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

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

[1] DeepMind

A DERIVATION OF THE COUPLING ESTIMATOR

Consider the expectation in Eq. 13. One choice to estimate the expectation is to build an Markov chain Monte Carlo (MCMC) kernel $\mathcal{K}(\cdot \,|\, u)$ that has $\pi(u)$ as its stationary distribution, run the Markov chain for some number of iterations T to obtain samples $u^{(0)}, u^{(1)}, \ldots, u^{(T)}$, and then approximate the expectation as $\mathbb{E}_{\pi(u)}\left[h(u)\right] \approx \frac{1}{T-t_0+1}\sum_{t=t_0}^{T} h(u^{(t)})$, where the initial t_0 samples are thrown away as they are part of the burn-in period. However, this estimator is biased when the number of iterations T is finite.Instead, MCMC couplings provide an unbiased estimator in finite time [Glynn and Rhee, 2014, Jacob et al., 2020b]. Coupling estimators use two MCMC chains, each with invariant distribution $\pi(\cdot)$, which evolve according to a marginal transition kernel $\mathcal{K}(\cdot \,|\, u)$ and a joint transition kernel $\mathcal{K}_{\mathrm{C}}(\cdot, \cdot \,|\, u, \bar{u})$. In our paper, we rely on a slight variation over the approach of Jacob et al. [2020b] proposed by Vanetti and Doucet [2020], which provides a construction also used by Biswas et al. [2019].

Consider an integer $L \geq 1$. We draw the first Markov chain as $u^{(0)} \sim \pi_0(u)$ and $u^{(t)} \sim \mathcal{K}(u \,|\, u^{(t-1)})$ for $t = 1, \ldots, L$. We then draw $\bar{u}^{(0)}$ (potentially conditionally upon $u^{(L-1)}, u^{(L)}$) such that marginally $\bar{u}^{(0)} \sim \pi_0(z)$. 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 meeting time is defined as $\tau = \inf\{t \geq L : u^{(t)} = \bar{u}^{(t-L)}\}$.

We now provide an informal derivation of the estimator. First, we write the expectation of interest as

$$\mathbb{E}_{\pi(u)}\left[h(u)\right] = \lim_{N\to\infty} \frac{1}{L} \sum_{t=N-L+1}^{N} \mathbb{E}[h(u^{(t)})]$$
$$= \lim_{N\to\infty} \frac{1}{L}\Big\{ \sum_{t=t_0}^{t_0+L-1} \mathbb{E}[h(u^{(t)})] + \sum_{t=t_0+L}^{N} \mathbb{E}[h(u^{(t)})]$$
$$- \sum_{t=t_0}^{N-L} \mathbb{E}[h(u^{(t)})]\Big\}. \tag{16}$$

Since $N \to \infty$, then the term within the limit can be equivalently rewritten as

$$\frac{1}{L} \sum_{t=N-L+1}^{N} \mathbb{E}[h(u^{(t)})] = \frac{1}{L}\Big\{ \sum_{t=t_0}^{t_0+L-1} \mathbb{E}[h(u^{(t)})]$$
$$+ \sum_{t=t_0+L}^{N} \mathbb{E}[h(u^{(t)})] - \sum_{t=t_0+L}^{N} \mathbb{E}[h(u^{(t-L)})]\Big\}. \tag{17}$$

Taking into account that $u^{(t)}$ and $\bar{u}^{(t)}$ have the same marginal distribution, then we can replace $u^{(t-L)}$ with $\bar{u}^{(t-L)}$,

$$\frac{1}{L} \sum_{t=N-L+1}^{N} \mathbb{E}[h(u^{(t)})] = \frac{1}{L}\Big\{ \sum_{t=t_0}^{t_0+L-1} \mathbb{E}[h(u^{(t)})]$$
$$+ \sum_{t=t_0+L}^{N} \mathbb{E}[h(u^{(t)})] - \sum_{t=t_0+L}^{N} \mathbb{E}[h(\bar{u}^{(t-L)})]\Big\}. \tag{18}$$

We now combine the two sums in the right,

$$\frac{1}{L} \sum_{t=N-L+1}^{N} \mathbb{E}[h(u^{(t)})] = \frac{1}{L}\Big\{ \sum_{t=t_0}^{t_0+L-1} \mathbb{E}[h(u^{(t)})]$$
$$+ \sum_{t=t_0+L}^{N} \mathbb{E}[h(u^{(t)}) - h(\bar{u}^{(t-L)})]\Big\}. \tag{19}$$

We now apply the fact that the two chains meet after some time τ, i.e., $u^{(t)} = \bar{u}^{(t-L)}$ for $t \geq \tau$. This gives

$$\frac{1}{L} \sum_{t=N-L+1}^{N} \mathbb{E}[h(u^{(t)})] = \frac{1}{L}\Big\{ \sum_{t=t_0}^{t_0+L-1} \mathbb{E}[h(u^{(t)})]$$
$$+ \sum_{t=t_0+L}^{N\wedge(\tau-1)} \mathbb{E}[h(u^{(t)}) - h(\bar{u}^{(t-L)})]\Big\}. \tag{20}$$

When we consider the limit $N \to \infty$, the sum in the r.h.s. no longer depends on N; instead, it only contains (at most)

$\tau - t_0 - L$ terms,

$$\begin{aligned}\mathbb{E}_{\pi(u)}\left[h(u)\right] &= \lim_{N\to\infty}\frac{1}{L}\sum_{t=N-L+1}^{N}\mathbb{E}[h(u^{(t)})] \qquad (21)\\ &= \frac{1}{L}\Big\{\sum_{t=t_0}^{t_0+L-1}\mathbb{E}[h(u^{(t)})] + \sum_{t=t_0+L}^{\tau-1}\mathbb{E}[h(u^{(t)}) - h(\bar{u}^{(t-L)})]\Big\}.\end{aligned}$$

From this expression, we can obtain the unbiased estimator of $\mathbb{E}_{\pi(u)}\left[h(u)\right]$ as given in Eq. 14.

B MAXIMAL COUPLING KERNEL FOR CATEGORICAL DISTRIBUTIONS

Here we describe the maximal coupling kernel for two categorical distributions [Lindvall, 2002], which is used by the coupled DISIR procedure from Section 4.2 (more specifically, in Line 7 of Algorithm 3).

The kernel $\mathcal{K}_{\text{C-Cat}}$ is summarized in Algorithm 5. It takes two (possibly unnormalized) probability vectors, $(w_1, \ldots, w_K)$ and $(v_1, \ldots, v_K)$, and returns a realization of two coupled categorical random variables $(\ell, \bar{\ell})$. Algorithm 5 is a maximal coupling scheme, in the sense that it achieves the theoretically maximum probability that $\ell = \bar{\ell}$.

Algorithm 5: Maximal coupling kernel for categoricals, $\mathcal{K}_{\text{C-Cat}}(\cdot, \cdot \,|\, (w_1, \ldots, w_K), (v_1, \ldots, v_K))$

```
Input: Two unnormalized probability vectors
       $(w_1, \ldots, w_K)$ and $(v_1, \ldots, v_K)$
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 $k = 1, \ldots, K$
Compute the total variation $\gamma = \frac{1}{2}\sum_{k=1}^{K}|\widetilde{w}_k - \widetilde{v}_k|$
Sample $u \sim$ Uniform$(0, 1)$
if $u \leq 1 - \gamma$ then coupling occurs
    Sample $\ell \sim$ Cat$(p_1, \ldots, p_K)$ with
      $p_k \propto \min(\widetilde{w}_k, \widetilde{v}_k)$
    Return $(\ell, \ell)$
else coupling does not occur
    Sample $\ell \sim$ Cat$(p_1, \ldots, p_K)$ with
      $p_k \propto \max(\widetilde{w}_k - \widetilde{v}_k, 0)$
    Sample $\bar{\ell} \sim$ Cat$(p_1, \ldots, p_K)$ with
      $p_k \propto \max(\widetilde{v}_k - \widetilde{w}_k, 0)$
    Return $(\ell, \bar{\ell})$
end
```

C PROOFS OF PROPOSITIONS AND LEMMAS

C.1 PROPOSITION 5 AND ITS PROOF

The results in this section were derived by Andrieu et al. [2010, Theorem 4] and Andrieu et al. [2018, Theorem 1], but we include them here for completeness.

Proposition 5 (ISIR is invariant w.r.t. the posterior)**.** *Under Assumption 1, for any $K \geq 2$, the ISIR transition kernel $\mathcal{K}_{\text{ISIR}}$ is invariant w.r.t. $p_{\theta,\phi}(z_{1:K}, \ell \,|\, x)$ and the corresponding Markov chain is ergodic. Additionally, if Assumption 3 is satisfied, then for any initial value $(z_{1:K}^{(0)}, \ell^{(0)})$, the total variation distance w.r.t. the target is upper bounded by*

$$\left|\left|\mathcal{K}_{\text{ISIR}}^{T}(\cdot, \cdot \,|\, z_{1:K}^{(0)}, \ell^{(0)}) - p_{\theta,\phi}(\cdot, \cdot \,|\, x)\right|\right|_{TV} \leq \rho_K^T, \qquad (22)$$

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

We now prove Proposition 5. Under Assumption 1, the extended target distribution

$$p_{\theta,\phi}(z_{1:K}, \ell \,|\, x) = \frac{1}{K}\, p_\theta(z_\ell \,|\, x) \prod_{k=1, k\neq\ell}^{K} q_\phi(z_k \,|\, x) \qquad (23)$$

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

$$\begin{aligned}\mathcal{K}_{\text{ISIR}}(z_{1:K}^{\star}, \ell^{\star} \,|\, z_{1:K}, \ell) &= \sum_{\ell_{\text{aux}}=1}^{K}\frac{1}{K}\delta_{z_\ell}(z_{\ell_{\text{aux}}}^{\star})\\ &\times \left(\prod_{k=1, k\neq\ell_{\text{aux}}}^{K} q(z_k^{\star} \,|\, x)\right)\frac{w_{\theta,\phi}(z_{\ell^\star}^{\star})}{\sum_{k=1}^{K} w_{\theta,\phi}(z_k^{\star})}. \qquad (24)\end{aligned}$$

To prove invariance, the marginalization $\sum_{\ell=1}^{K}\int p_{\theta,\phi}(z_{1:K}, \ell \,|\, x)\mathcal{K}_{\text{ISIR}}(z_{1:K}^{\star}, \ell^{\star} \,|\, z_{1:K}, \ell)\mathrm{d}z_{1:K}$ should be equal to $p_{\theta,\phi}(z_{1:K}^{\star}, \ell^{\star} \,|\, x)$. Indeed,

$$\begin{aligned}&\sum_{\ell=1}^{K}\int p_{\theta,\phi}(z_{1:K}, \ell \,|\, x)\mathcal{K}_{\text{ISIR}}(z_{1:K}^{\star}, \ell^{\star} \,|\, z_{1:K}, \ell)\mathrm{d}z_{1:K}\\ &= \sum_{\ell_{\text{aux}}=1}^{K} p_{\theta,\phi}(z_{1:K}^{\star}, \ell_{\text{aux}} \,|\, x)p_{\theta,\phi}(\ell^{\star} \,|\, z_{1:K}^{\star}, x)\\ &= p_{\theta,\phi}(z_{1:K}^{\star}, \ell^{\star} \,|\, x). \qquad (25)\end{aligned}$$

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 $p_{\theta,\phi}(z_{1:K}^{\star}, \ell_{\text{aux}} \,|\, x)$ (from Eq. 23), and we have applied that $p_{\theta,\phi}(\ell^{\star} \,|\, z_{1:K}^{\star}, x)$ is a categorical distribution with probability proportional to $w_{\theta,\phi}(z_{\ell^\star}^{\star})$. 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 $(z_{1:K}^{(t)}, \ell^{(t)})_{t\geq 0}$ according to $\mathcal{K}_{\text{ISIR}}$, the $\ell^{(t)}$-th latent variable, i.e., $(z^{(t)} := z_{\ell^{(t)}}^{(t)})_{t\geq 0}$, 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. 24 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

$$\begin{aligned}\mathcal{K}_{\text{ISIR,orig}}(z^\star \,|\, z) &= \frac{1}{K}\sum_{\ell_{\text{aux}}=1}^{K}\sum_{\ell^\star=1}^{K}\int \delta_z(z^\star_{\ell_{\text{aux}}}) \qquad (26)\\ &\times \left(\prod_{k\neq \ell_{\text{aux}}} q_\phi(z^\star_k \,|\, x)\right)\frac{w_{\theta,\phi}(z^\star_{\ell^\star})}{\sum_{k=1}^{K} w_{\theta,\phi}(z^\star_k)}\delta_{z^\star_{\ell^\star}}(z^\star)\mathrm{d}z^\star_{1:K}\\ &= \sum_{\ell^\star=1}^{K}\int \delta_z(z^\star_1)\left(\prod_{k=2}^{K} q_\phi(z^\star_k \,|\, x)\right)\\ &\times \frac{w_{\theta,\phi}(z^\star_{\ell^\star})}{\sum_{k=1}^{K} w_{\theta,\phi}(z^\star_k)}\delta_{z^\star_{\ell^\star}}(z^\star)\mathrm{d}z^\star_{1:K},\end{aligned}$$

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

$$\begin{aligned}\mathcal{K}_{\text{ISIR,orig}}(z^\star \,|\, z) &\geq \sum_{\ell^\star=2}^{K}\int \delta_z(z^\star_1)\left(\prod_{k=2}^{K} q_\phi(z^\star_k \,|\, x)\right) \qquad (27)\\ &\times \frac{w_{\theta,\phi}(z^\star_{\ell^\star})}{\sum_{k=1}^{K} w_{\theta,\phi}(z^\star_k)}\delta_{z^\star_{\ell^\star}}(z^\star)\mathrm{d}z^\star_{1:K}.\end{aligned}$$

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

$$\begin{aligned}\mathcal{K}_{\text{ISIR,orig}}(z^\star \,|\, z) &\geq \sum_{\ell^\star=2}^{K}\int \delta_z(z^\star_1)\left(\prod_{k=2,k\neq\ell^\star}^{K} q_\phi(z^\star_k \,|\, x)\right)\\ &\times \frac{p_\theta(x)p_\theta(z^\star_{\ell^\star} \,|\, x)}{\sum_{k=1}^{K} w_{\theta,\phi}(z^\star_k)}\delta_{z^\star_{\ell^\star}}(z^\star)\mathrm{d}z^\star_{1:K}. \qquad (28)\end{aligned}$$

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

$$\begin{aligned}\mathcal{K}_{\text{ISIR,orig}}(z^\star \,|\, z) &\geq \sum_{\ell^\star=2}^{K}\int \delta_z(z^\star_1)\left(\prod_{k=2,k\neq\ell^\star}^{K} q_\phi(z^\star_k \,|\, x)\right)\\ &\times \frac{p_\theta(x)p_\theta(z^\star_{\ell^\star} \,|\, x)}{2w^{\max}_{\theta,\phi} + \sum_{k=3}^{K} w_{\theta,\phi}(z^\star_k)}\delta_{z^\star_{\ell^\star}}(z^\star)\mathrm{d}z^\star_{1:K}. \qquad (29)\end{aligned}$$

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

$$\mathcal{K}_{\text{ISIR,orig}}(z^\star \,|\, z) \geq \mathbb{E}\left[\frac{(K-1)p_\theta(x)p_\theta(z^\star \,|\, x)}{2w^{\max}_{\theta,\phi} + \sum_{k=3}^{K} w_{\theta,\phi}(z^\star_k)}\right], \qquad (30)$$

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

$$\begin{aligned}\mathcal{K}_{\text{ISIR,orig}}(z^\star \,|\, z) &\geq \frac{(K-1)p_\theta(x)p_\theta(z^\star \,|\, x)}{\mathbb{E}\left[2w^{\max}_{\theta,\phi} + \sum_{k=3}^{K} w_{\theta,\phi}(z^\star_k)\right]}\\ &= \frac{(K-1)p_\theta(x)}{2w^{\max}_{\theta,\phi} + (K-2)p_\theta(x)}p_\theta(z^\star \,|\, x). \qquad (31)\end{aligned}$$

This kernel thus satisfies a minorization condition and thus

$$\left|\left|\mathcal{K}^T_{\text{ISIR,orig}}(\cdot \,|\, z^{(0)}) - p_\theta(\cdot \,|\, x)\right|\right|_{\text{TV}} \leq \rho_K^T, \qquad (32)$$

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

The bound in Eq. 22 follows directly, since $z^{(t)} := z^{(t)}_{\ell^{(t)}}$ 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. 23, so we have for any $T \geq 0$,

$$\begin{aligned}&\left|\left|\mathcal{K}^T_{\text{ISIR,orig}}(\cdot \,|\, z^{(0)}) - p_\theta(\cdot \,|\, x)\right|\right|_{\text{TV}} \qquad (33)\\ &= \left|\left|\mathcal{K}^T_{\text{ISIR}}(\cdot,\cdot \,|\, z^{(0)}_{1:K}, \ell^{(0)}) - p_{\theta,\phi}(\cdot,\cdot \,|\, x)\right|\right|_{\text{TV}}.\end{aligned}$$

C.2 PROOF OF PROPOSITION 1

We now prove here that the DISIR kernel admits $p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K}, \ell \,|\, x)$ defined in Eq. 10 as invariant distribution. The transition kernel $\mathcal{K}_{\text{DISIR}}(\cdot,\cdot \,|\, \xi_{1:K}, \ell)$ is defined through Algorithm 2 and can be written as

$$\begin{aligned}\mathcal{K}_{\text{DISIR}}(\xi^\star_{1:K}, \ell^\star \,|\, \xi_{1:K}, \ell) &= \sum_{\ell_{\text{aux}}=1}^{K}\frac{1}{K}\delta_{\xi_\ell}(\xi^\star_{\ell_{\text{aux}}}) \qquad (34)\\ &\times \prod_{k=1}^{\ell_{\text{aux}}-1} p_\beta(\xi^\star_k \,|\, \xi^\star_{k+1}) \prod_{k=\ell_{\text{aux}}+1}^{K} p_\beta(\xi^\star_k \,|\, \xi^\star_{k-1})\frac{w_{\theta,\phi}(z^\star_{\ell^\star})}{\sum_{k=1}^{K} w_{\theta,\phi}(z^\star_k)},\end{aligned}$$

for $z^\star_k = g_\phi(\xi^\star_k, x)$.

For the kernel to be invariant, the marginalization $\sum_{\ell=1}^{K}\int p^{\text{DISIR}}_{\theta,\phi}(\xi_{1:K}, \ell \,|\, x)\mathcal{K}_{\text{DISIR}}(\xi^\star_{1:K}, \ell^\star \,|\, \xi_{1:K}, \ell)\mathrm{d}\xi_{1:K}$ should be equal to $p^{\text{DISIR}}_{\theta,\phi}(\xi^\star_{1:K}, \ell^\star \,|\, x)$. We obtain this marginalization below. We first define

$$p^{\text{DISIR}}_{\theta,\phi}(\xi \,|\, x) := \frac{w_{\theta,\phi}(g_\phi(\xi, x))q(\xi)}{p_\theta(x)}. \qquad (35)$$

(Eq. 35 gives the marginal distribution of ξ_ℓ obtained after integrating out the rest of latent variables from Eq. 10.) Similarly to the proof in Appendix C.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

$$
\begin{aligned}
&\sum_{\ell=1}^{K} \int p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}, \ell \,|\, x) \mathcal{K}_{\text{DISIR}}(\xi_{1:K}^\star, \ell^\star \,|\, \xi_{1:K}, \ell) \mathrm{d}\xi_{1:K} \\
&= \sum_{\ell=1}^{K} \sum_{\ell_{\text{aux}}=1}^{K} \frac{1}{K} \int \frac{p_{\theta,\phi}^{\text{DISIR}}(\xi_\ell \,|\, x)}{K} \delta_{\xi_\ell}(\xi_{\ell_{\text{aux}}}^\star) \\
&\quad \times \prod_{k=1}^{\ell_{\text{aux}}-1} p_\beta(\xi_k^\star \,|\, \xi_{k+1}^\star) \prod_{k=\ell_{\text{aux}}+1}^{K} p_\beta(\xi_k^\star \,|\, \xi_{k-1}^\star) \frac{w_{\theta,\phi}(z_{\ell^\star}^\star)}{\sum_{k=1}^{K} w_{\theta,\phi}(z_k^\star)} \mathrm{d}\xi_\ell \\
&= \sum_{\ell_{\text{aux}}=1}^{K} p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}^\star, \ell_{\text{aux}} \,|\, x) \frac{w_{\theta,\phi}(z_{\ell^\star}^\star)}{\sum_{k=1}^{K} w_{\theta,\phi}(z_k^\star)} \\
&= \sum_{\ell_{\text{aux}}=1}^{K} p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}^\star, \ell_{\text{aux}} \,|\, x) p_{\theta,\phi}(\ell^\star \,|\, x, \xi_{1:K}) \\
&= p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}^\star, \ell^\star \,|\, x). \qquad (36)
\end{aligned}
$$

Here, we have additionally recognized the term $p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}^\star, \ell_{\text{aux}} \,|\, x)$ (see Eq. 10). For the second-to-last step, we have applied the fact that the conditional distribution of ℓ under Eq. 10, $p_{\theta,\phi}^{\text{DISIR}}(\ell \,|\, x, \xi_{1:K})$, 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

$$
\begin{aligned}
&p_{\theta,\phi}^{\text{DISIR}}(\ell \,|\, x, \xi_{1:K}) \propto w_{\theta,\phi}(g_\phi(\xi_\ell, x)) q(\xi_\ell) \qquad (37) \\
&\times \prod_{k=1}^{\ell-1} p_\beta(\xi_k \,|\, \xi_{k+1}) \prod_{k=\ell+1}^{K} p_\beta(\xi_k \,|\, \xi_{k-1}).
\end{aligned}
$$

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

$$
\begin{aligned}
&q(\xi_\ell) \prod_{k=1}^{\ell-1} p_\beta(\xi_k \,|\, \xi_{k+1}) \prod_{k=\ell+1}^{K} p_\beta(\xi_k \,|\, \xi_{k-1}) \\
&\quad = q(\xi_1) \prod_{k=2}^{K} p_\beta(\xi_k \,|\, \xi_{k-1}), \qquad (38)
\end{aligned}
$$

so this product is independent of ℓ and we have

$$
p_{\theta,\phi}^{\text{DISIR}}(\ell \,|\, x, \xi_{1:K}) \propto w_{\theta,\phi}(g_\phi(\xi_\ell, x)). \qquad (39)
$$

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

C.3 PROOF OF PROPOSITION 1

We prove here Lemma 1. From the DISIR invariant distribution in Eq. 10, it follows that the marginal distribution of ξ_ℓ is given by $p_{\theta,\phi}^{\text{DISIR}}(\xi_\ell \,|\, x)$ (Eq. 35).

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

$$
\begin{aligned}
&\mathbb{E}_{p_{\theta,\phi}^{\text{DISIR}}(\xi \,|\, x)} \left[f(g_\phi(\xi, x)) \right] \qquad (40) \\
&= \int f(g_\phi(\xi, x)) \frac{w_{\theta,\phi}(g_\phi(\xi, x)) q(\xi)}{p_\theta(x)} \mathrm{d}\xi \\
&= \int f(g_\phi(\xi, x)) \frac{p_\theta(x, g_\phi(\xi, x))}{q_\phi(g_\phi(\xi, x) \,|\, x)} \frac{q(\xi)}{p_\theta(x)} \mathrm{d}\xi.
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathbb{E}_{p_{\theta,\phi}^{\text{DISIR}}(\xi \,|\, x)} \left[f(g_\phi(\xi, x)) \right] = \int f(z) \frac{p_\theta(x, z)}{q_\phi(z \,|\, x)} \frac{q_\phi(z \,|\, x)}{p_\theta(x)} \mathrm{d}z \\
&= \mathbb{E}_{p_\theta(z \,|\, x)} \left[f(z) \right]. \qquad (41)
\end{aligned}
$$

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 $(\xi_{1:K}, \ell) \sim p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}, \ell \,|\, x)$ then, for $z_k = g_\phi(\xi_k, \ell)$, we have $(z_{1:K}, \ell) \sim p_{\theta,\phi}(z_{1:K}, \ell \,|\, x)$.

C.4 PROOF OF PROPOSITION 2

We establish here the identity in Eq 10. This is a generalization of Theorem 6 in Andrieu et al. [2010]. From the result in Lemma 1 and the definition of $p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}, \ell \,|\, x)$ (Eq. 10), it follows that

$$
\begin{aligned}
&\int h(z) p_\theta(z \,|\, x) \mathrm{d}z \qquad (42) \\
&= \sum_{\ell=1}^{K} \int h(g_\phi(\xi_\ell, x)) p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K}, \ell \,|\, x) \mathrm{d}\xi_{1:K} \\
&= \int \Big[\sum_{\ell=1}^{K} h(g_\phi(\xi_\ell, x)) p_{\theta,\phi}^{\text{DISIR}}(\ell \,|\, x, \xi_{1:K}) \Big] p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K} \,|\, x) \mathrm{d}\xi_{1:K} \\
&= \int \Big[\sum_{\ell=1}^{K} \widetilde{w}_{\theta,\phi}^{(\ell)} h(g_\phi(\xi_\ell, x)) \Big] p_{\theta,\phi}^{\text{DISIR}}(\xi_{1:K} \,|\, x) \mathrm{d}\xi_{1:K},
\end{aligned}
$$

where $\widetilde{w}_{\theta,\phi}^{(\ell)} \propto w_{\theta,\phi}(g_\phi(\xi_\ell, x))$ are the normalized importance weights. The last equality follows from Eq. 39 used in the proof of Appendix C.2. The result thus follows.

C.5 PROOF OF PROPOSITION 3

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. 14. The proof follows the approach of Jacob et al. [2020b, Proposition 3.1] and Middleton et al. [2020, Theorem 1] but some details differ.

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 $\mathbb{E}[\tau] < \infty$, so the estimator $\hat{H} := \hat{\pi}(h)$ from Eq. 14 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

$$\hat{\pi}_N(h) = \frac{1}{L}\left(\sum_{t=k}^{k+L-1} h(u^{(t)}) + \sum_{t=k+L}^{N} h(u^{(t)}) - h(\bar{u}^{(t-L)})\right) = \frac{1}{L}\sum_{t=k}^{N} \Delta_t, \qquad (43)$$

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 $\mathbb{E}[\tau] < \infty$, 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

$$\mathbb{E}\left[(\hat{\pi}_N(h) - \hat{\pi}_{N'}(h))^2\right] = \frac{1}{L^2}\sum_{s=N+1}^{N'}\sum_{t=N+1}^{N'} \mathbb{E}[\Delta_s \Delta_t]$$
$$\leq \frac{1}{L^2}\sum_{s=N+1}^{N'}\sum_{t=N+1}^{N'} \mathbb{E}\left[\Delta_s^2\right]^{1/2}\mathbb{E}\left[\Delta_t^2\right]^{1/2}$$
$$= \frac{1}{L^2}\left(\sum_{t=N+1}^{N'} \mathbb{E}\left[\Delta_t^2\right]^{1/2}\right)^2. \qquad (44)$$

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

$$\mathbb{E}\left[\Delta_t^2\right] \leq \mathbb{E}\left[|\Delta_t|^{2+\eta}\right]^{\frac{1}{1+\frac{\eta}{2}}} \mathbb{E}\left[\mathbb{I}_{\tau>t}\right]^{\frac{\eta}{2+\eta}}$$
$$\leq D^{\frac{1}{1+\frac{\eta}{2}}} \mathbb{P}(\tau > t)^{\frac{\eta}{2+\eta}}, \qquad (45)$$

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

$$\mathbb{E}\left[(\hat{\pi}_N(h) - \hat{\pi}_{N'}(h))^2\right] \qquad (46)$$
$$\leq \frac{1}{L^2}\left(\sum_{t=N+1}^{N'}\left(D^{\frac{1}{1+\frac{\eta}{2}}}\mathbb{P}(\tau > t)^{\frac{\eta}{2+\eta}}\right)^{\frac{1}{2}}\right)^2$$
$$= \frac{1}{L^2} D^{\frac{1}{1+\frac{\eta}{2}}}\left(\sum_{t=N+1}^{N'} \mathbb{P}(\tau > t)^{\frac{1}{2}\frac{\eta}{2+\eta}}\right)^2.$$

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

$$\sum_{t=N+1}^{\infty} \mathbb{P}(\tau > t)^{\lambda} \leq C \sum_{t=N+1}^{\infty} \frac{1}{t^{\lambda\kappa}} < \infty. \qquad (47)$$

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

$$\mathbb{E}[\hat{\pi}(h)] = \mathbb{E}\left[\lim_{N\to\infty} \hat{\pi}_N(h)\right] = \lim_{N\to\infty} \mathbb{E}[\hat{\pi}_N(h)], \qquad (48)$$

and, under condition (a), we have

$$\lim_{N\to\infty} \mathbb{E}[\hat{\pi}_N(h)] = \lim_{N\to\infty} \frac{1}{L}\sum_{t=N-L+1}^{N} \mathbb{E}\left[h(u^{(t)})\right] = \pi(h). \qquad (49)$$

C.6 PROOF OF PROPOSITION 4

To prove Proposition 4, we need to check that the conditions (a) to (d) of Proposition 3 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 5. 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 5 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 $(w_k)_{k=1,\ldots,K}$ and $(v_k)_{k=1,\ldots,K}$ 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 $k = 2, \ldots, K$. The normalized probabilities are thus

$$\widetilde{w}_k = \frac{w_k}{w_1 + S}, \quad \widetilde{v}_k = \frac{v_k}{v_1 + S}, \qquad (50)$$

where

$$S = \sum_{k=2}^{K} v_k = \sum_{k=2}^{K} w_k. \tag{51}$$

Therefore, the probability of coupling of ISIR is

$$P_{\text{meet}} \geq \mathbb{E}\left[(1-\gamma)\frac{\sum_{k=2}^{K}\min(\widetilde{w}_k, \widetilde{v}_k)}{\sum_{k=1}^{K}\min(\widetilde{w}_k, \widetilde{v}_k)}\right], \tag{52}$$

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

$$1-\gamma = 1 - \frac{1}{2}\sum_{k=1}^{K}|\widetilde{w}_k - \widetilde{v}_k| = \sum_{k=1}^{K}\min(\widetilde{w}_k, \widetilde{v}_k). \tag{53}$$

Thus, we have

$$P_{\text{meet}} \geq \mathbb{E}\left[\sum_{k=2}^{K}\min(\widetilde{w}_k, \widetilde{v}_k)\right]. \tag{54}$$

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

$$\begin{aligned} P_{\text{meet}} &\geq \mathbb{E}\left[\sum_{k=2}^{K}\min\left(\frac{w_k}{w_1+S}, \frac{v_k}{v_1+S}\right)\right] \\ &\geq \mathbb{E}\left[\frac{1}{K w^{\max}_{\theta,\phi}}\sum_{k=2}^{K} w_k\right] = \frac{p_\theta(x)}{w^{\max}_{\theta,\phi}} - \frac{p_\theta(x)}{K w^{\max}_{\theta,\phi}}. \end{aligned} \tag{55}$$

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 (as expected) increasing with K instead of decreasing. From this lower bound, we can also deduce directly an upper bound on $\mathbb{E}[\tau]$:

$$\mathbb{E}[\tau] \leq (L-1) + \frac{1}{P_{\text{meet}}} \tag{56}$$

where $1/P_{\text{meet}}$ is the expectation of a geometric random variable of success probability P_{meet}. By plugging the lower bound on the r.h.s. of (55) in (56), we obtain a decreasing upper bound on $\mathbb{E}[\tau]$ converging to $L-1+\frac{w^{\max}_{\theta,\phi}}{p_\theta(x)}$ as $K \to \infty$.

D ADDITIONAL EXPERIMENTAL DETAILS

D.1 GRADIENTS OF THE PPCA MODEL

Figure 5 shows the errors when estimating the gradient w.r.t. a randomly chosen weight term of the probabilistic principal component analysis (PPCA) model; it looks qualitatively similar to Figure 2.

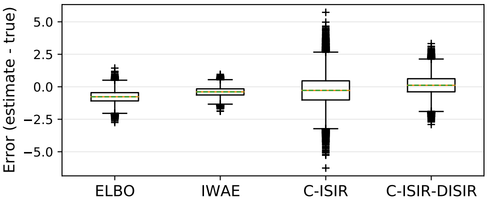

Figure 5: Boxplot representation of the error of different estimators for the gradient w.r.t. one of the weights of the PPCA model.

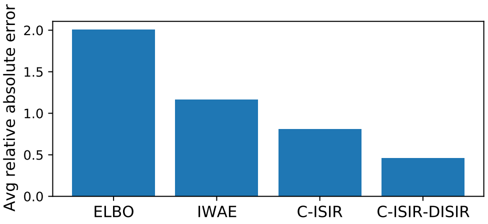

Figure 6: Relative error (in absolute value) for the PPCA model, averaged over all components of the gradient.

Figure 6 shows the relative error in absolute value, averaged across all the components of the gradient. C-ISIR-DISIR exhibits the smaller error.

D.2 EXPERIMENTAL SETUP FOR THE VAE

Binarized MNIST. For binarized MNIST, the model is $p_\theta(x,z) = \mathcal{N}(z;0,I)p_\theta(x\,|\,z)$, where $p_\theta(x\,|\,z)$ 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 $q_\phi(z\,|\,x)$ 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). The RMSProp learning rate is 5×10^{-4} and the batchsize is 100.

Fashion-MNIST and CIFAR-10. We define the likelihood $p_\theta(z\,|\,x)$ 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 described above (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 $16\times16\times1$ and ReLu activations, followed by three con-

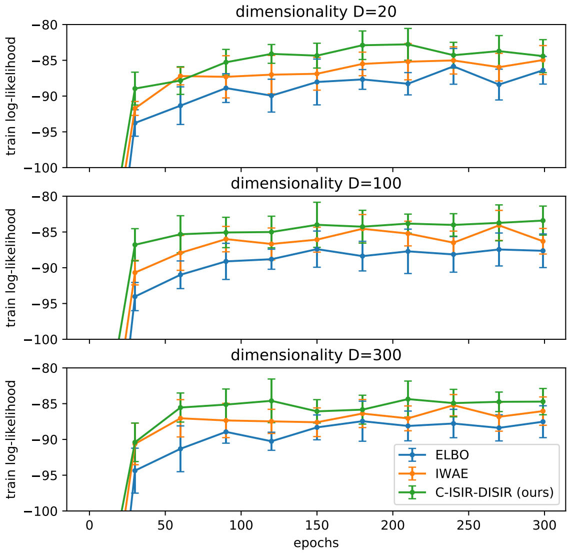

Figure 7: Train log-likelihood on a VAE fitted on binarized MNIST. The estimator from Algorithm 4 provides better performance for multiple values of the dimensionality D.

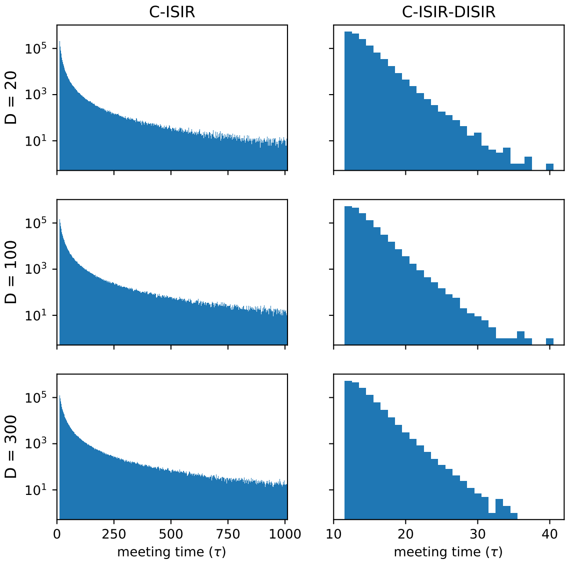

Figure 8: Histograms of the meeting time for a VAE fitted on binarized MNIST. The histograms corresponding to C-ISIR have significantly heavier tails, which results in higher computational complexity of the overall estimator.

volutional 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.

D.3 TRAIN LOG-LIKELIHOOD ON BINARIZED MNIST

Figure 7 shows (an estimate of) the evolution of the train log-likelihood for the VAE fitted on binarized MNIST for different values of the dimensionality D.

D.4 HISTOGRAMS OF THE MEETING TIME

Figure 8 shows the histograms of the meeting time for the experiment on binarized MNIST from Section 6.2. The meeting time behaves similarly across different values of the dimensionality D, although the histogram gets heavier-tailed for C-ISIR when D increases.

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.

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

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

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.

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

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.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. In *International Conference on Learning Representations*, 2017.

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

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.