Monte Carlo VAE

SUPPLEMENTARY DOCUMENT

Achille Thin [1] **Nikita Kotelevskii** [2] **Alain Durmus** [3] **Maxim Panov** [2] **Eric Moulines** [1 4 5] **Arnaud Doucet** [6]

1. Notations and definitions

Let $(\mathsf{X}, \mathcal{X})$ be a measurable space. A *Markov kernel* N on $\mathsf{X} \times \mathcal{X}$ is a mapping $N : \mathsf{X} \times \mathcal{X} \to [0, 1]$ satisfying the following conditions:

(i) for every $x \in \mathsf{X}$, the mapping $N(x, \cdot) : A \mapsto N(x, A)$ is a probability of on $\mathcal{X}$,

(ii) for every $A \in \mathcal{X}$, the mapping $N(\cdot, A) : x \mapsto N(x, A)$ is a measurable function from $(\mathsf{X}, \mathcal{X})$ to $([0, 1], \mathcal{B}([0, 1])$, where $\mathcal{B}([0, 1])$ denotes the borelian sets of $[0, 1]$.

Let λ be a positive σ-finite measure on $(\mathsf{X}, \mathcal{X})$ and $n : \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$ be a nonnegative function, measurable with respect to the product σ-field $\mathcal{X} \otimes \mathcal{X}$. Then, the application N defined on $\mathsf{X} \times \mathcal{X}$ by

$$N(x, A) = \int_A n(x, y)\lambda(\mathrm{d}y) ,$$

is a kernel. The function n is called the density of the kernel N w.r.t. the measure λ. The kernel N is Markovian if and only if $\int_{\mathsf{X}} n(x, y)\lambda(\mathrm{d}y) = 1$ for all $x \in \mathsf{X}$.

Let N be a kernel on $\mathsf{X} \times \mathcal{X}$ and f be a nonnegative function. A function $Nf : \mathsf{X} \to \mathbb{R}_+$ is defined by setting, for $x \in \mathsf{X}$,

$$Nf(x) = \int_{\mathsf{X}} N(x, \mathrm{d}y)f(y) .$$

Let μ be a probability on $(\mathsf{X}, \mathcal{X})$. For $A \in \mathcal{X}$, define

$$\mu N(A) = \int_{\mathsf{X}} \mu(\mathrm{d}x)\ N(x,\ A) .$$

If N is Markovian, then μN is a probability on $(\mathsf{X}, \mathcal{X})$.

2. Experiences

2.1. Toy example

We first describe additional experiments on the toy dataset introduced in Section 4.2.

Recall that we generate some i.i.d. data $x = (x_i)_{i=1}^N \in \mathbb{R}^N$ from the i.i.d. latent variables $z = (z_i)_{i=1}^N \in \mathbb{R}^{2N}$ as follows for $\eta > 0$: $z_i \sim \mathrm{N}(0; \mathrm{Id})$ and $x_i \mid z_i \sim \mathrm{N}(\eta \cdot (\|z_i\| + \zeta), \sigma^2) = p_\theta(x_i \mid z_i)$.

This example, presented for $z \in \mathbb{R}^2$, easily extends to the case where z lies in $\mathbb{R}^d$, with d increasing from 2 to 300. We tackle here the problem at estimating the parameter $\theta = (\eta, \zeta)$ when d varies.

We show in Figure S1 the error $\|\hat{\theta} - \theta\|^2$ for the different methods. The increased flexibility of the posterior proves more effective for estimating the true parameters of the generative model.

2.2. Probabilistic Principal Component Analysis

We detail the impact of the learnable reverse kernels on the variance of the estimator and looseness of the ELBO. In our experiments, reverse kernels were given by fully-connected neural networks. We train K different reverse kernels $\{l_k\}_{k=0}^{K-1}$

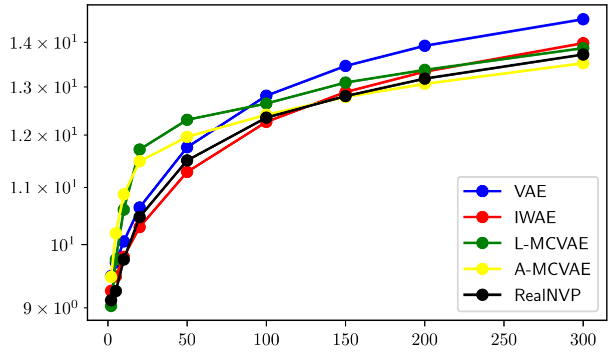

Figure S1. Squared error for parameter's estimates, obtained using different models.

for the K transitions, each given by a separate neural network, and amortized over the observation x, similarly to (Salimans et al., 2015; Huang et al., 2018). Given the parameters (θ, ϕ), we train these kernels for a large number of epochs using the SIS objective (14) and the Adam optimizer (Kingma & Ba, 2014). In particular, we display in Figure S2 the different estimators to be compared. It is easily seen that reverse kernels can not provide reasonable and stable density estimates. At the same time, we observe the variance of the gradient is higher in those models than in the ones we present in the main text. This motivates our approach bypassing the optimization of the reverse kernels.

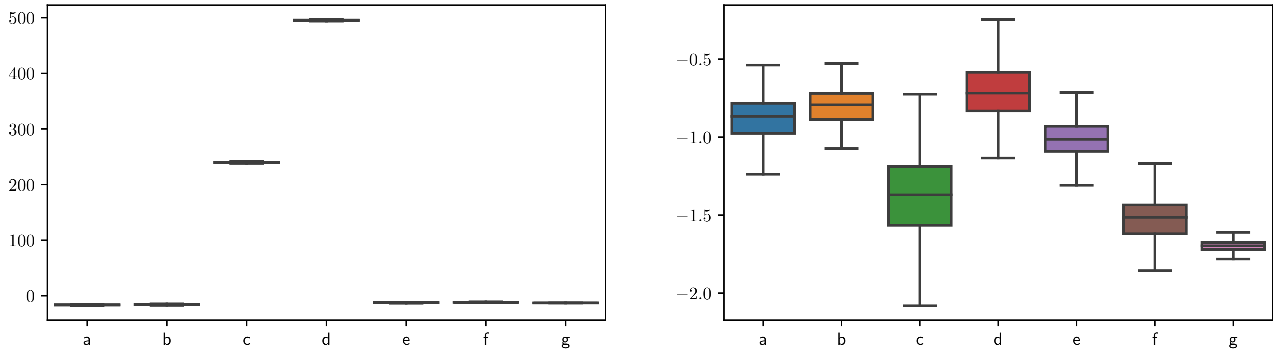

Figure S2. Representation of the different estimators (left) and their gradient (right) of the true log likelihood. From left to right, a/ L-MCVAE, $K = 5$, b/ L-MCVAE, $K = 10$, c/ L-MCVAE, $K = 1$, learnable reverse, d/ L-MCVAE, $K = 2$ learnable reverse, e/ A-MCVAE, $K = 5$, f/ A-MCVAE, $K = 10$, g/ A-MCVAE, $K = 5$ with control variates.

2.3. Additional experimental results

We display in this section the full results on MNIST, CelebA and CIFAR respectively of the different models as well as the effect of the different annealing schemes (respectively in Table 1, Table 2 and 3).

3. Proofs

3.1. Proof of SIS and AIS Identities

Proposition S1. *Let $\{\Gamma_k\}_{k=0}^K$ be a sequence of distributions on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $\{M_k\}_{k=1}^K$ and $\{L_k\}_{k=0}^{K-1}$ be Markov kernels. Assume that for each $k \in \{0, \ldots, K-1\}$, there exists a positive measurable function $w_k : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}_+$ such that*

$$\Gamma_k(\mathrm{d}z_k) L_{k-1}(z_k, \mathrm{d}z_{k-1}) = \Gamma_{k-1}(\mathrm{d}z_{k-1}) M_k(z_{k-1}, \mathrm{d}z_k) w_k(z_{k-1}, z_k) \; . \tag{S1}$$

Table 1. Results of the different models on MNIST with different annealing schemes.

number of epoches	*ELBO:* 10	30	100	*NLL:* 10	30	100
VAE	95.26 ± 0.5	91.58 ± 0.27	89.7 ± 0.19	89.83 ± 0.59	86.86 ± 0.26	85.22 ± 0.07
IWAE, K= 10	91.42 ± 0.21	88.56 ± 0.07	87.17 ± 0.19	88.54 ± 0.27	86.07 ± 0.1	84.82 ± 0.1
IWAE, K= 50	90.34 ± 0.27	87.5 ± 0.16	86.05 ± 0.11	89.4 ± 0.25	86.54 ± 0.15	85.05 ± 0.1
L-MCVAE Fixed, K= 5	96.6 ± 3.51	88.8 ± 0.46	87.77 ± 0.12	90.63 ± 2.19	85.85 ± 0.27	85.07 ± 0.04
L-MCVAE Sigmoidal, K= 5	95.48 ± 2.29	88.87 ± 0.82	87.81 ± 0.53	90.05 ± 1.63	85.92 ± 0.62	85.16 ± 0.38
L-MCVAE All learnable, K= 5	96.62 ± 3.24	88.58 ± 0.75	87.51 ± 0.41	90.59 ± 2.01	85.68 ± 0.49	84.92 ± 0.24
L-MCVAE Fixed, K= 10	95.98 ± 3.91	88.36 ± 0.7	87.38 ± 0.35	90.5 ± 2.23	85.75 ± 0.33	85.0 ± 0.11
L-MCVAE Sigmoidal, K= 10	96.78 ± 0.47	88.35 ± 0.63	87.17 ± 0.52	91.13 ± 0.27	85.72 ± 0.31	84.84 ± 0.26
L-MCVAE All learnable, K= 10	96.78 ± 1.06	87.99 ± 0.71	86.8 ± 0.66	91.33 ± 0.61	85.47 ± 0.46	84.58 ± 0.39
A-MCVAE Fixed, K= 3	96.21 ± 3.43	88.64 ± 0.78	87.63 ± 0.42	90.42 ± 2.34	85.77 ± 0.65	85.02 ± 0.37
A-MCVAE Sigmoidal, K= 3	96.59 ± 2.31	88.96 ± 0.4	87.86 ± 0.06	90.85 ± 1.62	85.97 ± 0.34	85.17 ± 0.1
A-MCVAE All learnable, K= 3	95.44 ± 2.68	88.79 ± 0.63	87.78 ± 0.37	89.9 ± 1.68	85.96 ± 0.59	85.23 ± 0.41
A-MCVAE Fixed, K= 5	95.55 ± 2.96	87.99 ± 0.57	87.03 ± 0.27	90.39 ± 2.21	85.6 ± 0.67	84.84 ± 0.38
A-MCVAE Sigmoidal, K= 5	96.56 ± 2.02	88.51 ± 0.31	87.46 ± 0.48	91.62 ± 1.55	85.96 ± 0.06	85.15 ± 0.21
A-MCVAE All learnable, K= 5	95.81 ± 1.72	88.11 ± 0.13	87.14 ± 0.18	90.79 ± 1.14	85.71 ± 0.28	84.95 ± 0.04
VAE with RealNVP	95.23 ± 0.33	91.69 ± 0.15	89.62 ± 0.17	89.98 ± 0.24	86.88 ± 0.05	85.23 ± 0.18

Table 2. Full results of the different models on CelebA. All scores must be added 11400 in this table.

number of epoches	*ELBO:* 10	30	100	*NLL:* 10	30	100
VAE	23.78 ± 1.95	17.99 ± 0.4	14.72 ± 0.16	17.35 ± 1.7	12.68 ± 0.62	10.11 ± 0.32
IWAE, K= 10	20.59 ± 0.71	15.45 ± 0.52	12.2 ± 0.3	18.25 ± 0.6	13.18 ± 0.42	10.14 ± 0.31
IWAE, K= 50	19.05 ± 0.39	13.59 ± 0.5	10.48 ± 0.89	19.08 ± 0.42	13.17 ± 0.54	10.12 ± 0.86
L-MCVAE Fixed, K= 5	21.93 ± 1.34	13.12 ± 1.27	12.03 ± 1.21	16.65 ± 1.55	10.12 ± 1.38	9.14 ± 1.27
L-MCVAE Sigmoidal, K= 5	21.61 ± 1.48	12.72 ± 0.43	11.6 ± 0.37	16.42 ± 1.47	9.62 ± 0.47	8.72 ± 0.4
L-MCVAE All learnable, K= 5	20.75 ± 0.65	12.99 ± 0.7	11.91 ± 0.61	16.16 ± 0.93	10.01 ± 0.72	9.03 ± 0.64
L-MCVAE Fixed, K= 10	21.49 ± 0.03	12.83 ± 0.57	11.76 ± 0.56	17.67 ± 0.75	10.26 ± 0.9	9.24 ± 0.79
L-MCVAE Sigmoidal, K= 10	19.44 ± 0.82	11.81 ± 0.45	10.7 ± 0.4	15.67 ± 1.48	9.24 ± 0.8	8.24 ± 0.73
L-MCVAE All learnable, K= 10	20.7 ± 1.15	11.81 ± 0.34	10.6 ± 0.23	17.0 ± 1.87	9.29 ± 0.73	8.26 ± 0.52
A-MCVAE Fixed, K= 3	21.59 ± 1.5	13.94 ± 0.42	12.84 ± 0.3	16.64 ± 1.37	10.98 ± 0.48	9.95 ± 0.3
A-MCVAE Sigmoidal, K= 3	23.63 ± 1.19	14.17 ± 0.26	12.96 ± 0.18	18.0 ± 0.54	11.09 ± 0.2	10.11 ± 0.13
A-MCVAE All learnable, K= 3	22.11 ± 1.66	14.62 ± 0.35	13.54 ± 0.18	17.38 ± 1.54	11.68 ± 0.33	10.67 ± 0.16
A-MCVAE Fixed, K= 5	20.13 ± 1.11	13.11 ± 0.38	11.99 ± 0.56	16.71 ± 1.47	10.64 ± 0.24	9.63 ± 0.32
A-MCVAE Sigmoidal, K= 5	20.95 ± 1.18	12.42 ± 0.42	11.13 ± 0.37	17.42 ± 1.49	9.97 ± 0.59	8.82 ± 0.57
A-MCVAE All learnable, K= 5	22.17 ± 0.17	12.73 ± 0.09	11.46 ± 0.15	18.97 ± 1.04	10.41 ± 0.28	9.22 ± 0.16
VAE with RealNVP	15.56 ± 0.29	13.60 ± 0.35	12.21 ± 0.27	10.69 ± 0.19	9.09 ± 0.26	8.98 ± 0.2

Table 3. Results of the different models on CIFAR-10 with different annealing schemes. All scores must be added 2800 in this table.

number of epoches	*ELBO:* 10	30	100	*NLL:* 10	30	100
VAE	69.57 ± 0.08	69.55 ± 0.51	68.84 ± 0.06	68.51 ± 0.07	68.41 ± 0.33	67.9 ± 0.03
IWAE, K= 10	69.82 ± 0.03	69.35 ± 0.03	69.36 ± 0.36	68.56 ± 0.03	68.0 ± 0.03	68.02 ± 0.4
IWAE, K= 50	69.94 ± 0.08	69.55 ± 0.04	69.43 ± 0.03	69.15 ± 0.15	68.37 ± 0.18	67.93 ± 0.02
L-MCVAE Fixed, K= 5	70.86 ± 0.53	68.44 ± 0.18	68.12 ± 0.11	69.37 ± 0.37	67.78 ± 0.1	67.53 ± 0.07
L-MCVAE Sigmoidal, K= 5	70.9 ± 0.59	68.46 ± 0.13	68.12 ± 0.11	69.42 ± 0.39	67.77 ± 0.11	67.51 ± 0.08
L-MCVAE All learnable, K= 5	70.62 ± 0.41	68.55 ± 0.18	68.09 ± 0.1	69.15 ± 0.38	67.73 ± 0.07	67.5 ± 0.07
L-MCVAE Fixed, K= 10	70.67 ± 0.42	68.37 ± 0.06	69.07 ± 1.49	69.62 ± 0.54	67.78 ± 0.06	67.51 ± 0.03
L-MCVAE Sigmoidal, K= 10	70.99 ± 0.59	68.36 ± 0.04	68.03 ± 0.0	69.8 ± 0.67	67.76 ± 0.04	67.51 ± 0.03
L-MCVAE All learnable, K= 10	71.19 ± 0.79	68.36 ± 0.03	68.01 ± 0.04	69.95 ± 0.62	67.78 ± 0.07	67.5 ± 0.05
A-MCVAE Fixed, K= 3	69.97 ± 0.99	68.48 ± 0.29	68.18 ± 0.16	69.26 ± 0.76	67.77 ± 0.18	67.55 ± 0.1
A-MCVAE Sigmoidal, K= 3	70.5 ± 1.18	68.45 ± 0.28	68.19 ± 0.18	69.18 ± 0.8	67.77 ± 0.19	67.56 ± 0.11
A-MCVAE All learnable, K= 3	70.69 ± 1.23	68.44 ± 0.3	68.17 ± 0.18	69.36 ± 0.89	67.76 ± 0.2	67.55 ± 0.11
A-MCVAE Fixed, K= 5	70.37 ± 1.04	68.31 ± 0.21	68.04 ± 0.1	69.36 ± 0.87	67.73 ± 0.17	67.51 ± 0.08
A-MCVAE Sigmoidal, K= 5	70.89 ± 0.38	68.4 ± 0.05	68.07 ± 0.04	69.71 ± 0.33	67.8 ± 0.04	67.53 ± 0.02
A-MCVAE All learnable, K= 5	70.1 ± 0.89	68.28 ± 0.2	68.01 ± 0.08	69.23 ± 0.75	67.71 ± 0.15	67.5 ± 0.07
VAE with RealNVP	70.01 ± 0.12	69.51 ± 0.07	69.19 ± 0.13	68.73 ± 0.05	68.35 ± 0.05	68.05 ± 0.02

Then,

$$\Gamma_0(\mathrm{d}z_0)\prod_{k=1}^{K} M_k(z_{k-1},\mathrm{d}z_k)\prod_{k=1}^{K} w_k(z_{k-1},z_k) = \Gamma_K(\mathrm{d}z_K)\prod_{k=K}^{1} L_{k-1}(z_k,\mathrm{d}z_{k-1})\;. \tag{S2}$$

Proof. We prove by induction that for $k \in \{1,\ldots,K\}$,

$$\Gamma_0(\mathrm{d}z_0)\prod_{i=1}^{k} M_i(z_{i-1},\mathrm{d}z_i)\prod_{i=1}^{k} w_i(z_{i-1},z_i) = \Gamma_k(\mathrm{d}z_k)\prod_{i=k}^{1} L_{i-1}(z_i,\mathrm{d}z_{i-1})\;. \tag{S3}$$

Eq. (S3) is satisfied for $k = 1$ by (S1). Assume that (S3) is satisfied for $k \leqslant K-1$. By (S1),

$$\begin{aligned}
\Gamma_{k+1}(\mathrm{d}z_{k+1})\prod_{i=k+1}^{1} L_{i-1}(z_i,\mathrm{d}z_{i-1}) &= \Gamma_{k+1}(\mathrm{d}z_{k+1})L_k(z_{k+1},\mathrm{d}z_k)\prod_{i=k}^{1} L_{i-1}(z_i,\mathrm{d}z_{i-1})\\
&= \Gamma_k(\mathrm{d}z_k)M_{k+1}(z_k,\mathrm{d}z_{k+1})w_{k+1}(z_k,z_{k+1})\prod_{i=k}^{1} L_{i-1}(z_i,\mathrm{d}z_{i-1})\\
&= M_{k+1}(z_k,\mathrm{d}z_{k+1})w_{k+1}(z_k,z_{k+1})\Gamma_0(\mathrm{d}z_0)\prod_{i=1}^{k} M_i(z_{i-1},\mathrm{d}z_i)\prod_{i=1}^{k} w_i(z_{i-1},z_i)
\end{aligned}$$

which concludes the proof. □

We now highlight conditions under which (S1) is satisfied.

1. Assume that $\{\Gamma_k\}_{k=0}^{K}$ have positive densities w.r.t. to the Lebesgue measure, *i.e.* $\Gamma_k(\mathrm{d}z_k) = \Gamma_k(z_k)\mathrm{d}z_k$ and that the kernels $\{M_k\}_{k=1}^{K}$ and $\{L_k\}_{k=0}^{K-1}$ have positive transition densities $M_k(z_{k-1},\mathrm{d}z_k) = m_k(z_{k-1},z_k)\mathrm{d}z_k$ and $L_{k-1}(z_k,\mathrm{d}z_{k-1}) = \ell_{k-1}(z_k,z_{k-1})\mathrm{d}z_{k-1}$, $k \in \{1,\ldots,K\}$. Then,

$$w_k(z_{k-1},z_k) = \frac{\gamma_k(z_k)\ell_{k-1}(z_k,z_{k-1})}{\gamma_{k-1}(z_{k-1})m_k(z_{k-1},z_k)}$$

2. Assume that for $k \in \{1, \ldots, K\}$, $\Gamma_k(\mathrm{d}z_{k-1})M_k(z_{k-1}, \mathrm{d}z_k) = \Gamma_k(\mathrm{d}z_k)L_{k-1}(z_k, \mathrm{d}z_{k-1})$, and that there exists a positive measurable function such that $\Gamma_k(\mathrm{d}z_{k-1}) = \tilde{w}_k(z_{k-1})\Gamma_{k-1}(\mathrm{d}z_{k-1})$. Then,

$$\Gamma_k(\mathrm{d}z_k)L_{k-1}(z_k, \mathrm{d}z_{k-1}) = \Gamma_k(\mathrm{d}z_{k-1})M_k(z_{k-1}, \mathrm{d}z_k) = \tilde{w}_k(z_{k-1})\Gamma_{k-1}(\mathrm{d}z_{k-1})M_k(z_{k-1}, \mathrm{d}z_k) \ .$$

   Hence, (S1) is satisfied with $w_k(z_{k-1}, z_k) = \tilde{w}_k(z_{k-1})$. In particular, if for all $k \in \{0, \ldots, K\}$, $\Gamma_k(z_k) = \gamma_k(z_k)\mathrm{d}z_k$, where γ_k is a positive p.d.f., then $\tilde{w}_k(z_k) = \gamma_k(z_k)/\gamma_{k-1}(z_{k-1})$.

3. Assume that for $k \in \{1, \ldots, K\}$, M_k is reversible w.r.t. Γ_k, *i.e.* $\Gamma_k(\mathrm{d}z_{k-1})M_k(z_{k-1}, \mathrm{d}z_k) = \Gamma_k(\mathrm{d}z_k)M_k(z_k, \mathrm{d}z_{k-1})$, and that there exists a positive measurable function such that $\Gamma_k(\mathrm{d}z_{k-1}) = \tilde{w}_k(z_{k-1})\Gamma_{k-1}(\mathrm{d}z_{k-1})$. Then, setting $L_{k-1} = M_k$, (S1) is satisfied.

3.2. Proof of (14)

For $k \in \{1, \ldots, K\}$, $z_{k-1} \in \mathbb{R}^d$, denote by $G_{k,z_{k-1}}$ the mapping $u_k \mapsto \mathrm{T}_{u_k}(z_{k-1})$. Our derivation below rely on the fact that for $k \in \{1, \ldots, K\}$, $z_{k-1} \in \mathbb{R}^d$, $G_{k,z_{k-1}}$ is a C^1-diffeomorphism. This is the case for the Langevin mappings. Note, similarly to the density considered in Section 3, that $m_k(z_{k-1}, z_k) = \varphi(G^{-1}_{k,z_{k-1}}(z_k))\mathrm{J}_{G^{-1}_{k,z_{k-1}}}(z_k)$. When $K = 1$, we have

$$\begin{aligned}\int \log\big(w_1(z_0, z_1)\big)q^1_\phi(z_{0:1} \mid x)\mathrm{d}z_{0:1} &= \int \log\big(w_1(z_0, z_1)\big)q_\phi(z_0 \mid x)\mathrm{J}_{G^{-1}_{1,z_0}}(z_1)\varphi\big(G^{-1}_{1,z_0}(z_1)\big)\mathrm{d}z_{0:1} \\ &= \int \log\big(w_1(z_0, \mathrm{T}_{1,u_1}(z_0))\big)q_\phi(z_0 \mid x)\varphi(u_1)\mathrm{d}z_0\mathrm{d}u_1 \ ,\end{aligned}$$

where we have performed the change of variables $u_1 = G^{-1}_{1,z_0}(z_1)$, hence $z_1 = G_{1,z_0}(u_1) = \mathrm{T}_{1,u_1}(z_0)$. Let now K be in $\mathbb{N}^*$. In general, we write

$$\begin{aligned}\mathcal{L}_{\mathrm{SIS}} &= \int \log\left(\prod_{k=1}^K w_k(z_{k-1}, z_k)\right) q^K_\phi(z_{0:K} \mid x)\mathrm{d}z_{0:K} = \int \log\left(\prod_{k=1}^K w_k(z_{k-1}, z_k)\right) q_\phi(z_0 \mid x) \prod_{k=1}^K m_k(z_{k-1}, z_k)\mathrm{d}z_{0:K-1}\mathrm{d}z_K \\ &= \int \log\left(\prod_{k=1}^K w_k(z_{k-1}, z_k)\right) q_\phi(z_0 \mid x) \prod_{k=1}^{K-1} m_k(z_{k-1}, z_k)\varphi(G^{-1}_{K,z_{K-1}}(z_K))\mathrm{J}_{G^{-1}_{K,z_{K-1}}}(z_K)\mathrm{d}z_{0:K-1}\mathrm{d}z_K \\ &= \int \log\left(\prod_{k=1}^{K-1} w_k(z_{k-1}, z_k)w_K(z_{K-1}, \mathrm{T}_{u_K}(z_{K-1}))q_\phi(z_0 \mid x)\right) \prod_{k=1}^{K-1} m_k(z_{k-1}, z_k)\varphi(u_K)\mathrm{d}z_{0:K-1}\mathrm{d}u_K\end{aligned}$$

using the change of variables $u_K = G^{-1}_{K,z_{K-1}}(z_K)$. By an immediate backwards induction, we write

$$\mathcal{L}_{\mathrm{SIS}} = \int \log\left(\prod_{k=1}^K w_k(\bigcirc_{i=1}^{k-1} \mathrm{T}_{i,u_i}(z_0), \bigcirc_{i=1}^{k} \mathrm{T}_{i,u_i}(z_0))\right) q_\phi(z_0 \mid x)\varphi(u_{1:K})\mathrm{d}z_0\mathrm{d}u_{1:K} \ .$$

3.3. Proof of Lemma 1

Let $\eta < \mathrm{L}^{-1}$ and $u \in \mathbb{R}^D$. First we show that $\mathrm{T}^{\mathrm{MALA}}_u$ is invertible. Consider, for each $(y, u) \in \mathbb{R}^{2d}$, the mapping $H_{y,u}(z) = y - \sqrt{2\eta}u - \eta\nabla \log \pi(z)$. We have, for $z_1, z_2 \in \mathbb{R}^d$,

$$\|H_{y,u}(z_1) - H_{y,u}(z_2)\| \leqslant \eta\|\nabla \log \pi(z_1) - \nabla \log \pi(z_2)\| \leqslant \eta\mathrm{L}\|z_1 - z_2\|$$

and $\eta\mathrm{L} < 1$. Hence $H_{y,u}$ is a contraction mapping and thus has a unique fixed point $z_{y,u}$. Hence, for all $(y, u) \in \mathbb{R}^{2d}$ there exists a unique $z_{y,u}$ satisfying

$$H_{y,u}(z_{y,u}) = z_{y,u} \Rightarrow y = z_{y,u} + \eta\nabla \log \pi(z_{y,u}) + \sqrt{2\eta}u = \mathrm{T}^{\mathrm{MALA}}_u(z_{y,u}).$$

This establishes the invertibility of $\mathrm{T}^{\mathrm{MALA}}_u$. The fact that the inverse of $\mathrm{T}^{\mathrm{MALA}}_u$ is C^1 follows from a simple application of the local inverse function theorem.

4. ELBO AIS

4.1. Construction of the control variates

We prove in this section that the variance reduced objective we consider is valid. Sample now n samples $u_{0:K}^{1:n} \overset{\text{i.i.d.}}{\sim} \varphi_{d,K+1}$. For an index $i \in \{1, \ldots, n\}$, given the initial point $z_0^i = V_{\phi,x}(u_0^i)$ and the innovation noise $u_{1:K}^i$, we sample the A/R booleans $a_{1:K}^i$. We introduce, in the main text, for $i \in \{1, \ldots, n\}$

$$\tilde{W}_{n,i} = \frac{1}{n-1} \sum_{j \neq i} W(V_{\phi,x}(u_0^j), a_{1:K}^j, u_{1:K}^j) \ ,$$

$\tilde{W}_i$ provides a reasonable estimate of the AIS ELBO but is independent from the i-th trajectory. We use this quantity as a control variate to reduce the variance of our gradient estimator by introducing

$$\begin{aligned} \widehat{\nabla \mathcal{L}_{\text{AIS}}}_n = n^{-1} & \sum_{i=1}^{n} \nabla W(V_{\phi,x}(u_0^i), a_{1:K}^i, u_{1:K}^i) \\ &+ n^{-1} \sum_{i=1}^{n} \left[W(V_{\phi,x}(u_0^i), a_{1:K}^i, u_{1:K}^i) - \tilde{W}_{n,i} \right] \\ &\times \nabla \log A(V_{\phi,x}(u_0^i), a_{1:K}^i, u_{1:K}^i) \ . \end{aligned} \tag{S4}$$

Proving its unbiasedness boils down to proving that the term $n^{-1} \sum_{i=1}^{n} \tilde{W}_{n,i} \nabla \log A(V_{\phi,x}(u_0^i), a_{1:K}^i, u_{1:K}^i)$ has expectation zero. Let us compute for $i \in \{1 \ldots, n\}$,

$$\int \sum_{a_{1:K}^i} \varphi_{d,K+1}(u_{0:K}^i) A(V_{\phi,x}(u_0^i), a_{1:K}^i, u_{1:K}^i) \tilde{W}_{n,i} \nabla \log A(V_{\phi,x}(u_0^i), a_{1:K}^i, u_{1:K}^i) \mathrm{d}u_{0:K}^i =$$

$$\int \sum_{a_{1:K-1}^i} \varphi_{d,K+1}(u_{0:K}^i) \prod_{k=1}^{K} \alpha_{k,u_k^i}^{a_k^i}(z_{k-1}^i) \tilde{W}_{n,i} \left[\nabla \sum_{k=1}^{K-1} \log \alpha_{k,u_k^i}^{a_k^i}(z_{k-1}^i) + \sum_{a_K^i} \nabla \log \alpha_{K,u_K^i}^{a_K^i}(z_{K-1}^i) \right] \mathrm{d}u_{0:K}^i \ ,$$

denoting $z_0^j = V_{\phi,x}(u_0^j)$, $z_k^j = \bigcirc_{i=1}^{k} \mathrm{T}_{i,u_i^j}^{a_i^j}(z_0^j)$ by simplicity of notation. Yet, $\sum_{a_K^i} \alpha_{K,u_K^i}^{a_K^i}(z_{K-1}^i) = 1$ exactly, thus $\sum_{a_K^i} \alpha_{K,u_K^i}^{a_K^i}(z_{K-1}^l) \nabla \log \alpha_{K,u_K^i}^{a_K^i}(z_{K-1}^i) = 0$. We can thus show by an immediate induction that $\int \sum_{a_{1:K}^i} \varphi_d(u_{0:K}^i) \tilde{W}_{n,i} \nabla \log A(V_{\phi,x}(u_0^i), a_{1:K}^i, u_{1:K}^i) \mathrm{d}u_{i:K}^i = 0$, as $\tilde{W}_{n,i}$ is a constant in that integral by independence of the samples for $i \in \{1 \ldots, n\}$. Moreover, as

$$\int \sum_{a_{1:K}^{1:n}} \sum_{i=1}^{n} \tilde{W}_{n,i} \nabla \log A(V_{\phi,x}(u_0^i), a_{1:K}^i, u_{1:K}^i) \prod_{\ell=1}^{n} \varphi_{d,K+1}(u_{0:K}^{\ell}) \mathrm{d}u_{0:K}^{1:n} =$$

$$\int \sum_{i=1}^{n} \sum_{a_{1:K}^{-i}} \left[\int \sum_{a_{1:K}^i} \tilde{W}_{n,i} \nabla \log A(V_{\phi,x}(u_0^i), a_{1:K}^i, u_{1:K}^i) \varphi_{d,K+1}(u_{0:K}^i) \mathrm{d}u_{0:K}^i \right] \prod_{\ell \neq i} \varphi_{d,K+1}(u_{0:K}^{\ell}) \mathrm{d}u_{1:K}^{-i} \ ,$$

then $n^{-1} \sum_{i=1}^{n} \tilde{W}_{n,i} \nabla \log A(V_{\phi,x}(u_0^i), a_{1:K}^i, u_{1:K}^i)$ is of zero expectation, and (S4) is an unbiased estimator of the gradient.

4.2. Discussion of (Wu et al., 2020)

In (Wu et al., 2020), authors consider a MCMC VAE inspired by AIS. The model used however is quite different in spirit to what is performed in this work. (Wu et al., 2020) use Langevin mappings and accept reject steps in their VAE. Note however that the A/R probabilities defined are written as

$$\alpha(x, y) = 1 \wedge \pi(y)/\pi(x) \ ,$$

different from (21). Moreover, even though accept/reject steps are considered, the score function estimator (25) is not taken into account.

Finally, the initial density of the sequence is not taken to be some variational mean field initialization but directly the prior in the latent space. As a result, the scores obtained by the MCMC VAE are less competitive than that of the RNVP VAE presented in (Wu et al., 2020, Table 3.), contrary to what is presented here.

References

Huang, C.-W., Tan, S., Lacoste, A., and Courville, A. C. Improving explorability in variational inference with annealed variational objectives. In *Advances in Neural Information Processing Systems*, pp. 9701–9711, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Salimans, T., Kingma, D., and Welling, M. Markov chain Monte Carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pp. 1218–1226, 2015.

Wu, H., Köhler, J., and Noé, F. Stochastic normalizing flows. *Advances in Neural Information Processing Systems*, 2020.