DeepMind

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

Francisco J. R. Ruiz, Michalis K. Titsias, Taylan Cemgil, Arnaud Doucet

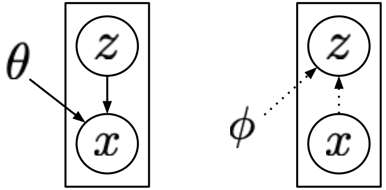
24/11/2020

DeepMind

Motivation and Goal

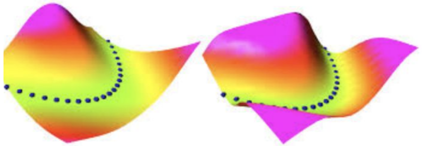
Deep Generative Models

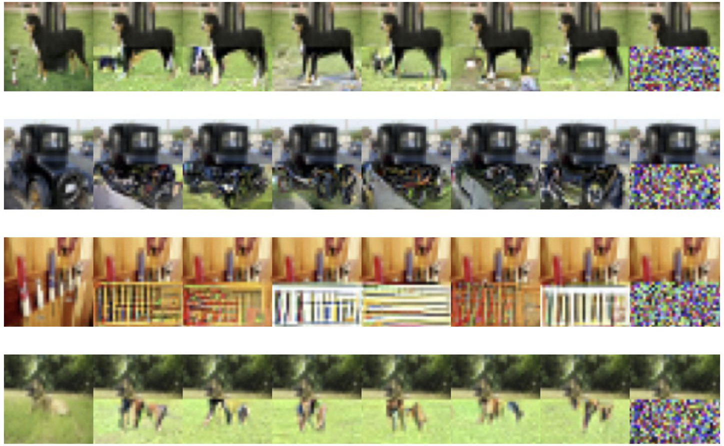
Variational Auto-Encoder (VAE)



[Anonymous, ICLR 2021]

Energy-Based Model (EBM)





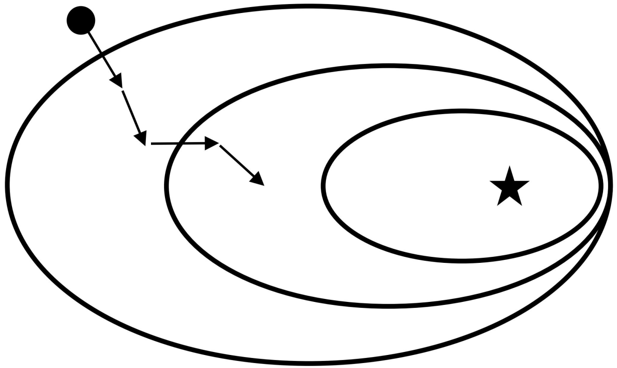
[OpenAI]

Fitting Deep Generative Models

- Stochastic gradient descent (SGD) is a **powerful tool** in Machine Learning

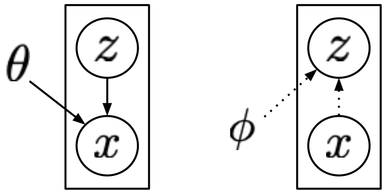
- SGD obtains and follows **unbiased estimates** of the log-likelihood gradient

- For some deep generative models, unbiased gradient estimates are **not directly available**

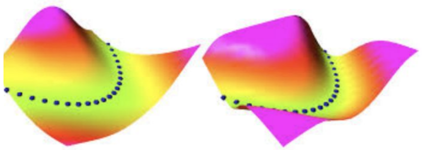


Why Fitting These Models is Hard

VAE



EBM



The model's likelihood is intractable

log-likelihood *intractable integral*

$$\mathcal{L}(\theta) := \log p_\theta(x) = \log\left(\int p_\theta(x, z)dz\right)$$

likelihood *intractable integral*

$$p_\theta(x) = \frac{\exp(E_\theta(x))}{Z_\theta}, \quad Z_\theta = \int \exp(E_\theta(x))\mathrm{d}x$$

Its gradient is intractable

gradient *an expectation*

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

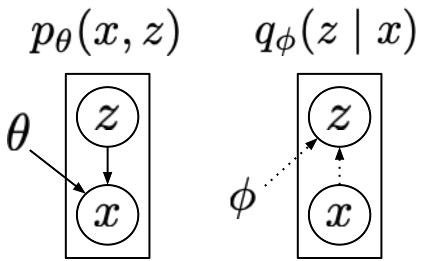
gradient *an expectation*

$$\nabla_\theta \log p_\theta(x) = \nabla_\theta E_\theta(x) - \mathbb{E}_{p_\theta(x')}\left[\nabla_\theta E_\theta(x')\right]$$

(but it can be written in terms of an expectation)

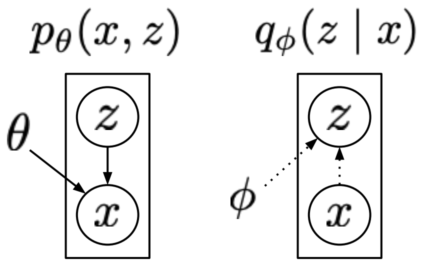
Summary of Contributions

- An algorithm to obtain **unbiased gradient estimates** for VAEs
  - **Avoid approximation gap** of previous approaches (based on bounds)
  - Lead to VAEs with **better predictive performance**
  - Applicable to other deep generative models **beyond VAEs**
  - Main **limitation**: increased computational complexity



Technical Tools

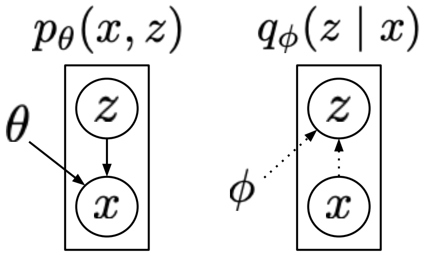
- Two main ideas to develop the unbiased gradient estimates:
  - **Augmented latent space** (similarly to IWAE)
  - **MCMC couplings** based on importance sampling



DeepMind

Preliminaries

Review: VAE / IWAE



- The VAE log-likelihood

log-likelihood (intractable)

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

- Optimize the ELBO (a lower bound)

importance weights

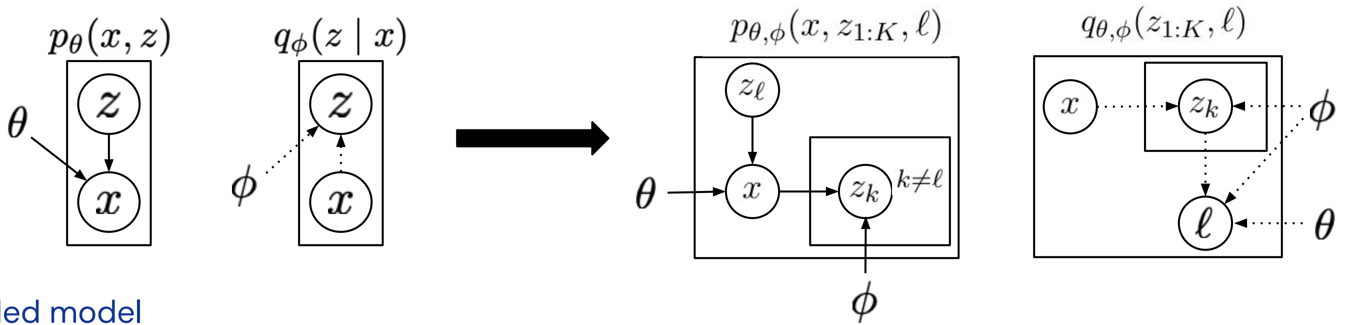
$$\mathcal{L}_{\text{ELBO}}(\theta, \phi) = \mathbb{E}_{q_\phi(z \mid x)}\left[\log w_{\theta,\phi}(z)\right] \qquad w_{\theta,\phi}(z) := \frac{p_\theta(x, z)}{q_\phi(z \mid x)}$$

- Or optimize the IWAE (a tighter lower bound)

auxiliary random variables

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

Background: The IWAE as an Augmented Space



- Extended model

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

- Extended variational distribution

$$q_{\theta,\phi}(z_{1:K}, \ell) = \text{Categorical}\left(\ell \mid \widetilde{w}_{\theta,\phi}^{(1)}, \ldots, \widetilde{w}_{\theta,\phi}^{(K)}\right) \prod_{k=1}^{K} q_\phi(z_k \mid x)$$

normalized importance weights

$$w_{\theta,\phi}^{(k)} = w_{\theta,\phi}(z_k), \quad \widetilde{w}_{\theta,\phi}^{(k)} = \frac{w_{\theta,\phi}^{(k)}}{\sum_{k'=1}^{K} w_{\theta,\phi}^{(k')}}$$

- Its ELBO coincides with the IWAE bound

$$\mathcal{L}_{\text{ELBO}}^{\text{augmented}}(\theta, \phi) = \mathbb{E}_{q_{\theta,\phi}(z_{1:K}, \ell)}\left[\log \frac{p_{\theta,\phi}(x, z_{1:K}, \ell)}{q_{\theta,\phi}(z_{1:K}, \ell)}\right] = \mathbb{E}_{q_{\theta,\phi}(z_{1:K})}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K} w_{\theta,\phi}^{(k)}\right)\right]$$

The Roadmap to Unbiased Estimation

- Both the ELBO and IWAE bounds are **biased** approximators of gradient of the log-likelihood

- If we could **sample** from the posterior, we could easily form an unbiased estimator

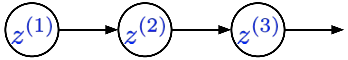
  gradient

  expectation

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

  (but we cannot sample in practice)

- **MCMC couplings** provide unbiased estimators by design without the need to obtain exact samples from the posterior



MCMC Couplings

$$u^{(1)} \to u^{(2)} \to u^{(3)} \to$$

- Consider estimating an expectation of the form

generic r.v.

$$H = \mathbb{E}_{\pi(u)}[h(u)]$$

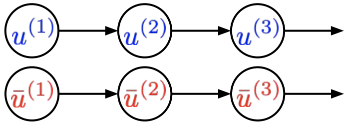
- Typical MCMC approach: sample from a kernel $\mathcal{K}(\cdot \mid u)$ that targets $\pi(u)$

MCMC Couplings

- Consider estimating an expectation of the form

generic r.v.

$$H = \mathbb{E}_{\pi(u)}[h(u)]$$



- Typical MCMC approach: sample from a kernel $\mathcal{K}(\cdot \mid u)$ that targets $\pi(u)$
- Coupling MCMC: use **two** MCMC chains with the same stationary distribution that are *coupled*
  - There is a joint MCMC kernel $\mathcal{K}_C(\cdot, \cdot \mid u, \bar{u})$

MCMC Couplings

- Consider estimating an expectation of the form

generic r.v.

$$H = \mathbb{E}_{\pi(u)}[h(u)]$$

$$\begin{tikzcd} u^{(1)} \arrow[r] & u^{(2)} \arrow[r] & u^{(3)} \arrow[r] & {} \\ \bar{u}^{(1)} \arrow[r] & \bar{u}^{(2)} \arrow[r] & \bar{u}^{(3)} \arrow[r] & {} \end{tikzcd}$$

- Typical MCMC approach: sample from a kernel $\mathcal{K}(\cdot \mid u)$ that targets $\pi(u)$
- Coupling MCMC: use **two** MCMC chains with the same stationary distribution that are *coupled*
  - There is a joint MCMC kernel $\mathcal{K}_C(\cdot, \cdot \mid u, \bar{u})$
- An unbiased estimator is

finite sum

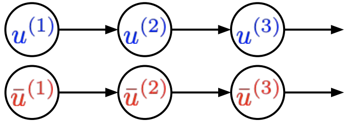
estimates from chain #1

estimates from chain #2

$$H \approx h(u^{(t_0)}) + \sum_{t=t_0+1}^{\tau-1} \left(h(u^{(t)}) - h(\bar{u}^{(t-1)}) \right)$$

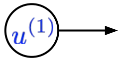
MCMC Couplings: Sampling

- Coupling MCMC: use **two** MCMC chains with the same stationary distribution that are *coupled*
  - Marginally evolving according to $\mathcal{K}(\cdot \mid u)$
  - There is a joint MCMC kernel $\mathcal{K}_C(\cdot, \cdot \mid u, \bar{u})$

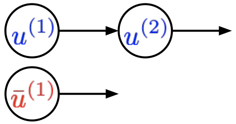


MCMC Couplings: Sampling

- Coupling MCMC: use **two** MCMC chains with the same stationary distribution that are *coupled*
  - Marginally evolving according to $\mathcal{K}(\cdot \mid u)$
  - There is a joint MCMC kernel $\mathcal{K}_C(\cdot, \cdot \mid u, \bar{u})$
  - Initialize the first chain $u^{(1)} \sim \mathcal{K}(u \mid u^{(0)})$

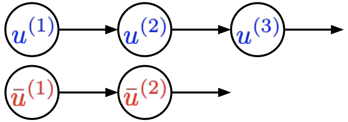


MCMC Couplings: Sampling



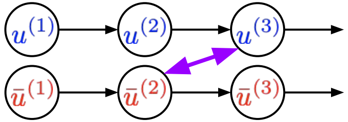
- Coupling MCMC: use **two** MCMC chains with the same stationary distribution that are *coupled*
  - Marginally evolving according to $\mathcal{K}(\cdot \mid u)$
  - There is a joint MCMC kernel $\mathcal{K}_C(\cdot, \cdot \mid u, \bar{u})$
  - Initialize the first chain $u^{(1)} \sim \mathcal{K}(u \mid u^{(0)})$
  - Then sample both chains from the joint kernel $u^{(t+1)}, \bar{u}^{(t)} \sim \mathcal{K}_C(u, \bar{u} \mid u^{(t)}, \bar{u}^{(t-1)})$

MCMC Couplings: Sampling



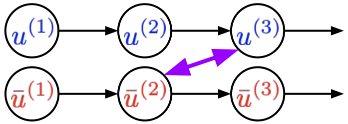
- Coupling MCMC: use **two** MCMC chains with the same stationary distribution that are *coupled*
  - Marginally evolving according to $\mathcal{K}(\cdot \mid u)$
  - There is a joint MCMC kernel $\mathcal{K}_C(\cdot, \cdot \mid u, \bar{u})$
  - Initialize the first chain $u^{(1)} \sim \mathcal{K}(u \mid u^{(0)})$
  - Then sample both chains from the joint kernel $u^{(t+1)}, \bar{u}^{(t)} \sim \mathcal{K}_C(u, \bar{u} \mid u^{(t)}, \bar{u}^{(t-1)})$

MCMC Couplings: Sampling



- Coupling MCMC: use **two** MCMC chains with the same stationary distribution that are *coupled*
  - Marginally evolving according to $\mathcal{K}(\cdot \mid u)$
  - There is a joint MCMC kernel $\mathcal{K}_C(\cdot, \cdot \mid u, \bar{u})$
  - Initialize the first chain $u^{(1)} \sim \mathcal{K}(u \mid u^{(0)})$
  - Then sample both chains from the joint kernel $u^{(t+1)}, \bar{u}^{(t)} \sim \mathcal{K}_C(u, \bar{u} \mid u^{(t)}, \bar{u}^{(t-1)})$

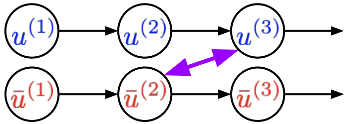
MCMC Couplings: Sampling



- Coupling MCMC: use **two** MCMC chains with the same stationary distribution that are *coupled*
  - Marginally evolving according to $\mathcal{K}(\cdot \mid u)$
  - There is a joint MCMC kernel $\mathcal{K}_C(\cdot, \cdot \mid u, \bar{u})$
  - Initialize the first chain $u^{(1)} \sim \mathcal{K}(u \mid u^{(0)})$
  - Then sample both chains from the joint kernel $u^{(t+1)}, \bar{u}^{(t)} \sim \mathcal{K}_C(u, \bar{u} \mid u^{(t)}, \bar{u}^{(t-1)})$

- Define the **meeting time** $\tau = \inf\{t \geq 1 : u^{(t)} = \bar{u}^{(t-1)}\}$
  - We design the kernel such that: (i) the two chains **meet each other** (τ is random but finite), and (ii) they **remain equal** to each other afterwards

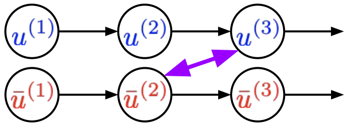
MCMC Couplings: Unbiased Estimation



- Use the samples from both chains to form an estimator

$$H \approx h(u^{(t_0)}) + \sum_{t=t_0+1}^{\tau-1} \left(h(u^{(t)}) - h(\bar{u}^{(t-1)}) \right) \qquad \tau = \inf\{t \geq 1 : u^{(t)} = \bar{u}^{(t-1)}\}$$

MCMC Couplings: Unbiased Estimation



- Use the samples from both chains to form an estimator

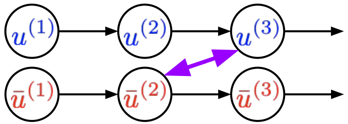
$$H \approx h(u^{(t_0)}) + \sum_{t=t_0+1}^{\tau-1} \left(h(u^{(t)}) - h(\bar{u}^{(t-1)}) \right) \qquad \tau = \inf\{t \geq 1 : u^{(t)} = \bar{u}^{(t-1)}\}$$

- Proof

$$H = \mathbb{E}_{\pi(u)}[h(u)] = \mathbb{E}[h(u^{(t_0)})] + \sum_{t=t_0+1}^{\infty} \left(\mathbb{E}[h(u^{(t)})] - \mathbb{E}[h(u^{(t-1)})] \right)$$

telescoping sum

MCMC Couplings: Unbiased Estimation



- Use the samples from both chains to form an estimator

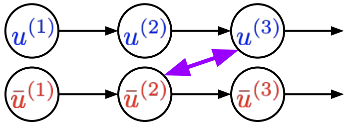
$$H \approx h(u^{(t_0)}) + \sum_{t=t_0+1}^{\tau-1} \left(h(u^{(t)}) - h(\bar{u}^{(t-1)}) \right) \qquad \tau = \inf\{t \geq 1 : u^{(t)} = \bar{u}^{(t-1)}\}$$

- Proof

$$H = \mathbb{E}_{\pi(u)}[h(u)] = \mathbb{E}[h(u^{(t_0)})] + \sum_{t=t_0+1}^{\infty} \left(\mathbb{E}[h(u^{(t)})] - \mathbb{E}[h(u^{(t-1)})] \right) \qquad \textit{telescoping sum}$$

$$= \mathbb{E}[h(u^{(t_0)})] + \sum_{t=t_0+1}^{\infty} \left(\mathbb{E}[h(u^{(t)})] - \mathbb{E}[h(\bar{u}^{(t-1)})] \right) \qquad \textit{same marginals}$$

MCMC Couplings: Unbiased Estimation



- Use the samples from both chains to form an estimator

$$H \approx h(u^{(t_0)}) + \sum_{t=t_0+1}^{\tau-1} \left(h(u^{(t)}) - h(\bar{u}^{(t-1)}) \right) \qquad \tau = \inf\{t \geq 1 : u^{(t)} = \bar{u}^{(t-1)}\}$$

- Proof

$$H = \mathbb{E}_{\pi(u)}[h(u)] = \mathbb{E}[h(u^{(t_0)})] + \sum_{t=t_0+1}^{\infty} \left(\mathbb{E}[h(u^{(t)})] - \mathbb{E}[h(u^{(t-1)})] \right)$$

telescoping sum

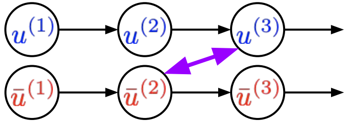
$$= \mathbb{E}[h(u^{(t_0)})] + \sum_{t=t_0+1}^{\infty} \left(\mathbb{E}[h(u^{(t)})] - \mathbb{E}[h(\bar{u}^{(t-1)})] \right)$$

same marginals

$$= \mathbb{E}\left[h(u^{(t_0)}) + \sum_{t=t_0+1}^{\infty} \left(h(u^{(t)}) - h(\bar{u}^{(t-1)}) \right) \right]$$

swap expectation and limit

MCMC Couplings: Unbiased Estimation



- Use the samples from both chains to form an estimator

$$H \approx h(u^{(t_0)}) + \sum_{t=t_0+1}^{\tau-1} \left(h(u^{(t)}) - h(\bar{u}^{(t-1)}) \right) \qquad \tau = \inf\{t \geq 1 : u^{(t)} = \bar{u}^{(t-1)}\}$$

- Proof

$$H = \mathbb{E}_{\pi(u)}[h(u)] = \mathbb{E}[h(u^{(t_0)})] + \sum_{t=t_0+1}^{\infty} \left(\mathbb{E}[h(u^{(t)})] - \mathbb{E}[h(u^{(t-1)})] \right)$$

telescoping sum

$$= \mathbb{E}[h(u^{(t_0)})] + \sum_{t=t_0+1}^{\infty} \left(\mathbb{E}[h(u^{(t)})] - \mathbb{E}[h(\bar{u}^{(t-1)})] \right)$$

same marginals

$$= \mathbb{E}\left[h(u^{(t_0)}) + \sum_{t=t_0+1}^{\infty} \left(h(u^{(t)}) - h(\bar{u}^{(t-1)}) \right) \right]$$

swap expectation and limit

$$= \mathbb{E}\left[h(u^{(t_0)}) + \sum_{t=t_0+1}^{\tau-1} \left(h(u^{(t)}) - h(\bar{u}^{(t-1)}) \right) + \sum_{t=\tau}^{\infty} \left(\cancel{h(u^{(t)})} - \cancel{h(\bar{u}^{(t-1)})} \right) \right]$$

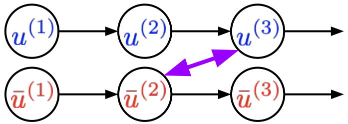
by design

DeepMind

Unbiased Estimators on an Extended Space

Our Proposal

- Start with $\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{p_\theta(z \mid x)} \left[\nabla_\theta \log p_\theta(x, z) \right]$
- Form the augmented model and augmented proposal



$$u = [z_{1:K}, \ell]$$

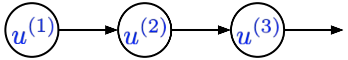
$$\bar{u} = [\bar{z}_{1:K}, \bar{\ell}]$$

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

$$q_{\theta,\phi}(z_{1:K}, \ell) = \mathrm{Categorical}\left(\ell \mid \widetilde{w}^{(1)}_{\theta,\phi}, \ldots, \widetilde{w}^{(K)}_{\theta,\phi}\right) \prod_{k=1}^{K} q_\phi(z_k \mid x)$$

- Run a coupled MCMC kernel on the extended space, targeting the augmented posterior
  - How to form the kernel?

PIMH Algorithm (Non-Coupled Version)



Algorithm 1: Particle independent Metropolis-Hastings (PIMH) kernel, $\mathcal{K}_{\text{PIMH}}(\cdot, \cdot \mid z_{1:K}, \ell)$

Input: Current state of the chain, $(z_{1:K}, \ell)$

Output: Next state of the chain

Sample a candidate $(z^{\star}_{1:K}, \ell^{\star}) \sim q_{\theta,\phi}(\cdot, \cdot)$

Sample $u \sim \mathcal{U}([0, 1])$

if $u \leq \alpha(z^{\star}_{1:K}, \ell^{\star} \mid z_{1:K}, \ell)$ **then**

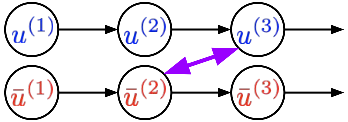
  Return $(z^{\star}_{1:K}, \ell^{\star})$ ▷ `the proposal is accepted`

else

  Return $(z_{1:K}, \ell)$ ▷ `the proposal is rejected`

end

PIMH Algorithm (Coupled Version)



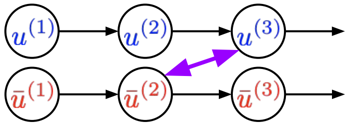
Algorithm 4: Coupled PIMH kernel, $\mathcal{K}_{\text{C-PIMH}}((\cdot,\cdot),(\cdot,\cdot) \mid (z_{1:K},\ell),(\bar{z}_{1:K},\bar{\ell}))$

Input: Current state of both chains, $(z_{1:K},\ell)$ and $(\bar{z}_{1:K},\bar{\ell})$

Output: New state of both chains

Sample $(z^\star_{1:K},\ell^\star) \sim q_{\theta,\phi}(\cdot,\cdot)$
Sample $u \sim \mathcal{U}([0,1])$
if $u \leq \alpha(z^\star_{1:K},\ell^\star \mid z_{1:K},\ell)$ and $u \leq \alpha(z^\star_{1:K},\ell^\star \mid \bar{z}_{1:K},\bar{\ell})$ **then**
  Return $((z^\star_{1:K},\ell^\star),(z^\star_{1:K},\ell^\star))$ ▷ both chains accept the proposal
else if $u \leq \alpha(z^\star_{1:K},\ell^\star \mid z_{1:K},\ell)$ and $u > \alpha(z^\star_{1:K},\ell^\star \mid \bar{z}_{1:K},\bar{\ell})$ **then**
  Return $((z^\star_{1:K},\ell^\star),(\bar{z}_{1:K},\bar{\ell}))$ ▷ the first chain accepts the proposal
else if $u > \alpha(z^\star_{1:K},\ell^\star \mid z_{1:K},\ell)$ and $u \leq \alpha(z^\star_{1:K},\ell^\star \mid \bar{z}_{1:K},\bar{\ell})$ **then**
  Return $((z_{1:K},\ell),(z^\star_{1:K},\ell^\star))$ ▷ the second chain accepts the proposal
else
  Return $((z_{1:K},\ell),(\bar{z}_{1:K},\bar{\ell}))$ ▷ neither chain accepts the proposal
end

PIMH Algorithm (Coupled Version)



Algorithm 4: Coupled PIMH kernel, $\mathcal{K}_{\text{C-PIMH}}((\cdot,\cdot),(\cdot,\cdot) \mid (z_{1:K},\ell),(\bar{z}_{1:K},\bar{\ell}))$

Input: Current state of both chains, $(z_{1:K},\ell)$ and $(\bar{z}_{1:K},\bar{\ell})$

Output: New state of both chains

Sample $(z^\star_{1:K},\ell^\star) \sim q_{\theta,\phi}(\cdot,\cdot)$

Sample $u \sim \mathcal{U}([0,1])$

if $u \leq \alpha(z^\star_{1:K},\ell^\star \mid z_{1:K},\ell)$ and $u \leq \alpha(z^\star_{1:K},\ell^\star \mid \bar{z}_{1:K},\bar{\ell})$ **then**

  Return $((z^\star_{1:K},\ell^\star),(z^\star_{1:K},\ell^\star))$ ▷ both chains accept the proposal

else if $u \leq \alpha(z^\star_{1:K},\ell^\star \mid z_{1:K},\ell)$ and $u > \alpha(z^\star_{1:K},\ell^\star \mid \bar{z}_{1:K},\bar{\ell})$ **then**

  Return $((z^\star_{1:K},\ell^\star),(\bar{z}_{1:K},\bar{\ell}))$ ▷ the first chain accepts the proposal

else if $u > \alpha(z^\star_{1:K},\ell^\star \mid z_{1:K},\ell)$ and $u \leq \alpha(z^\star_{1:K},\ell^\star \mid \bar{z}_{1:K},\bar{\ell})$ **then**

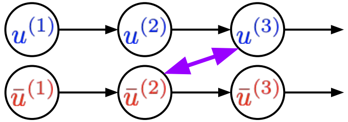
  Return $((z_{1:K},\ell),(z^\star_{1:K},\ell^\star))$ ▷ the second chain accepts the proposal

else

  Return $((z_{1:K},\ell),(\bar{z}_{1:K},\bar{\ell}))$ ▷ neither chain accepts the proposal

end

PIMH Algorithm (Coupled Version)



Algorithm 4: Coupled PIMH kernel, $\mathcal{K}_{\text{C-PIMH}}((\cdot,\cdot),(\cdot,\cdot) \mid (z_{1:K},\ell),(\bar{z}_{1:K},\bar{\ell}))$

Input: Current state of both chains, $(z_{1:K},\ell)$ and $(\bar{z}_{1:K},\bar{\ell})$

Output: New state of both chains

Sample $(z^\star_{1:K},\ell^\star) \sim q_{\theta,\phi}(\cdot,\cdot)$

Sample $u \sim \mathcal{U}([0,1])$

if $u \leq \alpha(z^\star_{1:K},\ell^\star \mid z_{1:K},\ell)$ and $u \leq \alpha(z^\star_{1:K},\ell^\star \mid \bar{z}_{1:K},\bar{\ell})$ **then**

  Return $((z^\star_{1:K},\ell^\star),(z^\star_{1:K},\ell^\star))$ ▷ both chains accept the proposal

else if $u \leq \alpha(z^\star_{1:K},\ell^\star \mid z_{1:K},\ell)$ and $u > \alpha(z^\star_{1:K},\ell^\star \mid \bar{z}_{1:K},\bar{\ell})$ **then**

  Return $((z^\star_{1:K},\ell^\star),(\bar{z}_{1:K},\bar{\ell}))$ ▷ the first chain accepts the proposal

else if $u > \alpha(z^\star_{1:K},\ell^\star \mid z_{1:K},\ell)$ and $u \leq \alpha(z^\star_{1:K},\ell^\star \mid \bar{z}_{1:K},\bar{\ell})$ **then**

  Return $((z_{1:K},\ell),(z^\star_{1:K},\ell^\star))$ ▷ the second chain accepts the proposal

else

  Return $((z_{1:K},\ell),(\bar{z}_{1:K},\bar{\ell}))$ ▷ neither chain accepts the proposal

end

PIMH Algorithm (Coupled Version)

- After collecting samples, obtain the unbiased gradient estimator as

$$\nabla_\theta \log p_\theta(x) \approx h(z_{1:K}^{(t_0)}) + \sum_{t=t_0+1}^{\tau-1} \left(h(z_{1:K}^{(t)}) - h(\bar{z}_{1:K}^{(t-1)}) \right)$$

The function h is

$$h(z_{1:K}) = \sum_{k=1}^{K} \tilde{w}_{\theta,\phi}^{(k)} \nabla_\theta \log p_\theta(x, z_k)$$

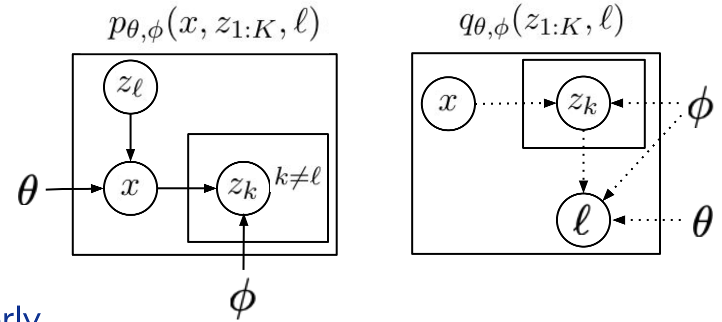
Our Contributions

- Our MCMC algorithm is **based on ISIR** (rather than PIMH)
- We propose an extension of ISIR, called DISIR, that significantly **reduces the estimator variance**
- We derive **sufficient conditions** that guarantee a finite-variance unbiased estimator in finite time
- Our estimator is based on a **lagged coupling estimator**, which further reduces the variance

Importance Sampling in High-Dimensional Spaces

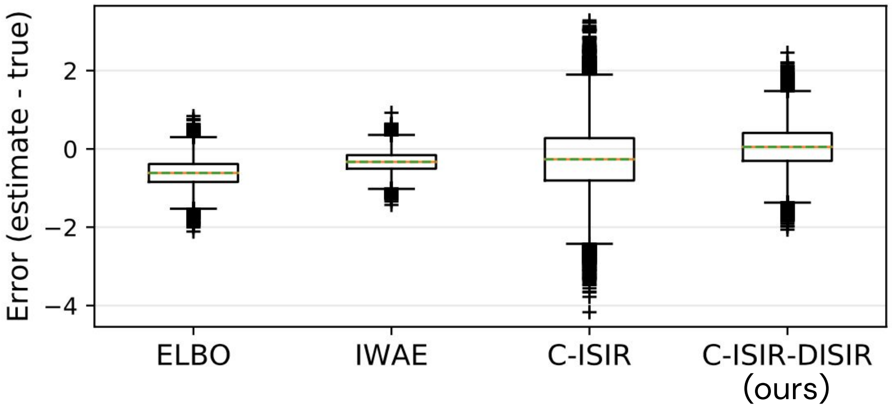
- IS typically fails in high dimensions, when one weight dominates the others

- We augment the dimensionality with *K-1* particles
  - So an IS-based MCMC algorithm should perform poorly (and the MCMC chains would never meet)
  - However, **performance improves with dimensionality** (and meeting occurs earlier) as the model and proposals become closer to each other when *K* increases



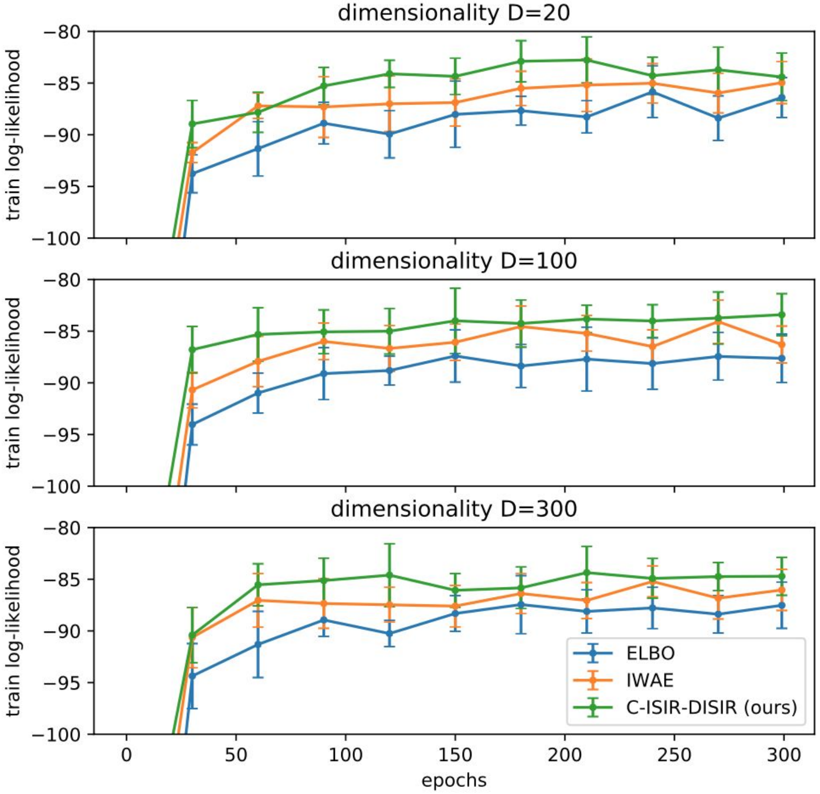
Experiments and Results

PPCA: Analysis of Unbiasedness



VAE on Binarized MNIST

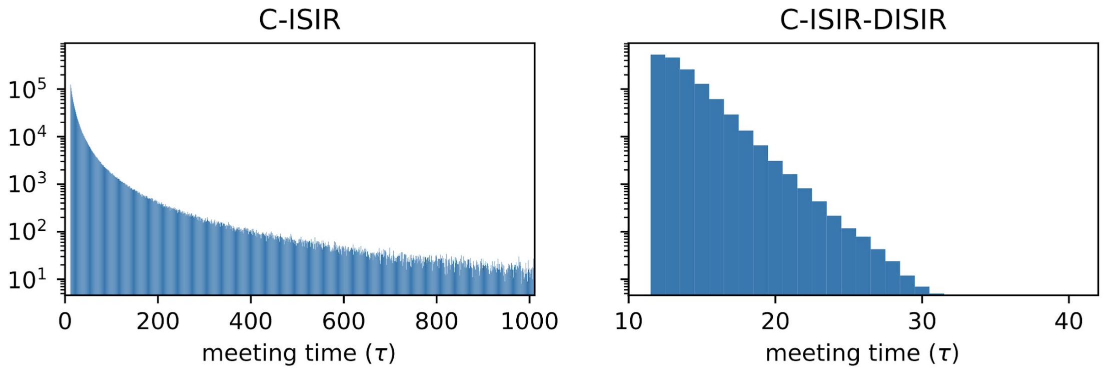
train log-likelihood



test log-likelihood

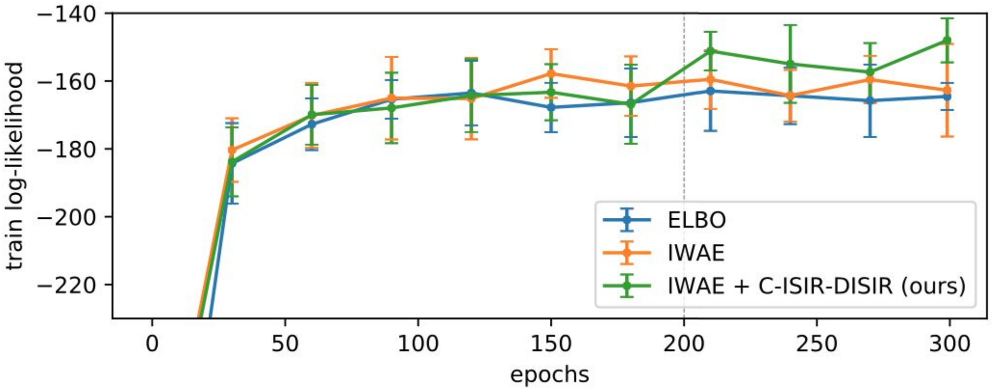
	dimensionality of z 20	100	300
ELBO	-90.05 ± 0.21	-89.96 ± 0.14	-90.63 ± 0.12
IWAE	-88.06 ± 0.08	-88.07 ± 0.06	-89.05 ± 0.08
C-ISIR-DISIR	$\mathbf{-87.29 \pm 0.08}$	$\mathbf{-86.75 \pm 0.10}$	$\mathbf{-88.10 \pm 0.08}$

Analysis of the Meeting Time



VAE on Fashion-MNIST and CIFAR-10

train log-likelihood (fashion-MNIST)



test log-likelihood

	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	$\mathbf{-168.19 \pm 0.32}$	$\mathbf{-148.40 \pm 0.27}$

Conclusions

- The combination of latent space augmentation and coupling estimators gives **practical unbiased gradients**
- Unbiased gradient estimation **improves the model's predictive performance** for VAEs
- The **computational time is higher**, but we can use this method to refine model fits
- Future work on improving coupling estimators will also reduce the computational complexity