $\xi^\star_k = \beta \xi^\star_{k+1} + \sqrt{1-\beta^2}\xi^{\text{new}}_k$ for $k = \ell_{\text{aux}} - 1, \dots, 1$
Set $z^\star_k = g_\phi(\xi^\star_k, x)$ for $k = 1, \dots, K$
Sample $\ell^\star \sim \text{Cat}(p_1, \dots, p_K)$ with $p_k \propto w_{\theta,\phi}(z^\star_k)$
Return $(\xi^\star_{1:K}, \ell^\star)$

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 2 (Invariant distribution of DISIR)**.** *Let Assumptions 1, 2 and 4 hold. For any $K \geq 2$ and any $\beta \in (-1, 1)$, the DISIR transition kernel $\mathcal{K}_{DISIR}$ admits*

$$p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K}, \ell \,|\, x) = \frac{1}{K} \frac{w_{\theta,\phi}(g_\phi(\xi_\ell, x)) q(\xi_\ell)}{p_\theta(x)} \times \prod_{k=1}^{\ell-1} p_\beta(\xi_k \,|\, \xi_{k+1}) \prod_{k=\ell+1}^{K} p_\beta(\xi_k \,|\, \xi_{k-1}) \tag{14}$$

as invariant distribution and is ergodic.

Proof. See Appendix A.2. □

This target distribution has two desired properties. First, when the correlation strength $\beta \in (-1, 1)$, the ℓ-th sample z_ℓ is distributed according to the posterior $p_\theta(z \,|\, x)$. That is, DISIR defines a Markov chain that targets Eq. 14, and $(z^{(t)}_{\ell^{(t)}})_{t \geq 0}$ is a Markov chain that converges to the posterior $p_\theta(z \,|\, x)$. 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 in Lemma 1.

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

Proof. See Appendix A.3. □

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 $p_\theta(z \,|\, x)$. We denote this 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 $\beta \leftarrow \beta - 0.01(\text{ESS} - 0.3K)$. We also constrain the resulting $\beta \in [10^{-6}, 1 - 10^{-6}]$ to avoid numerical issues.

3.3 Estimates of Expectations with DISIR

The DISIR samples $(z^{(t)}_{\ell^{(t)}})_{t \geq 0}$, with $z^{(t)}_k = g_\phi(\xi^{(t)}_k, x)$, are distributed asymptotically according to $p_\theta(z \,|\, x)$. Thus, in the limit of infinite samples, we can estimate expectations under the posterior, since $\frac{1}{T+1} \sum_{t=0}^T h(z^{(t)}_{\ell^{(t)}}) \to \mathbb{E}_{p_\theta(z \,|\, x)}[h(z)]$ 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 when estimating an expectation. The proposition below, generalizing Theorem 6 by Andrieu et al. (2010), shows that it is possible to use all the K importance samples for estimating expectations. Specifically, we can estimate an expectation of the form $\mathbb{E}_{p_\theta(z \,|\, x)}[h(z)]$ with a weighted average of the importance samples.

Proposition 3 (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*

$$\mathbb{E}_{p_\theta(z \,|\, x)}[h(z)] = \mathbb{E}_{p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K}, \ell \,|\, x)} \left[\sum_{k=1}^K \widetilde{w}^{(k)}_{\theta,\phi} h(z_k) \right], \tag{15}$$

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

$$\frac{1}{T+1} \sum_{t=0}^T \left[\sum_{k=1}^K \widetilde{w}^{(k,t)}_{\theta,\phi} \nabla_\theta \log p_\theta(x, z^{(t)}_k) \right] \to \nabla_\theta \mathcal{L} \tag{16}$$

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

Proof. See Appendix A.4. □

3.4 Choice of the Proposal

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

VAEs use a variational distribution $q_\phi(z \mid x)$ to approximate the posterior; this distribution is defined through an encoder network. Like for the IWAE, in our case $q_\phi(z \mid x)$ is a proposal distribution rather than a variational approximation. We fit ϕ via stochastic optimization of the IWAE bound in Eq. 6. 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).

4 UNBIASED GRADIENT ESTIMATION WITH MCMC COUPLINGS

Section 4.1 reviews how to obtain an unbiased gradient estimator using two coupled Markov chains. Then, Section 4.2 describes the main algorithm of this paper: a method to couple two chains, where each one is marginally evolving according to the DISIR transition kernel from Section 3.2.

4.1 Unbiased Estimation with Couplings

In this section, 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

$$H := \mathbb{E}_{\pi(u)}\left[h(u)\right], \tag{17}$$

for some distribution $\pi(u)$ and function $h(u)$. As discussed in Section 2.2, 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 $\mathcal{K}_\mathrm{C}$. More formally, let $\mathcal{K}(\cdot \mid u)$ be the marginal transition kernel of each Markov chain, and let $\mathcal{K}_\mathrm{C}(\cdot, \cdot \mid u, \bar{u})$ be a joint kernel that takes the state of both chains (denoted u and $\bar{u}$) and produces the new state of both chains. The joint kernel is such that $\mathcal{K}_\mathrm{C}(A, \mathcal{U} \mid u, \bar{u}) = \mathcal{K}(A \mid u)$ and $\mathcal{K}_\mathrm{C}(\mathcal{U}, A \mid u, \bar{u}) = \mathcal{K}(A \mid \bar{u})$ for any measurable set A.

We next review the unbiased estimator of Vanetti and Doucet (2020), a variation of the estimator of Jacob et al. (2020b) that exhibits smaller variance in practice and that applies a construction also used by Biswas et al. (2019). 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., $(u^{(t-1)}, \bar{u}^{(t-L-1)})$. 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., $u^{(t)} \sim \mathcal{K}(u \mid u^{(t-1)})$ 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 $(u^{(t)}, \bar{u}^{(t-L)}) \sim \mathcal{K}_\mathrm{C}(u, \bar{u} \mid u^{(t-1)}, \bar{u}^{(t-L-1)})$.

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 chains meet, $\tau = \inf\{t \geq L : u^{(t)} = \bar{u}^{(t-L)}\}$ (it could be infinite, but in practice 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. 17 by Vanetti and Doucet (2020) is

$$\hat{H} = \frac{1}{L}\left(\sum_{t=t_0}^{t_0+L-1} h(u^{(t)}) + \sum_{t=t_0+L}^{\tau-1}\left(h(u^{(t)}) - h(\bar{u}^{(t-L)})\right)\right), \tag{18}$$

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. 18 requires 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 $\hat{\pi}(\cdot)$ of $\pi(\cdot)$,

$$\hat{\pi}(u) = \frac{1}{L}\left(\sum_{t=t_0}^{t_0+L-1} \delta_{u^{(t)}}(u) + \sum_{t=t_0+L}^{\tau-1}\left(\delta_{u^{(t)}}(u) - \delta_{\bar{u}^{(t-L)}}(u)\right)\right). \tag{19}$$

Eq. 19 is a signed measure, i.e., we can have $\mathbb{E}_{\hat{\pi}(u)}[h(u)] < 0$ even for a positive function $h(\cdot)$.

We next provide sufficient conditions that ensure that the estimator in Eq. 18 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), which were weakened by Middleton et al. (2020).

Proposition 4 (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 $\mathbb{E}\left[h(u^{(t)})\right] \to \mathbb{E}_{\pi(u)}\left[h(u)\right]$ as $t \to \infty$.

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

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 constants* $C < \infty$ *and* $\kappa > 2(2\eta^{-1} + 1)$*, where* η *appears in the previous condition.*

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

Then, Eq. 18 is an unbiased estimator of $\mathbb{E}_{\pi(u)}[h(u)]$ *that can be computed in finite expected time and has finite variance.*

Proof. See Appendix A.5. □

Conditions (a) and (d) can be satisfied by careful design of the joint kernel $\mathcal{K}_{\mathrm{C}}$. 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 $\mathcal{K}_{\mathrm{C}}$ hold (Middleton et al., 2020).

4.2 Coupling DISIR

Here, we build a joint kernel $\mathcal{K}_{\mathrm{C}}$ for unbiased gradient estimation. The joint kernel is based on the DISIR algorithm (Algorithm 2) developed in Section 3.2. We denote the joint kernel as $\mathcal{K}_{\text{C-DISIR}}((\cdot,\cdot),(\cdot,\cdot)\,|\,(\xi_{1:K},\ell),(\bar{\xi}_{1:K},\bar{\ell}))$. 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 $\mathcal{K}_{\text{DISIR}}(\cdot,\cdot\,|\,\xi_{1:K},\ell)$ marginally if we ignore one of the two Markov chains. Thus, Algorithm 3 guarantees that the marginal stationary distribution of each chain is $p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K},\ell\,|\,x)$ (Eq. 14).

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 Algorithm 4. This corresponds to the maximal coupling kernel[1] 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 chains, i.e., $\xi^\star_k = \bar{\xi}^\star_k$ for $k \neq \ell_{\text{aux}}$, while the ℓ_{aux}-th importance sample is set to the current state of each chain, i.e., $\xi^\star_{\ell_{\text{aux}}} = \xi_\ell$ and $\bar{\xi}^\star_{\ell_{\text{aux}}} = \bar{\xi}_{\bar{\ell}}$. 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 $\beta \in (-1,1)$, Algorithm 3 guarantees that the chains remain equal to each other once they have previously met.

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

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

Input: Current state of both chains, $(\xi_{1:K},\ell)$ and $(\bar{\xi}_{1:K},\bar{\ell})$, and correlation strength β
Output: New state of both chains

1 Sample $\ell_{\text{aux}} \sim \text{Cat}(\frac{1}{K},\dots,\frac{1}{K})$
2 Set $\xi^\star_{\ell_{\text{aux}}} = \xi_\ell$ and $\bar{\xi}^\star_{\ell_{\text{aux}}} = \bar{\xi}_{\bar{\ell}}$
3 Sample $\xi^{\text{new}}_k \sim q(\xi)$ for $k \in \{1,\dots,K\}\backslash\{\ell_{\text{aux}}\}$
4 Set $\xi^\star_k = \beta\xi^\star_{k-1} + \sqrt{1-\beta^2}\xi^{\text{new}}_k$ and $\bar{\xi}^\star_k = \beta\bar{\xi}^\star_{k-1} + \sqrt{1-\beta^2}\xi^{\text{new}}_k$ for $k = \ell_{\text{aux}}+1,\dots,K$
5 Set $\xi^\star_k = \beta\xi^\star_{k+1} + \sqrt{1-\beta^2}\xi^{\text{new}}_k$ and $\bar{\xi}^\star_k = \beta\bar{\xi}^\star_{k+1} + \sqrt{1-\beta^2}\xi^{\text{new}}_k$ for $k = \ell_{\text{aux}}-1,\dots,1$
6 Set $z^\star_k = g_\phi(\xi^\star_k, x)$ and $\bar{z}^\star_k = g_\phi(\bar{\xi}^\star_k, x)$ for $k = 1,\dots,K$
7 Sample $\ell^\star,\bar{\ell}^\star \sim \mathcal{K}_{\text{C-Cat}}(\ell,\bar{\ell}\,|\,(w_{\theta,\phi}(z^\star_1),\dots,w_{\theta,\phi}(z^\star_K)),(w_{\theta,\phi}(\bar{z}^\star_1),\dots,w_{\theta,\phi}(\bar{z}^\star_K)))$ from the maximal coupling kernel (Algorithm 4)
8 Return $((\xi^\star_{1:K},\ell^\star),(\bar{\xi}^\star_{1:K},\bar{\ell}^\star))$

In practice, 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 $\beta > 0$; this step favours exploitation. As discussed in Section 3.2, it is valid to combine these two kernels as both can be interpreted as Markov kernels targeting the posterior $p_\theta(z\,|\,x)$. We denote this composed joint kernel as $\mathcal{K}_{\text{C-ISIR-DISIR}}$.

Unbiased gradient estimation with C-ISIR-DISIR. Algorithm 5 describes the procedure that provides an unbiased estimator of the gradient of the marginal log-likelihood, $\nabla_\theta\mathcal{L}$. It samples two Markov chains, $(\xi^{(t)}_{1:K},\ell^{(t)})$ and $(\bar{\xi}^{(t)}_{1:K},\bar{\ell}^{(t)})$, 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. 18 for the function $h(\xi_{1:K},\ell) := \sum_{k=1}^K \widetilde{w}^{(k)}_{\theta,\phi}\nabla_\theta \log p_\theta(x,z_k)$ (applying Proposition 3), where $z_k = g_\phi(\xi_k,x)$ and the normalized importance weights are $\widetilde{w}^{(k)}_{\theta,\phi} \propto w_{\theta,\phi}(z_k)$.

Under some assumptions, the joint kernel guarantees that we obtain an unbiased estimator of finite variance in expected finite time, as we show next.

Proposition 5. *Let Assumptions 1 to 4 hold and condition (b) of Proposition 4 be satisfied. For any* $K \geq 2$*, Algorithm 5 returns an unbiased estimator of* $\nabla_\theta\mathcal{L}$ *of finite variance that can be computed in finite expected time.*[2]

[2] The proof of this proposition also provides a lower bound on the probability of meeting at each iteration that increases with K.

Algorithm 4: Maximal coupling kernel for categoricals, $\mathcal{K}_{\text{C-Cat}}(\cdot, \cdot \,|\, (w_1, \dots, w_K), (v_1, \dots, v_K))$

Input: Two unnormalized probability vectors $(w_1, \dots, w_K)$ and $(v_1, \dots, v_K)$
Output: A sample $\ell, \bar{\ell}$ from the maximal coupling kernel

Normalize the input vectors, obtaining $\widetilde{w}_k \propto w_k$ and $\widetilde{v}_k \propto v_k$ for $k = 1, \dots, K$
Compute the total variation $\gamma = \frac{1}{2} \sum_{k=1}^{K} |\widetilde{w}_k - \widetilde{v}_k|$
Sample $u \sim \text{Uniform}(0, 1)$
if $u \leq 1 - \gamma$ **then** coupling occurs
  Sample $\ell \sim \text{Cat}(p_1, \dots, p_K)$ with $p_k \propto \min(\widetilde{w}_k, \widetilde{v}_k)$
  Return (ℓ, ℓ)
else coupling does not occur
  Sample $\ell \sim \text{Cat}(p_1, \dots, p_K)$ with $p_k \propto \max(\widetilde{w}_k - \widetilde{v}_k, 0)$
  Sample $\bar{\ell} \sim \text{Cat}(p_1, \dots, p_K)$ with $p_k \propto \max(\widetilde{v}_k - \widetilde{w}_k, 0)$
  Return $(\ell, \bar{\ell})$
end

Proof. See Appendix A.6. □

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. Finally, an alternative unbiased gradient estimator, 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. In our experiments, we use RMSProp and no gradient clipping is needed.

Algorithm 5: Unbiased estimation with C-ISIR-DISIR

Input: The constant t_0 and the lag L
Output: An unbiased estimator of $\nabla_\theta \mathcal{L}$

Initialize $\xi_k \sim q(\xi)$ and $\bar{\xi}_k \sim q(\xi)$ for $k = 1, \dots, K$
Initialize $\ell \sim \text{Cat}(\frac{1}{K}, \dots, \frac{1}{K})$ and $\bar{\ell} \sim \text{Cat}(\frac{1}{K}, \dots, \frac{1}{K})$
for $t = 1, \dots, L$ **do**
  Sample $(\xi_{1:K}^{(t)}, \ell^{(t)}) \sim \mathcal{K}_{\text{ISIR-DISIR}}(\cdot, \cdot \,|\, \xi_{1:K}^{(t-1)}, \ell^{(t-1)})$ (two steps of Algorithm 2)
end
Set the iteration $t = L$
while $t < t_0 + L - 1$ or the two chains have not met **do**
  Sample $((\xi_{1:K}^{(t+1)}, \ell^{(t+1)}), (\bar{\xi}_{1:K}^{(t-L+1)}, \bar{\ell}^{(t-L+1)})) \sim \mathcal{K}_{\text{C-ISIR-DISIR}}((\cdot, \cdot), (\cdot, \cdot) \,|\, (\xi_{1:K}^{(t)}, \ell^{(t)}), (\bar{\xi}_{1:K}^{(t-L)}, \bar{\ell}^{(t-L)}))$ (two steps of Algorithm 3)
  Increase $t \leftarrow t + 1$
end
Return the estimator from Eq. 18 using the function $h(\xi_{1:K}, \ell) := \sum_{k=1}^{K} \widetilde{w}_{\theta,\phi}^{(k)} \nabla_\theta \log p_\theta(x, z_k)$

6 EXPERIMENTS

In Section 6.1, we study the bias and variance of different estimators on an experiment where we have access to the exact gradient $\nabla_\theta \mathcal{L}$. In Section 6.2, we study the predictive performance of different VAEs trained with coupled DISIR, obtaining that models fitted with unbiased estimators outperform those fitted via ELBO or IWAE maximization.

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 $\nabla_\theta \mathcal{L}$. Specifically, 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 an amortized variational distribution $q_\phi(z \,|\, x)$ by maximizing the IWAE bound w.r.t. ϕ with $K = 100$ importance samples on the binarized MNIST dataset (Salakhutdinov and Murray, 2008). The distribution $q_\phi(z \,|\, x)$ is a mean-field Gaussian distribution whose parameters are linear functions of 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 5 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 5. 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

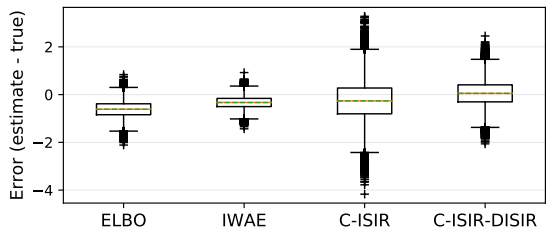


Figure 2: Boxplot representation of the error of different gradient estimators for PPCA. The estimators based on variational bounds (ELBO and IWAE) are biased. Among the two unbiased estimators based on couplings, the one based on Algorithm 5 (C-ISIR-DISIR) exhibits lower variance.

the (signed) error $\hat{\nabla}_\theta \mathcal{L} - \nabla_\theta \mathcal{L}$ for each one. We show in Figure 2 the boxplot representation of the error for a particular component of the gradient w.r.t. the intercept term. 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.

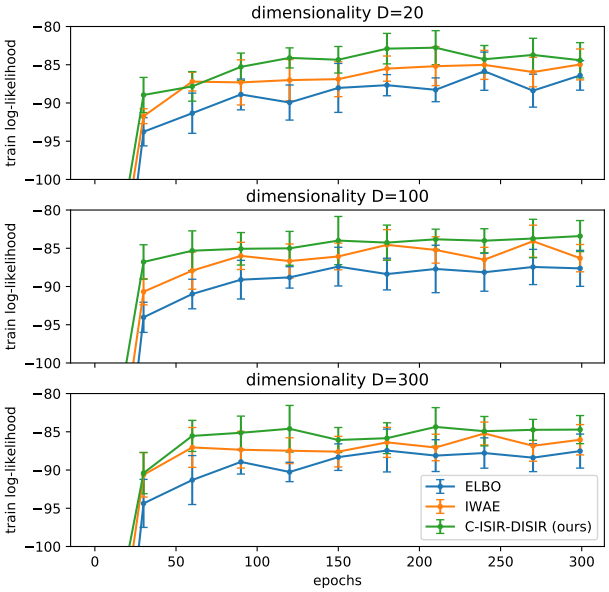
6.2 Variational auto-encoder

Now we apply the coupling estimators to fit VAEs and compare the predictive performance to the maximization of the ELBO and IWAE objectives.

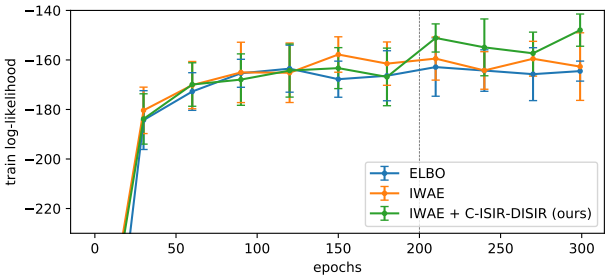
Binarized MNIST. We first consider the binarized MNIST dataset and define a model $p_\theta(x, z) = \mathcal{N}(z; 0, I) p_\theta(x \mid z)$, where $p_\theta(x \mid z)$ is a product of Bernoulli distributions whose parameters are obtained as the output of a neural network with 2 hidden layers of 200 hidden units each and ReLu activations (the third layer is the output layer and has sigmoid activations). We set the distribution $q_\phi(z \mid x)$ as a fully factorized Gaussian, and the encoder network has an analogous architecture (in this case, the output layer implements a linear transformation for the variational means and a softplus transformation for the standard deviations). We set $K = 10$ importance samples and explore the dimensionality $D = \{20, 100, 300\}$ of $z \in \mathbb{R}^D$. We use RMSProp (Tieleman and Hinton, 2012) and set the learning rate to 5×10^{-4} and the batchsize to 100. 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 proposed Hamiltonian Monte Carlo (HMC) trajectory with 10 leapfrog steps and adaptive acceptance rate tuned to 0.65.[3] 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.

[3]For CIFAR-10, the number of AIS chains is 4, the number of intermediate annealing distributions is 7,500, and the number of HMC leapfrog steps is 5.



(a) Results on binarized MNIST. The unbiased estimator from Algorithm 5 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.

Figure 3(a) shows the evolution of the train log-likelihood; the error bars correspond to the standard deviation of 10 independent runs, each with a different random seed. Table 1(a) shows the test log-likelihood after 300 epochs. The VAE models fitted with the unbiased estimator of Algorithm 5 have better predictive performance.

The $\mathcal{K}_{\text{C-ISIR-DISIR}}$ kernel in Algorithm 5 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.

Table 1: Test log-likelihood (higher is better). The unbiased estimators obtained via the coupled DISIR kernel provide VAE 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	$\mathbf{-168.19 \pm 0.32}$	$\mathbf{-148.40 \pm 0.27}$

The improved performance over IWAE comes at the expense of computational complexity. The cost of Algorithm 5 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 5 leads to improved models but it is also computationally more expensive. We now study the effect of switching to Algorithm 5 after fitting a VAE using the IWAE objective. That is, we first fit the VAE using the IWAE bound for 200 epochs, and we 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). We set $D = 100$ and define the likelihood $p_\theta(z \mid x)$ using a mixture of 10 discretized logistic distributions (Salimans et al., 2017). For fashion-MNIST, we use the same encoder and decoder architecture as for binarized MNIST (except for the output layer of the decoder, which implements a linear transformation for the location parameters, a softplus transformation for the scale parameters, and a softmax transformation for the mixture weights). For CIFAR-10, we use convolutional networks instead. The decoder consists of a fully connected layer with hidden size $16 \times 16 \times 1$ and ReLu activations, followed by three convolutional layers with 200, 50, and 30 channels (the filter size is 4×4 with stride of 2) and ReLu activations (except for the last layer). The encoder network has three convolutional layers with 64, 128, and 512 channels, followed by a fully connected hidden layer with output size 128 and by the output layer, which is the same as for fashion-MNIST. The RMSProp learning rate is 10^{-4}, and the batchsize is 100 for fashion-MNIST and 50 for CIFAR-10.

Figure 3(b) shows the train log-likelihood during optimization for fashion-MNIST. After switching from the IWAE objective to the unbiased estimator of Algorithm 5, 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.

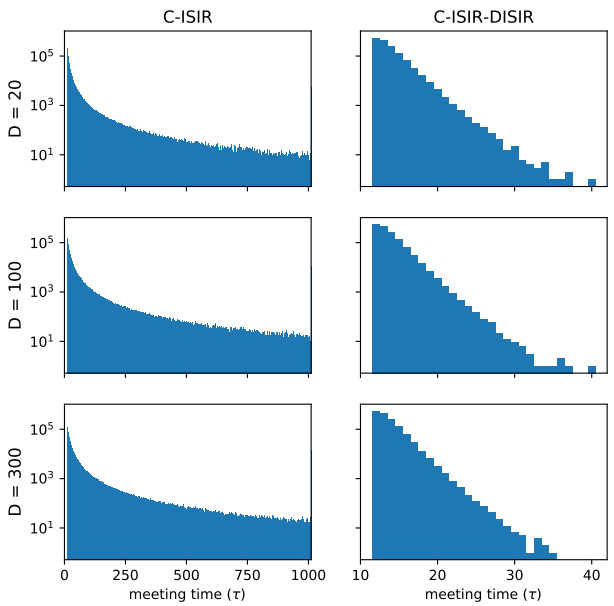


Figure 4: Histograms of the meeting time for a VAE fitted on binarized MNIST. The histogram corresponding to C-ISIR has significantly heavier tails, which results in higher computational complexity of the overall estimator. Moreover, C-ISIR occasionally (1%) reaches the maximum allowed number of MCMC iterations (hard-coded at around $1{,}000$), which induces a small bias in the estimator.

7 DISCUSSION

We have developed a practical algorithm to obtain unbiased estimators of the gradient of the log-likelihood for intractable models. The estimator is based on coupling two Markov chains that evolve according to a joint MCMC kernel that is based on an improvement of ISIR. We have provided the conditions under which the estimator can be computed in finite time and has finite variance. Moreover, we have shown empirically that VAEs fitted with unbiased estimators exhibit better predictive performance.

Compared to ELBO or IWAE gradients, the main practical limitation of the coupling approach is the relatively higher computational cost of the unbiased estimator and the fact that the running time is random, as the meeting time is random. Nevertheless, the topic of coupling estimators is currently an active research field. We expect future work will improve such estimators. For instance, Craiu and Meng (2020) recently use control variates to reduce the variance of the lagged estimator of Eq. 18.

As future work, we can also investigate the use of more sophisticated samplers other than ISIR (which is a special case of a particle MCMC algorithm limited to importance sampling), such as AIS (Neal, 2011) or sequential Monte Carlo (SMC) (Del Moral et al., 2006). For example, some SMC schemes, such as controlled SMC (Heng et al., 2020),

rely on differentiable transition kernels and can perform well without using resampling schemes, so they may be applicable in the context of VAEs.

Bibliography

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

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

Biswas, N., Jacob, P. E., and Vanetti, P. (2019). Estimating convergence of Markov chains with L-lag couplings. In *Advances in Neural Information Processing Systems*, pages 7391–7401.

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

Burda, Y., Grosse, R., and Salakhutdinov, R. (2016). Importance weighted autoencoders. In *International Conference on Learning Representations*.

Craiu, R. V. and Meng, X.-L. (2020). Double happiness: Enhancing the coupled gains of L-lag coupling via control variates. In *arXiv:2008.12662*.

Cremer, C., Morris, Q., and Duvenaud, D. (2017). Reinterpreting importance-weighted autoencoders. In *International Conference on Learning Representations*.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3):411–436.

Deligiannidis, G., Doucet, A., and Pitt, M. K. (2018). The correlated pseudomarginal method. *Journal of the Royal Statistical Society: Series B*, 80(5):839–870.

Domke, J. and Sheldon, D. R. (2018). Importance weighting and variational inference. In *Advances in Neural Information Processing Systems*.

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

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

Heng, J., Bishop, A. N., Deligiannidis, G., and Doucet, A. (2020). Controlled sequential Monte Carlo. *The Annals of Statistics*, 48(5):2904–2929.

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

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

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

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

Lindvall, T. (2002). *Lectures on the Coupling Method*. Courier Corporation.

Luo, Y., Beatson, A., Norouzi, M., Zhu, J., Duvenaud, D., Adams, R. P., and Chen, R. T. (2020). SUMO: Unbiased estimation of log marginal probability for latent variable models. In *International Conference on Learning Representations*.

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2019). Unbiased smoothing using particle independent Metropolis–Hastings. In *Artificial Intelligence and Statistics*.

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2020). Unbiased Markov chain Monte Carlo for intractable target distributions. *Electronic Journal of Statistics*, 14(2):2842–2891.

Naesseth, C. A., Lindsten, F., and Blei, D. (2020). Markovian score climbing: Variational inference with KL(p || q). *arXiv preprint arXiv:2003.10374*.

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

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC.

Qiu, Y., Zhang, L., and Wang, X. (2019). Unbiased contrastive divergence algorithm for training energy-based latent variable models. In *International Conference on Learning Representations*.

Rainforth, T., Kosiorek, A. R., Le, T. A., Maddison, C. J., Igl, M., Wood, F., and Teh, Y. W. (2019). Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*.

Salakhutdinov, R. and Murray, I. (2008). On the quantitative analysis of deep belief networks. In *International Conference on Machine Learning*.

Salimans, T., Karpathy, A., Chen, X., and Kingma, D. P. (2017). PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. In *International Conference on Learning Representations*.

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

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

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728.

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

Tucker, G., Lawson, D., Gu, S., and Maddison, C. J. (2019). Doubly reparameterized gradient estimators for Monte Carlo objectives. In *International Conference on Learning Representations*.

Vanetti, P. and Doucet, A. (2020). 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.

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

Wu, Y., Burda, Y., Salakhutdinov, R., and Grosse, R. (2017). On the quantitative analysis of decoder-based generative models. In *International Conference on Learning Representations*.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. In *arXiv:1708.07747*.

A Appendix

A.1 Proof of Proposition 1

These results have been derived in Andrieu et al. (2010, Theorem 4) and Andrieu et al. (2018, Theorem 1), but we include them here for completeness.

Under Assumption 1, the extended target distribution

$$p_{\theta,\phi}(z_{1:K}, \ell \,|\, x) = \frac{1}{K}\, p_\theta(z_\ell \,|\, x) \prod_{k=1, k\neq\ell}^{K} q_\phi(z_k \,|\, x) \quad (20)$$

is well-defined. The transition kernel of ISIR is defined by Algorithm 1 and given by

$$\mathcal{K}_{\text{ISIR}}(z^\star_{1:K}, \ell^\star \,|\, z_{1:K}, \ell) = \sum_{\ell_{\text{aux}}=1}^{K} \frac{1}{K} \delta_{z_\ell}(z^\star_{\ell_{\text{aux}}}) \times \left(\prod_{k=1, k\neq\ell_{\text{aux}}}^{K} q(z^\star_k \,|\, x) \right) \frac{w_{\theta,\phi}(z^\star_{\ell^\star})}{\sum_{k=1}^K w_{\theta,\phi}(z^\star_k)}. \quad (21)$$

To prove invariance, the marginalization

$$\sum_{\ell=1}^{K} \int p_{\theta,\phi}(z_{1:K}, \ell \,|\, x) \mathcal{K}_{\text{ISIR}}(z^\star_{1:K}, \ell^\star \,|\, z_{1:K}, \ell) \mathrm{d}z_{1:K},$$

should be equal to $p_{\theta,\phi}(z^\star_{1:K}, \ell^\star \,|\, x)$. Indeed,

$$\begin{aligned}
&\sum_{\ell=1}^{K} \int p_{\theta,\phi}(z_{1:K}, \ell \,|\, x) \mathcal{K}_{\text{ISIR}}(z^\star_{1:K}, \ell^\star \,|\, z_{1:K}, \ell) \mathrm{d}z_{1:K} \\
&= \sum_{\ell=1}^{K} \sum_{\ell_{\text{aux}}=1}^{K} \frac{1}{K} \int \frac{p_\theta(z_\ell \,|\, x)}{K} \delta_{z_\ell}(z^\star_{\ell_{\text{aux}}}) \\
&\quad \times \left(\prod_{k=1, k\neq\ell_{\text{aux}}}^{K} q(z^\star_k \,|\, x) \right) \frac{w_{\theta,\phi}(z^\star_{\ell^\star})}{\sum_{k=1}^K w_{\theta,\phi}(z^\star_k)} \mathrm{d}z_\ell \\
&= \sum_{\ell_{\text{aux}}=1}^{K} p_{\theta,\phi}(z^\star_{1:K}, \ell_{\text{aux}} \,|\, x) p_{\theta,\phi}(\ell^\star \,|\, z^\star_{1:K}, x) \\
&= p_{\theta,\phi}(z^\star_{1:K}, \ell^\star \,|\, x).
\end{aligned}$$

Here, we have first integrated out the latent variables $z_{1:K}$ except the ℓ-th one. Then, we have integrated out z_ℓ applying the properties of the Dirac delta function; this makes the resulting expression independent of ℓ and therefore we can easily get rid of the sum over ℓ. Next, we have recognized the term $p_{\theta,\phi}(z^\star_{1:K}, \ell_{\text{aux}} \,|\, x)$ (from Eq. 20), and we have applied that $p_{\theta,\phi}(\ell^\star \,|\, z^\star_{1:K}, x)$ is a categorical distribution with probability proportional to $w_{\theta,\phi}(z^\star_{\ell^\star})$. Finally, we have marginalized out ℓ_{aux}, leading to the final expression.

This transition kernel is ϕ-irreducible and aperiodic under Assumption 1; therefore, the Markov chain is ergodic (Tierney, 1994).

When simulating a Markov chain $(z^{(t)}_{1:K}, \ell^{(t)})_{t\geq 0}$ according to $\mathcal{K}_{\text{ISIR}}$, the $\ell^{(t)}$-th latent variable, i.e., $(z^{(t)} := z^{(t)}_{\ell^{(t)}})_{t\geq 0}$, is also a Markov chain with the transition kernel originally described by Andrieu et al. (2010), which we denote $\mathcal{K}_{\text{ISIR,orig}}$. We can obtain this kernel from Eq. 21 by setting $z_\ell = z$, $z^\star_{\ell^\star} = z^\star$, and marginalizing out the variables $\ell^\star$ and $z^\star_{1:K}$. This gives

$$\begin{aligned}
\mathcal{K}_{\text{ISIR,orig}}(z^\star \,|\, z) &= \frac{1}{K} \sum_{\ell_{\text{aux}}=1}^{K} \sum_{\ell^\star=1}^{K} \int \delta_z(z^\star_{\ell_{\text{aux}}}) \\
&\quad \times \left(\prod_{k\neq\ell_{\text{aux}}} q_\phi(z^\star_k \,|\, x) \right) \frac{w_{\theta,\phi}(z^\star_{\ell^\star})}{\sum_{k=1}^K w_{\theta,\phi}(z^\star_k)} \delta_{z^\star_{\ell^\star}}(z^\star) \mathrm{d}z^\star_{1:K} \\
&= \sum_{\ell^\star=1}^{K} \int \delta_z(z^\star_1) \left(\prod_{k=2}^{K} q_\phi(z^\star_k \,|\, x) \right) \\
&\quad \times \frac{w_{\theta,\phi}(z^\star_{\ell^\star})}{\sum_{k=1}^K w_{\theta,\phi}(z^\star_k)} \delta_{z^\star_{\ell^\star}}(z^\star) \mathrm{d}z^\star_{1:K}, \quad (22)
\end{aligned}$$

where we have used the symmetry of the kernel w.r.t. ℓ_{aux} and have arbitrarily considered the term with $\ell_{\text{aux}} = 1$.

We next prove the bound on the total variation distance. Given that each term is non-negative, we can lower bound $\mathcal{K}_{\text{ISIR,orig}}(z^\star \,|\, z)$ by getting rid of the term corresponding to $\ell^\star = 1$ from the summation. This gives

$$\begin{aligned}
\mathcal{K}_{\text{ISIR,orig}}(z^\star \,|\, z) &\geq \sum_{\ell^\star=2}^{K} \int \delta_z(z^\star_1) \left(\prod_{k=2}^{K} q_\phi(z^\star_k \,|\, x) \right) \\
&\quad \times \frac{w_{\theta,\phi}(z^\star_{\ell^\star})}{\sum_{k=1}^K w_{\theta,\phi}(z^\star_k)} \delta_{z^\star_{\ell^\star}}(z^\star) \mathrm{d}z^\star_{1:K}. \quad (23)
\end{aligned}$$

Using the definition of the importance weights, $w_{\theta,\phi}(z) = p_\theta(x, z)/q_\phi(z \,|\, x) = p_\theta(x) p_\theta(z \,|\, x)/q_\phi(z \,|\, x)$, this yields

$$\begin{aligned}
\mathcal{K}_{\text{ISIR,orig}}(z^\star \,|\, z) &\geq \sum_{\ell^\star=2}^{K} \int \delta_z(z^\star_1) \left(\prod_{k=2, k\neq\ell^\star}^{K} q_\phi(z^\star_k \,|\, x) \right) \\
&\quad \times \frac{p_\theta(x) p_\theta(z^\star_{\ell^\star} \,|\, x)}{\sum_{k=1}^K w_{\theta,\phi}(z^\star_k)} \delta_{z^\star_{\ell^\star}}(z^\star) \mathrm{d}z^\star_{1:K}. \quad (24)
\end{aligned}$$

By Assumption 3, we have $w_{\theta,\phi}(z^\star_1) + w_{\theta,\phi}(z^\star_2) \leq 2w^{\max}_{\theta,\phi}$, and we can further lower bound $\mathcal{K}_{\text{ISIR,orig}}$ as

$$\begin{aligned}
\mathcal{K}_{\text{ISIR,orig}}(z^\star \,|\, z) &\geq \sum_{\ell^\star=2}^{K} \int \delta_z(z^\star_1) \left(\prod_{k=2, k\neq\ell^\star}^{K} q_\phi(z^\star_k \,|\, x) \right) \\
&\quad \times \frac{p_\theta(x) p_\theta(z^\star_{\ell^\star} \,|\, x)}{2w^{\max}_{\theta,\phi} + \sum_{k=3}^K w_{\theta,\phi}(z^\star_k)} \delta_{z^\star_{\ell^\star}}(z^\star) \mathrm{d}z^\star_{1:K}. \quad (25)
\end{aligned}$$

Next, by using the symmetry of the integrand w.r.t. $\ell^\star$, it follows that

$$\mathcal{K}_{\text{ISIR,orig}}(z^\star \,|\, z) \geq \mathbb{E}\left[\frac{(K-1) p_\theta(x) p_\theta(z^\star \,|\, x)}{2w^{\max}_{\theta,\phi} + \sum_{k=3}^K w_{\theta,\phi}(z^\star_k)} \right], \quad (26)$$

where the expectation is w.r.t. $z_k^\star \sim q_\phi(\cdot \mid x)$ for $k = 3, ..., K$. We finally apply Jensen's inequality, $\mathbb{E}_q[f(\cdot)] \geq f(\mathbb{E}_q[\cdot])$ for the convex function $f(x) = 1/x$, obtaining

$$\begin{aligned}\mathcal{K}_{\text{ISIR,orig}}(z^\star \mid z) &\geq \frac{(K-1)p_\theta(x)p_\theta(z^\star \mid x)}{\mathbb{E}\left[2w_{\theta,\phi}^{\max} + \sum_{k=3}^K w_{\theta,\phi}(z_k^\star)\right]} \\ &= \frac{(K-1)p_\theta(x)}{2w_{\theta,\phi}^{\max} + (K-2)p_\theta(x)} p_\theta(z^\star \mid x).\end{aligned} \tag{27}$$

This kernel thus satisfies a minorization condition and thus

$$\left|\left|\mathcal{K}_{\text{ISIR,orig}}^T(\cdot \mid z^{(0)}) - p_\theta(\cdot \mid x)\right|\right|_{\text{TV}} \leq \rho_K^T, \tag{28}$$

where $\rho_K := 1 - \frac{K-1}{2w_{\theta,\phi}^{\max}/p_\theta(x) + K - 2}$.

The bound in Eq. 12 follows directly, since $z^{(t)} := z_{\ell^{(t)}}^{(t)}$ and, under the transition kernel $\mathcal{K}_{\text{ISIR}}$, the remaining variables are sampled from the full conditional distribution of the extended target from Eq. 20, so we have for any $T \geq 0$,

$$\begin{aligned}&\left|\left|\mathcal{K}_{\text{ISIR,orig}}^T(\cdot \mid z^{(0)}) - p_\theta(\cdot \mid x)\right|\right|_{\text{TV}} \\ &= \left|\left|\mathcal{K}_{\text{ISIR}}^T(\cdot, \cdot \mid z_{1:K}^{(0)}, \ell^{(0)}) - p_{\theta,\phi}(\cdot, \cdot \mid x)\right|\right|_{\text{TV}}.\end{aligned} \tag{29}$$

A.2 Proof of Proposition 2

We now prove here that the DISIR kernel admits $p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}, \ell \mid x)$ defined in Eq. 14 as invariant distribution. The transition kernel $\mathcal{K}_{\text{DISIR}}(\cdot, \cdot \mid \xi_{1:K}, \ell)$ is defined through Algorithm 3 and can be written as

$$\begin{aligned}\mathcal{K}_{\text{DISIR}}(\xi_{1:K}^\star, \ell^\star \mid \xi_{1:K}, \ell) &= \sum_{\ell_{\text{aux}}=1}^K \frac{1}{K} \delta_{\xi_\ell}(\xi_{\ell_{\text{aux}}}^\star) \\ &\times \prod_{k=1}^{\ell_{\text{aux}}-1} p_\beta(\xi_k^\star \mid \xi_{k+1}^\star) \prod_{k=\ell_{\text{aux}}+1}^K p_\beta(\xi_k^\star \mid \xi_{k-1}^\star) \\ &\times \frac{w_{\theta,\phi}(z_{\ell^\star}^\star)}{\sum_{k=1}^K w_{\theta,\phi}(z_k^\star)},\end{aligned}$$

for $z_k^\star = g_\phi(\xi_k^\star, x)$.

For the kernel to be invariant, the marginalization

$$\sum_{\ell=1}^K \int p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}, \ell \mid x) \mathcal{K}_{\text{DISIR}}(\xi_{1:K}^\star, \ell^\star \mid \xi_{1:K}, \ell) \mathrm{d}\xi_{1:K}$$

should be equal to $p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}^\star, \ell^\star \mid x)$. We obtain this marginalization below. We first define

$$p_{\theta,\phi}^{\text{DISIR}}(\xi \mid x) := \frac{w_{\theta,\phi}(g_\phi(\xi, x))q(\xi)}{p_\theta(x)}. \tag{30}$$

(Eq. 30 gives the marginal distribution of ξ_ℓ obtained after integrating out the rest of latent variables from Eq. 14.) Similarly to the proof in Appendix A.1, we first integrate out the variables $\xi_{1:K}$ except the ℓ-th one, and then we marginalize out ξ_ℓ taking into account the integration property of the Dirac delta function; this allows us to get rid of the sum over ℓ. Specifically, we have

$$\begin{aligned}&\sum_{\ell=1}^K \int p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}, \ell \mid x) \mathcal{K}_{\text{DISIR}}(\xi_{1:K}^\star, \ell^\star \mid \xi_{1:K}, \ell) \mathrm{d}\xi_{1:K} \\ &= \sum_{\ell=1}^K \sum_{\ell_{\text{aux}}=1}^K \frac{1}{K} \int \frac{p_{\theta,\phi}^{\text{DISIR}}(\xi_\ell \mid x)}{K} \delta_{\xi_\ell}(\xi_{\ell_{\text{aux}}}^\star) \\ &\quad \times \prod_{k=1}^{\ell_{\text{aux}}-1} p_\beta(\xi_k^\star \mid \xi_{k+1}^\star) \prod_{k=\ell_{\text{aux}}+1}^K p_\beta(\xi_k^\star \mid \xi_{k-1}^\star) \\ &\quad \times \frac{w_{\theta,\phi}(z_{\ell^\star}^\star)}{\sum_{k=1}^K w_{\theta,\phi}(z_k^\star)} \mathrm{d}\xi_\ell \\ &= \sum_{\ell_{\text{aux}}=1}^K p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}^\star, \ell_{\text{aux}} \mid x) \frac{w_{\theta,\phi}(z_{\ell^\star}^\star)}{\sum_{k=1}^K w_{\theta,\phi}(z_k^\star)} \\ &= \sum_{\ell_{\text{aux}}=1}^K p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}^\star, \ell_{\text{aux}} \mid x) p_{\theta,\phi}(\ell^\star \mid x, \xi_{1:K}) \\ &= p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}^\star, \ell^\star \mid x).\end{aligned}$$

Here, we have additionally recognized the term $p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}^\star, \ell_{\text{aux}} \mid x)$ (see Eq. 14). For the second-to-last step, we have applied the fact that the conditional distribution of ℓ under Eq. 14, $p_{\theta,\phi}^{\text{DISIR}}(\ell \mid x, \xi_{1:K})$, is a categorical with probability proportional to $w_{\theta,\phi}(g_\phi(\xi_\ell, x))$ (this posterior is analogous to the ISIR case). To establish this result, we note that

$$\begin{aligned}p_{\theta,\phi}^{\text{DISIR}}(\ell \mid x, \xi_{1:K}) &\propto w_{\theta,\phi}(g_\phi(\xi_\ell, x))q(\xi_\ell) \\ &\quad \times \prod_{k=1}^{\ell-1} p_\beta(\xi_k \mid \xi_{k+1}) \prod_{k=\ell+1}^K p_\beta(\xi_k \mid \xi_{k-1}).\end{aligned}$$

Since p_β is reversible with respect to q, it follows that

$$\begin{aligned}&q(\xi_\ell) \prod_{k=1}^{\ell-1} p_\beta(\xi_k \mid \xi_{k+1}) \prod_{k=\ell+1}^K p_\beta(\xi_k \mid \xi_{k-1}) \\ &= q(\xi_1) \prod_{k=2}^K p_\beta(\xi_k \mid \xi_{k-1}),\end{aligned}$$

so this product is independent of ℓ and we have

$$p_{\theta,\phi}^{\text{DISIR}}(\ell \mid x, \xi_{1:K}) \propto w_{\theta,\phi}(g_\phi(\xi_\ell, x)). \tag{31}$$

This establishes the proof of invariance. The DISIR kernel is ergodic for the same reasons as ISIR.

A.3 Proof of Lemma 1

We prove here Lemma 1. From the DISIR invariant distribution in Eq. 14, it follows that the marginal distribution of ξ_ℓ is given by $p_{\theta,\phi}^{\text{DISIR}}(\xi_\ell \mid x)$ (Eq. 30).

We need to show that for $\xi \sim p^{\text{DISIR}}_{\theta,\phi}(\xi \,|\, x)$, then $z := g_\phi(\xi, x) \sim p_\theta(z \,|\, x)$. For any test function $f(\cdot)$,

$$\begin{aligned}&\mathbb{E}_{p^{\text{DISIR}}_{\theta,\phi}(\xi \,|\, x)}\left[f(g_\phi(\xi, x))\right] \\ &= \int f(g_\phi(\xi, x)) \frac{w_{\theta,\phi}(g_\phi(\xi, x))q(\xi)}{p_\theta(x)} \mathrm{d}\xi \\ &= \int f(g_\phi(\xi, x)) \frac{p_\theta(x, g_\phi(\xi, x))}{q_\phi(g_\phi(\xi, x) \,|\, x)} \frac{q(\xi)}{p_\theta(x)} \mathrm{d}\xi.\end{aligned}$$

Under Assumption 2, we know that if $\xi \sim q(\xi)$ then $z = g_\phi(\xi, x) \sim q_\phi(z \,|\, x)$. Thus, by using the change of variables $z = g_\phi(\xi, x)$, we have

$$\begin{aligned}\mathbb{E}_{p^{\text{DISIR}}_{\theta,\phi}(\xi \,|\, x)}\left[f(g_\phi(\xi, x))\right] &= \int f(z) \frac{p_\theta(x, z)}{q_\phi(z \,|\, x)} \frac{q_\phi(z \,|\, x)}{p_\theta(x)} \mathrm{d}z \\ &= \mathbb{E}_{p_\theta(z \,|\, x)}\left[f(z)\right]. \end{aligned} \tag{32}$$

This completes the proof of the first part of Lemma 1.

The second part says that, for $\beta = 0$, the variables $(z_{1:K}, \ell)$ are distributed according to the augmented posterior. In this case, the variables ξ_k for $k \neq \ell$ are independent and identically distributed according to $q(\xi)$. Thus, if $(\xi_{1:K}, \ell) \sim p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K}, \ell \,|\, x)$ then, for $z_k = g_\phi(\xi_k, \ell)$, we have $(z_{1:K}, \ell) \sim p_{\theta,\phi}(z_{1:K}, \ell \,|\, x)$.

A.4 Proof of Proposition 3

We establish here the identity in Eq. 15. From the result in Lemma 1 and the definition of $p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K}, \ell \,|\, x)$ (Eq. 14), it follows that

$$\begin{aligned}&\int h(z) p_\theta(z \,|\, x) \mathrm{d}z \\ &= \sum_{\ell=1}^K \int h(g_\phi(\xi_\ell, x)) p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K}, \ell \,|\, x) \mathrm{d}\xi_{1:K} \\ &= \int \Big[\sum_{\ell=1}^K h(g_\phi(\xi_\ell, x)) p^{\text{DISIR}}_{\theta,\phi}(\ell \,|\, x, \xi_{1:K})\Big] p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K} \,|\, x) \mathrm{d}\xi_{1:K} \\ &= \int \Big[\sum_{\ell=1}^K \widetilde{w}^{(\ell)}_{\theta,\phi} h(g_\phi(\xi_\ell, x))\Big] p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K} \,|\, x) \mathrm{d}\xi_{1:K},\end{aligned}$$

where $\widetilde{w}^{(\ell)}_{\theta,\phi} \propto w_{\theta,\phi}(g_\phi(\xi_\ell, x))$ are the normalized importance weights. The last equality follows from Eq. 31 used in the proof of Appendix A.2. The result thus follows.

A.5 Proof of Proposition 4

The following shows that the conditions established by Middleton et al. (2020) to establish the fact that the estimator of Jacob et al. (2020b) can be computed in finite expected time and admit a finite variance are also applicable to the estimator of Eq. 18. The proof mimics the approach followed by Jacob et al. (2020b, Proposition 3.1) and Middleton et al. (2020, Theorem 1).

Here, we use the notation $\mu(h) := \int h(u)\mu(u)\mathrm{d}u$ for any test function $h(u)$ and probability density $\mu(u)$. Our goal is to estimate $H := \pi(h)$.

Firstly, by condition (c), we have $\mathbb{E}\left[\tau\right] < \infty$, so the estimator $\hat{H} := \hat{\pi}(h)$ from Eq. 18 can be computed in finite expected time.

Now let us denote by L_2 the complete space of random variables with finite second moment. We consider the sequence of random variables $(\hat{\pi}_N(h))_{N \geq k+L}$ defined by

$$\begin{aligned}\hat{\pi}_N(h) &= \frac{1}{L}\left(\sum_{t=k}^{k+L-1} h(u^{(t)}) + \sum_{t=k+L}^{N} h(u^{(t)}) - h(\bar{u}^{(t-L)})\right) \\ &= \frac{1}{L}\sum_{t=k}^{N} \Delta_t,\end{aligned}$$

where $\Delta_t := h(u^{(t)}) - h(\bar{u}^{(t-L)})$ for $t \geq k+L$ and $\Delta_t := h(u^{(t)})$ for $k \leq t < k+L$. We next show that this sequence is a Cauchy sequence in L_2 converging to $\hat{\pi}(h)$.

As $\mathbb{E}\left[\tau\right] < \infty$, we have $\mathbb{P}(\tau < \infty) = 1$ and $u^{(t)} = \bar{u}^{(t-L)}$ for $t \geq \tau$ under condition (d). Thus, it follows that $\hat{\pi}_N(h) \to \hat{\pi}(h)$ almost surely. For positive integers N, N' such that $k + L \leq N < N'$, we have

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{\pi}_N(h) - \hat{\pi}_{N'}(h)\right)^2\right] \\ &= \frac{1}{L^2}\sum_{s=N+1}^{N'}\sum_{t=N+1}^{N'} \mathbb{E}\left[\Delta_s \Delta_t\right] \\ &\leq \frac{1}{L^2}\sum_{s=N+1}^{N'}\sum_{t=N+1}^{N'} \mathbb{E}\left[\Delta_s^2\right]^{1/2} \mathbb{E}\left[\Delta_t^2\right]^{1/2} \\ &= \frac{1}{L^2}\left(\sum_{t=N+1}^{N'} \mathbb{E}\left[\Delta_t^2\right]^{1/2}\right)^2.\end{aligned}$$

Since $\mathbb{E}\left[\Delta_t^2\right] = \mathbb{E}\left[\Delta_t^2 \mathbb{I}_{\tau > t}\right]$, where $\mathbb{I}$ is the indicator function, by Holder's inequality we have

$$\begin{aligned}\mathbb{E}\left[\Delta_t^2\right] &\leq \mathbb{E}\left[\left|\Delta_t\right|^{2+\eta}\right]^{\frac{1}{1+\frac{\eta}{2}}} \mathbb{E}\left[\mathbb{I}_{\tau > t}\right]^{\frac{\eta}{2+\eta}} \\ &\leq D^{\frac{1}{1+\frac{\eta}{2}}} \mathbb{P}(\tau > t)^{\frac{\eta}{2+\eta}},\end{aligned}$$

where $\mathbb{E}\left[\left|\Delta_t\right|^{2+\eta}\right] < D$ for all t as $\mathbb{E}\left[\left|h(u^{(t)})\right|^{2+\eta}\right] < D$ by condition (b). Consequently, we have

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{\pi}_N(h) - \hat{\pi}_{N'}(h)\right)^2\right] \\ &\leq \frac{1}{L^2}\left(\sum_{t=N+1}^{N'} \left(D^{\frac{1}{1+\frac{\eta}{2}}} \mathbb{P}(\tau > t)^{\frac{\eta}{2+\eta}}\right)^{\frac{1}{2}}\right)^2 \\ &= \frac{1}{L^2} D^{\frac{1}{1+\frac{\eta}{2}}} \left(\sum_{t=N+1}^{N'} \mathbb{P}(\tau > t)^{\frac{1}{2}\frac{\eta}{2+\eta}}\right)^2.\end{aligned}$$

Defining $\lambda := \frac{1}{2}\frac{\eta}{2+\eta}$, it follows from condition (c) that

$\mathbb{P}(\tau > t) \leq Ct^{-\kappa}$ for $\kappa > 1/\lambda$, which yields

$$\sum_{t=N+1}^{\infty} \mathbb{P}(\tau > t)^{\lambda} \leq C \sum_{t=N+1}^{\infty} \frac{1}{t^{\lambda\kappa}} < \infty.$$

Thus, we have $\lim_{N\to\infty} \sum_{t=N+1}^{\infty} \mathbb{P}(\tau > t)^{\lambda} = 0$. Hence, we have proved $\hat{\pi}_N(h)$ is a Cauchy sequence in L_2, and has finite first and second moments, so $\hat{\pi}(h)$ has finite variance. As Cauchy sequences are bounded, the dominated convergence theorem shows that

$$\mathbb{E}\left[\hat{\pi}(h)\right] = \mathbb{E}\left[\lim_{N\to\infty} \hat{\pi}_N(h)\right] = \lim_{N\to\infty} \mathbb{E}\left[\hat{\pi}_N(h)\right],$$

and, under condition (a), we have

$$\lim_{N\to\infty} \mathbb{E}\left[\hat{\pi}_N(h)\right] = \lim_{N\to\infty} \frac{1}{L} \sum_{t=N-L+1}^{N} \mathbb{E}\left[h(u^{(t)})\right] = \pi(h).$$

A.6 Proof of Proposition 5

To prove Proposition 5, we need to check that the conditions (a) to (d) of Proposition 4 are satisfied for the coupled ISIR-DISIR kernel from Algorithm 3. Condition (a) is satisfied as the ISIR kernel is ϕ-irreducible and aperiodic (see, e.g., Tierney, 1994). Condition (b) is satisfied by assumption. Condition (d) is also satisfied by design of Algorithm 3—once the chains are coupled, they remain equal to each other forever. We next check that condition (c) is also satisfied.

We recall that the transition kernel we couple is a composition of the ISIR kernel followed by the DISIR kernel. Here we show that, at each iteration, the coupled ISIR kernel couples with a probability that is lower bounded by a quantity strictly positive independent of the current states of the two chains. This ensures that the distribution of the meeting time τ has tails decreasing geometrically fast.

As discussed in Section 4.2, the coupled ISIR kernel from Algorithm 3 couples when (i) both indicators sampled in Line 7 are equal, i.e., $\ell^\star = \bar{\ell}^\star$, and (ii) these indicators are different from ℓ_{aux}. Event (i) is driven by the joint kernel $\mathcal{K}_{\text{C-Cat}}$, whose probability of coupling is $1-\gamma$, where γ is defined in Algorithm 4. The probability of event (ii) is equal to the probability that the sampled indicators are different from ℓ_{aux}, which we assume equal to 1 (without loss of generality).

When we use Algorithm 4 to couple the ISIR chains, we have $w_k = w_{\theta,\phi}(z_k^\star)$ and $v_k = w_{\theta,\phi}(\bar{z}_k^\star)$ (see Line 7 of Algorithm 3). Thus, the unnormalized weights $(w_k)_{k=1,\dots,K}$ and $(v_k)_{k=1,\dots,K}$ before coupling differ at most by a single entry, which we assumed above to be the first entry, i.e., $w_1 \neq v_1$ and $w_k = v_k$ for $k = 2, \dots, K$. The normalized probabilities are thus

$$\widetilde{w}_k = \frac{w_k}{w_1 + S}, \quad \widetilde{v}_k = \frac{v_k}{v_1 + S}, \tag{33}$$

where

$$S = \sum_{k=2}^{K} v_k = \sum_{k=2}^{K} w_k. \tag{34}$$

Therefore, the probability of coupling of ISIR is

$$P_{\text{meet}} \geq \mathbb{E}\left[(1-\gamma)\frac{\sum_{k=2}^{K} \min(\widetilde{w}_k, \widetilde{v}_k)}{\sum_{k=1}^{K} \min(\widetilde{w}_k, \widetilde{v}_k)}\right],$$

where the expectation is w.r.t. to the joint distribution of the two chains at time t. Using the identity $|a-b| = a + b - 2\min(a,b)$, the term $1-\gamma$ can be simplified as

$$1 - \gamma = 1 - \frac{1}{2}\sum_{k=1}^{K} |\widetilde{w}_k - \widetilde{v}_k| = \sum_{k=1}^{K} \min(\widetilde{w}_k, \widetilde{v}_k).$$

Thus, we have

$$P_{\text{meet}} \geq \mathbb{E}\left[\sum_{k=2}^{K} \min(\widetilde{w}_k, \widetilde{v}_k)\right].$$

By Assumption 3, we have $w_1 + S \leq K w_{\theta,\phi}^{\max}$ (and similarly for $v_1 + S$); thus we can further lower bound the probability of coupling,

$$\begin{aligned} P_{\text{meet}} &\geq \mathbb{E}\left[\sum_{k=2}^{K} \min\left(\frac{w_k}{w_1+S}, \frac{v_k}{v_1+S}\right)\right] \\ &\geq \mathbb{E}\left[\frac{1}{K w_{\theta,\phi}^{\max}} \sum_{k=2}^{K} w_k\right] \\ &= \frac{(K-1)p_\theta(x)}{K w_{\theta,\phi}^{\max}}. \end{aligned}$$

We have used above that $S = \sum_{k=2}^{K} w_k$ only depends on the $K-1$ proposals common to the two chains at any iteration and so $\mathbb{E}[S] = \sum_{k=2}^{K} \mathbb{E}[w_k]$, where $\mathbb{E}[w_k] = \mathbb{E}_{q_\phi}[w_{\theta,\phi}(z_k^\star)] = p_\theta(x)$. Contrary to Jacob et al. (2020a), the lower bound we obtain on P_{meet} is increasing with K instead of decreasing.