Institut
des algorithmes
d'apprentissage
de Montréal

MILA

Variational Autoencoders

Chin-Wei Huang

Autoencoders as Generative Models

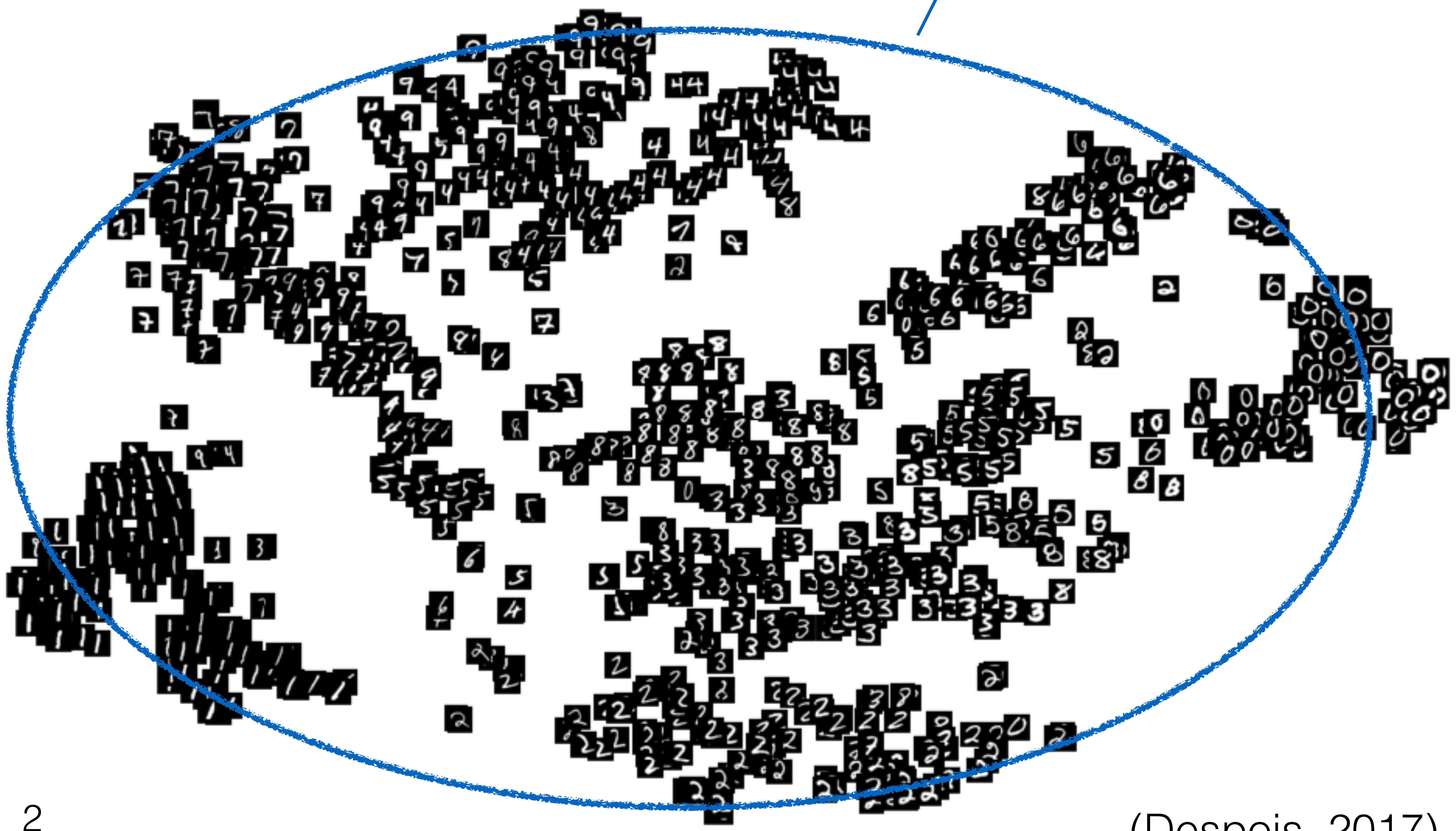
- Regularized autoencoder's objective:

$$\sum_{i=1}^{N} \underbrace{l(h_\theta(g_\phi(x_i)), x_i)}_{\text{reconstruction term}} + \lambda \overbrace{\Omega(\phi, \theta, g(x_i))}^{\text{regulerization term}}$$

code

parameters

$$\underline{z \sim p(z)}; \quad x = h_\theta(z)$$



(Despois, 2017)

Overview

- **Background**
- **Variational Inference and VAE**
- **Better model assumption**
- **Better inference**

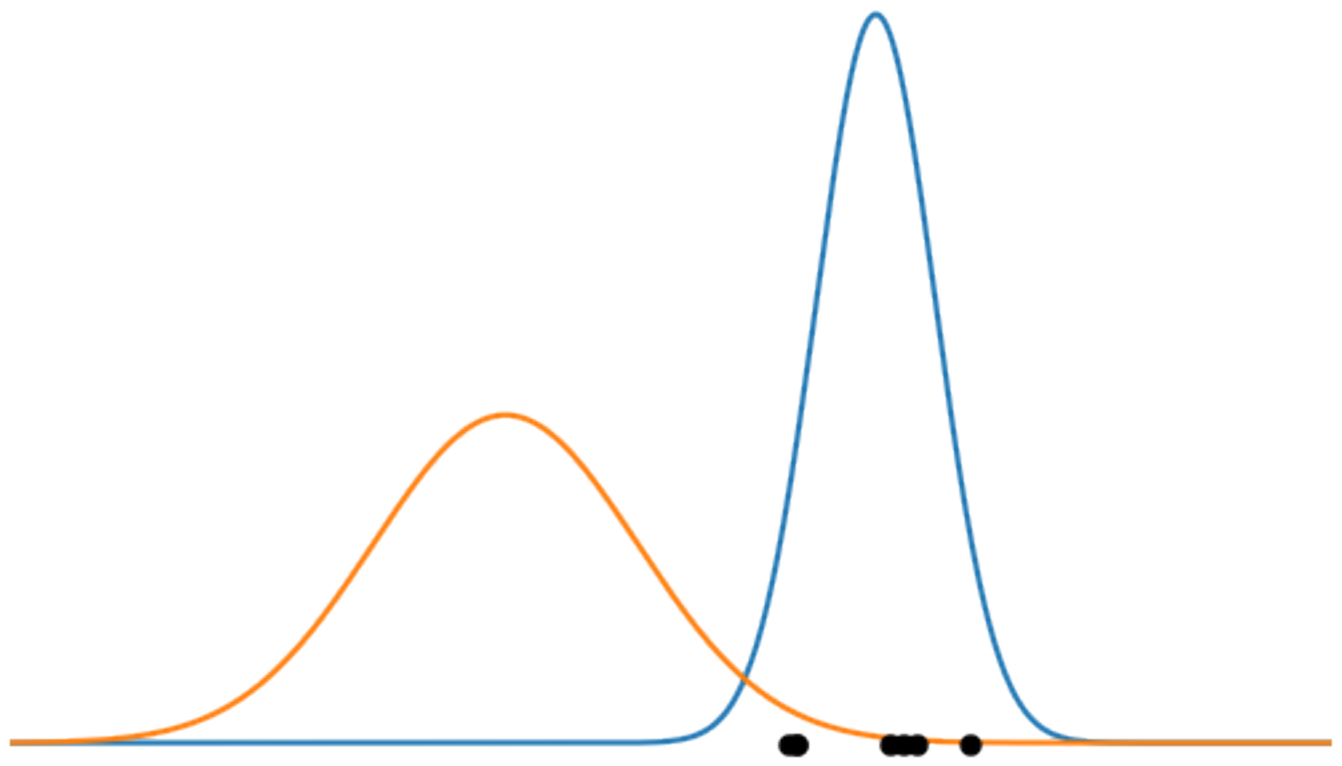
- Recap: MLE, maximum marginal likelihood, EM
- Approximate inference
- Derivation of ELBO
- Amortized variational inference
- Stochastic variational inference
- What's a good prior p(z)?
- What's a good likelihood p(x|z)?
- How to improve q(z|x)?

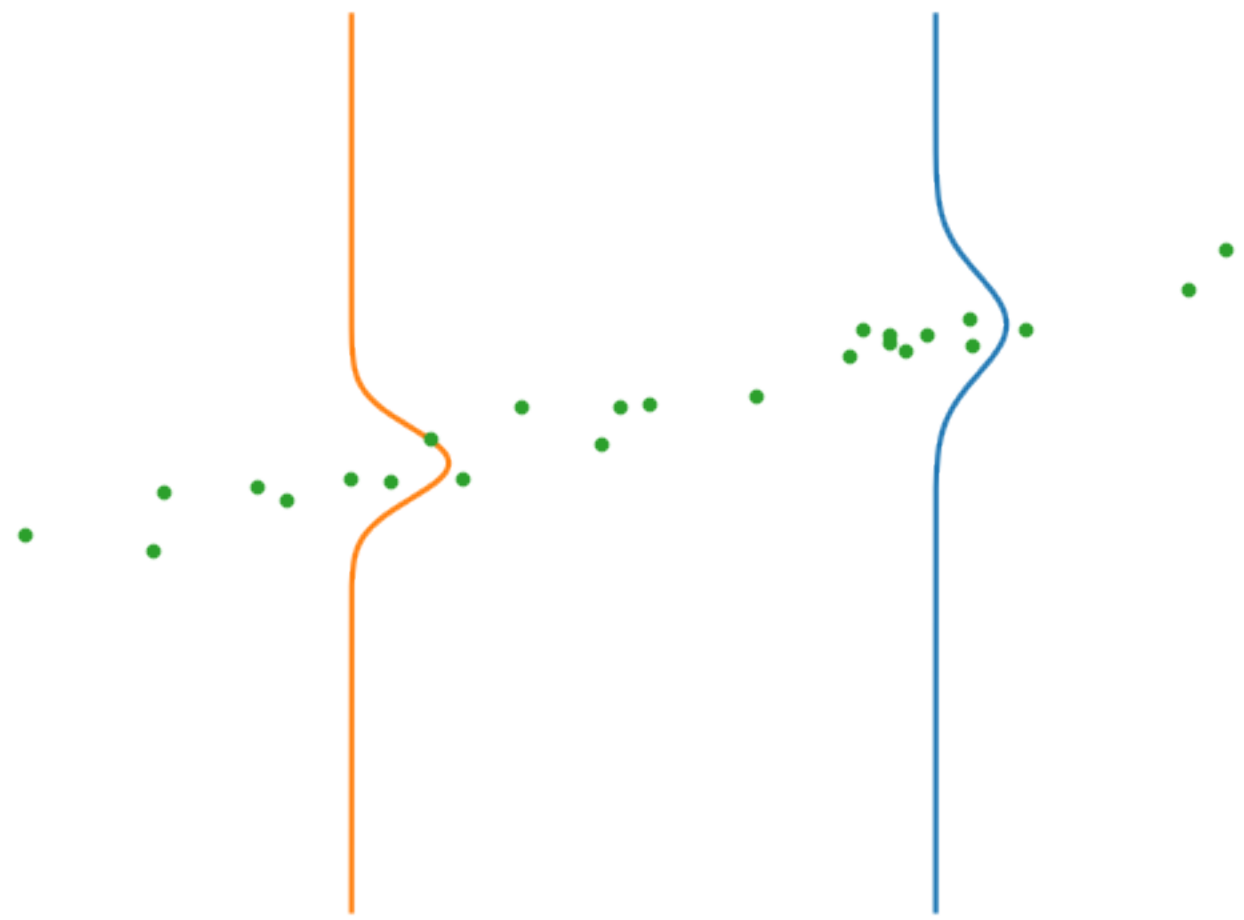
Background

Maximum Likelihood Recap

- **Maximum Likelihood**: Maximize the likelihood of the data being generated by your model; $\log p(x)$

- **Maximum Conditional Likelihood**: Given pairs of data, maximize the likelihood of the target data following the mapping defined by your model; $\log p(y|x)$





Training Latent Variable Models

- **Maximum Marginal Likelihood**: In the face of incomplete data, maximize the observed data's likelihood governed by the marginal distribution;

$$\log p(x) = \log \int_{\lambda \in \Lambda} p(x, \lambda) \mathrm{d}\lambda$$

- Gradient descent-based methods work for some discrete cases, when the latent variable is enumerable:

$$\log p(x) = \log \sum_{z \in Z} p(x, z) = \log \sum_{z \in Z} p(x|z)p(z)$$

Expectation-Maximization

- EM is a coordinate ascent method that maximizes the following objective, known as the **Evidence Lower Bound** (ELBO):

$$\mathcal{L}(\theta, \phi) = \underbrace{\mathbb{E}_{z \sim q_\phi(z)} [\log p_\theta(x, z)]}_{\text{Expected complete data likelihood}} + \underbrace{\mathcal{H}(q_\phi(z))}_{\text{Entropy}}$$

$$\text{E step:} \qquad q_{t+1} \leftarrow \underset{q}{\operatorname{argmax}}\, \mathcal{L}(p_t, q) = p_t(z|x)$$

$$\text{M step:} \qquad p_{t+1} \leftarrow \underset{p}{\operatorname{argmax}}\, \mathcal{L}(p, q_{t+1})$$

- What is Bayes Rule? What is the prior and posterior probability?

Approximate Inference

The need of approximate inference

$$\mathbb{E}_{z \sim q_\phi(z)} \left[\log p_\theta(x, z) \right]$$

make a guess!

Markov Chain Monte Carlo

- Sampling method
- Unbiased but high variance

Variational Inference

- Optimization method
- **Low variance** but biased

Variational Inference and VAE

Evidence Lower BOund (ELBO)

- Define the joint probability as p(x,z) = p(x|z)p(z)

1. $z \sim p(z)$; e.g. $p(z) = \mathcal{N}(z; 0, \mathrm{I})$
2. $x \sim p(x|z) = p(x; \mathrm{dec}(z))$; e.g. $p(x|z) = \mathrm{Bern}(x; \pi(z))$

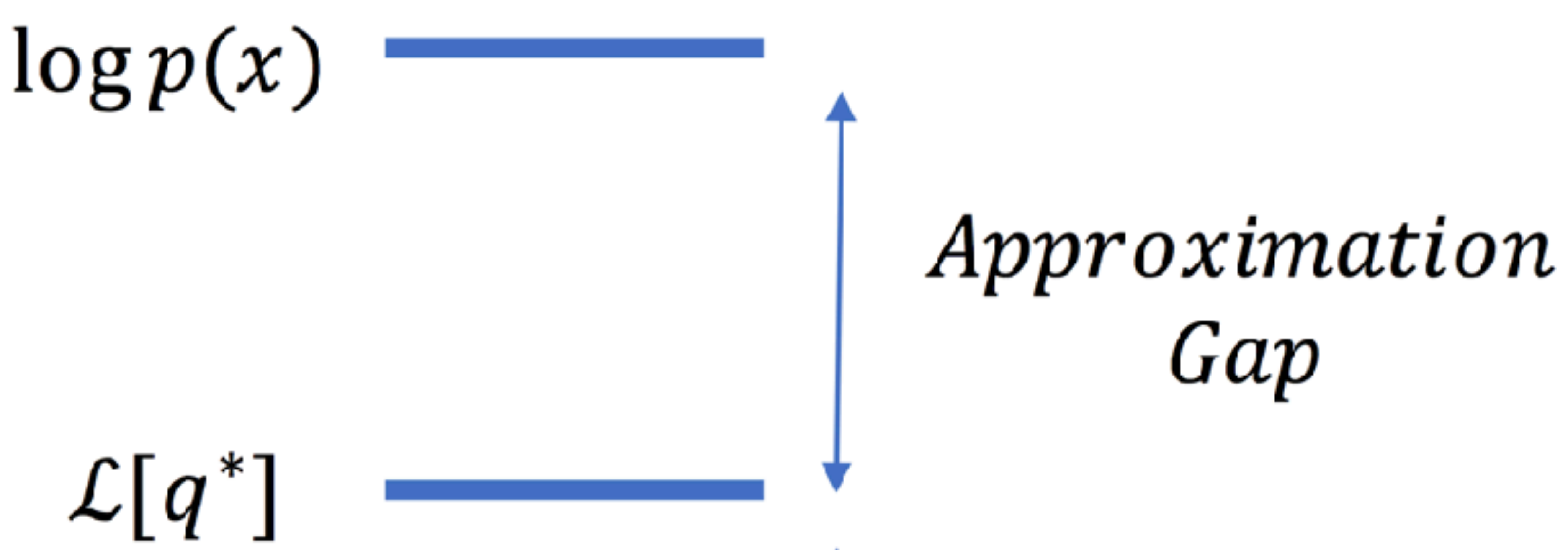
Variational Gap

$$\log p(x) - \overbrace{\mathbb{E}_{z\sim q(z)}\left[\log p(x,z) - \log q(z)\right]}^{\mathcal{L}(p,q)}$$

$$= \mathbb{E}_{z\sim q(z)}[\log q(z) - \log \frac{p(x,z)}{p(x)}]$$

$$= \mathbb{E}_{z\sim q(z)}[\log q(z) - \log p(z|x)]$$

$$= D_{\mathrm{KL}}\left(q(z) \| p(z|x)\right)$$



(Cremer et al. 2017)

Amortized Variational Inference

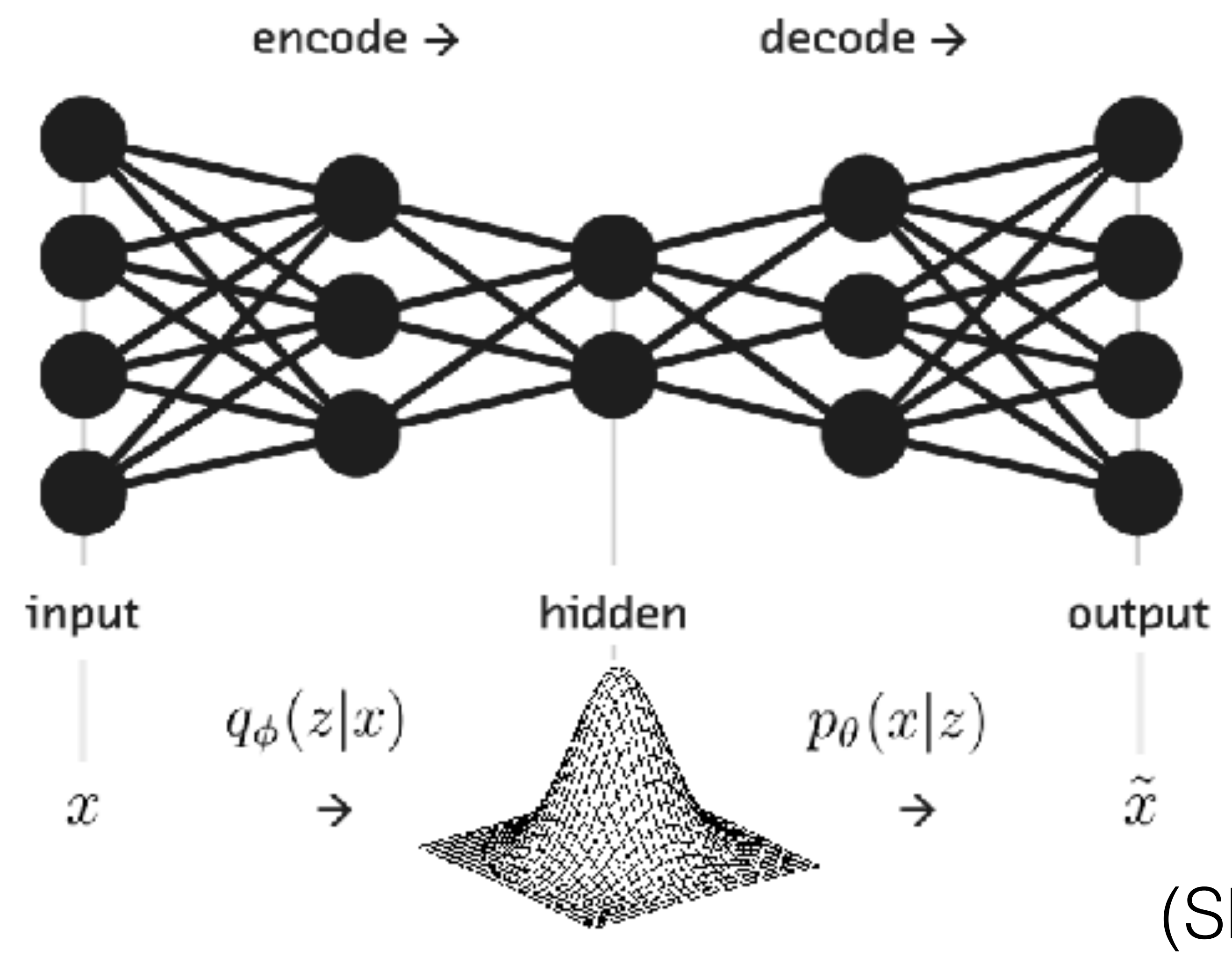
Traditional VI

- For each data instance x_i there is a variational approximation q_i
- When the true posterior $p(z|x_i)$ is not tractable, we need to keep track of the q_i for each i

Amortized VI

- Use an **inference network** (encoder) $q_\phi(z|x_i)$

$$\boxed{\begin{aligned}\mathcal{L}_i(\theta,\phi) &= \mathbb{E}_{z\sim q_\phi(z|x_i)}\left[\log p_\theta(x_i,z) - \log q_\phi(z|x_i)\right] \\ &= \underbrace{\mathbb{E}_{z\sim q_\phi(z|x_i)}\left[\log p_\theta(x_i|z)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}\left(\log q_\phi(z|x_i)\|p(z)\right)}_{\text{regularization term}}\end{aligned}}$$



(Shiffman et al., 2016)

Amortized inference (Gershman & Goodman, 2014)
Auto-encoding variational bayes, AEVB (Kingma & Welling, 2014)
Recognition model (Rezende et al, 2014)

Stochastic Variational Inference

- We want to estimate the gradient of the ELBO (Monte Carlo)

$$\nabla \mathcal{L}_i = \mathbb{E}_z \left[(\log p(x_i, z) - \log q(z|x_i)) \nabla \log q(z|x_i) \right]$$

$$\approx \frac{1}{M} \sum_{m=1}^{M} (\log p(x_i, z_m) - \log q(z_m|x_i)) \nabla \log q(z_m|x_i) \qquad z_m \sim q(z|x_i)$$

unbiased, high variance

Remedies:
- More samples
- Rao-Blackwellization
- Control Variate
- Baseline …

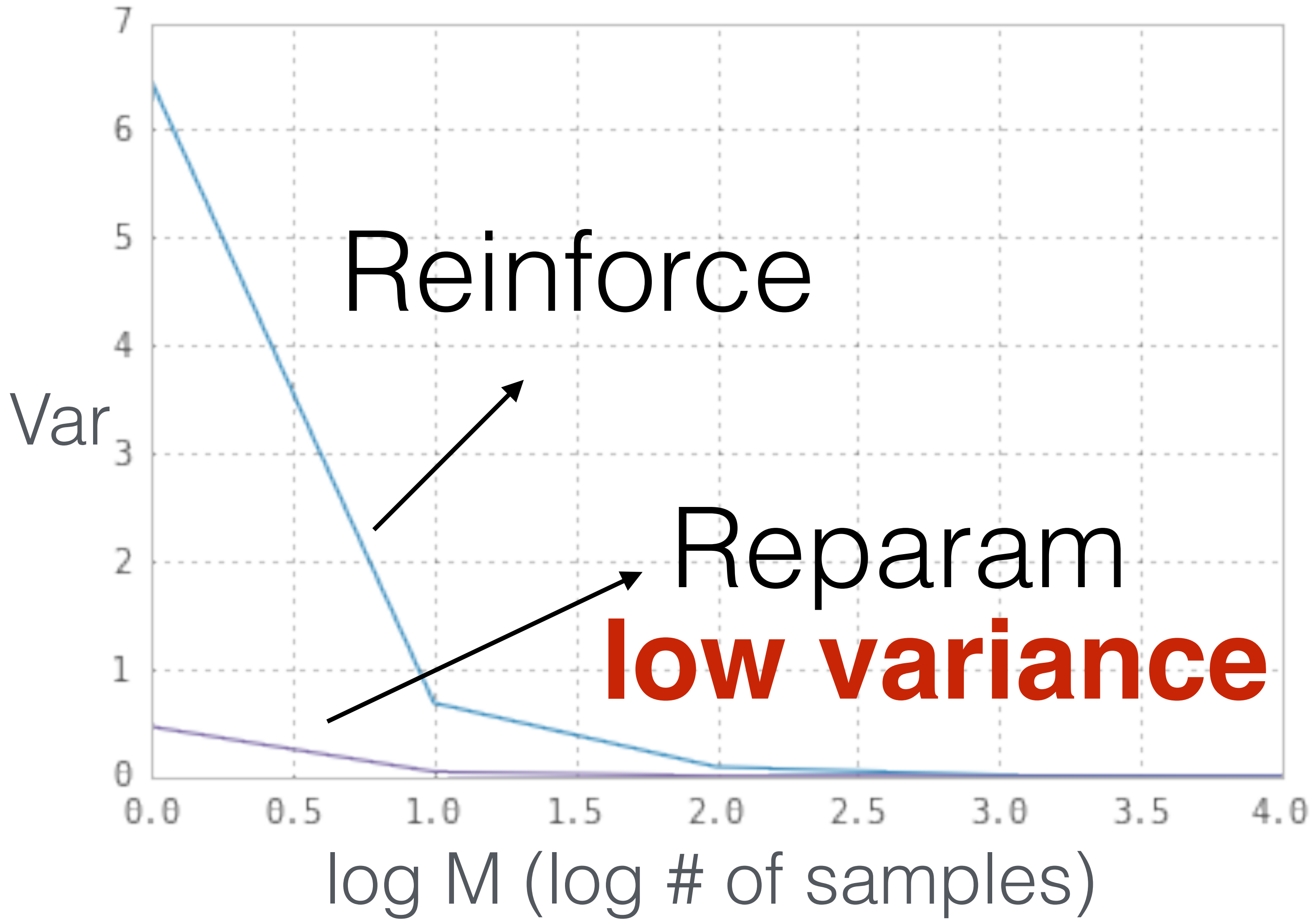
Other gradient estimator?

Black-Box Variational Inference (Ranganath et al., 2014)
NVIL (Mnih & Gregor, 2014)

Reparameterization Trick

- Decompose the sampling process into

$$z \sim q_\phi(z|x) \;\Leftrightarrow\; \epsilon \sim q(\epsilon);$$
$$z = g_\phi(\epsilon, x)$$
$$z \sim \mathcal{N}(z; \mu_\phi(x), \sigma^2_\phi(x)) \;\Leftrightarrow\; \epsilon \sim \mathcal{N}(\epsilon; 0, \mathrm{I});$$
$$z = \sigma(x) \odot \epsilon + \mu(x)$$



How to estimate the gradient then?

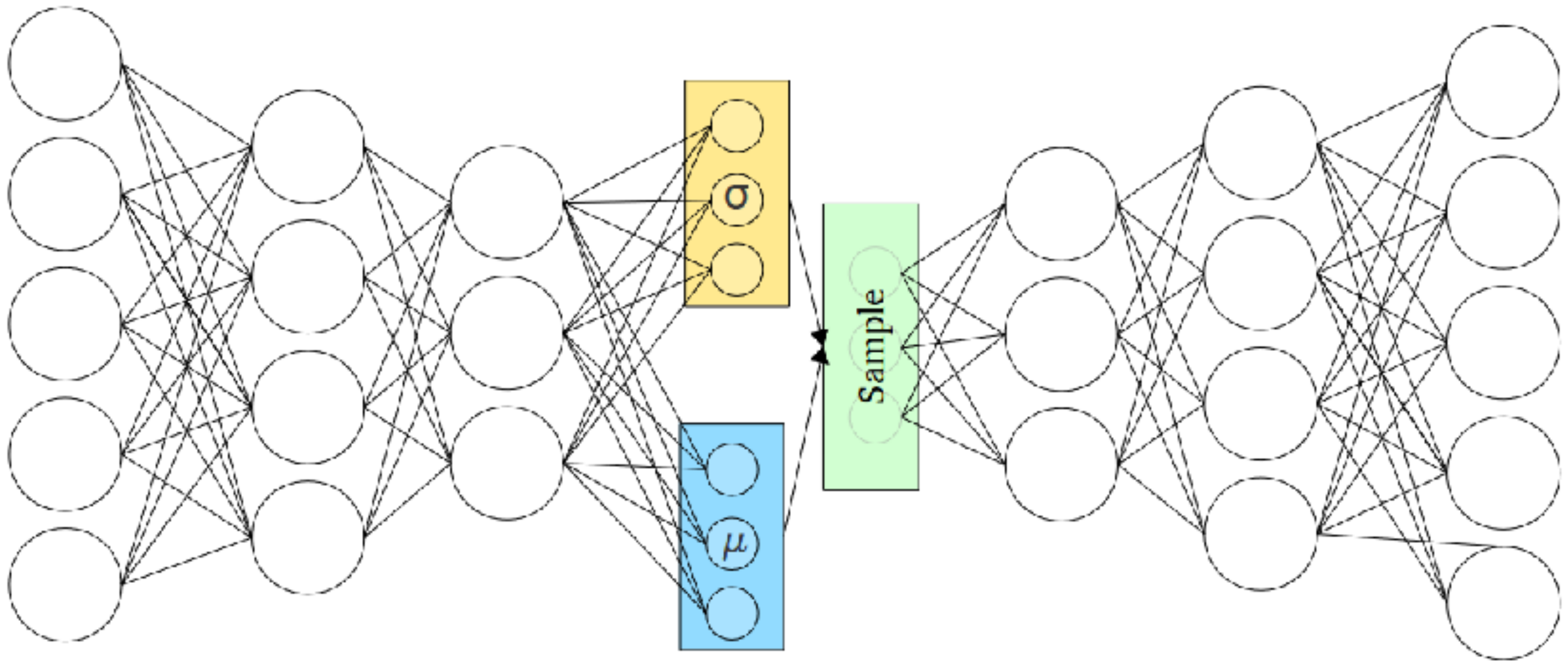
Doubly Stochastic Variational Bayes (Titsias & Lazaro-Gredilla, 2014)
Stochastic Gradient Variational Bayes, SGVB (Kingma & Welling, 2014)
Stochastic Backpropagation (Rezende et al, 2014)

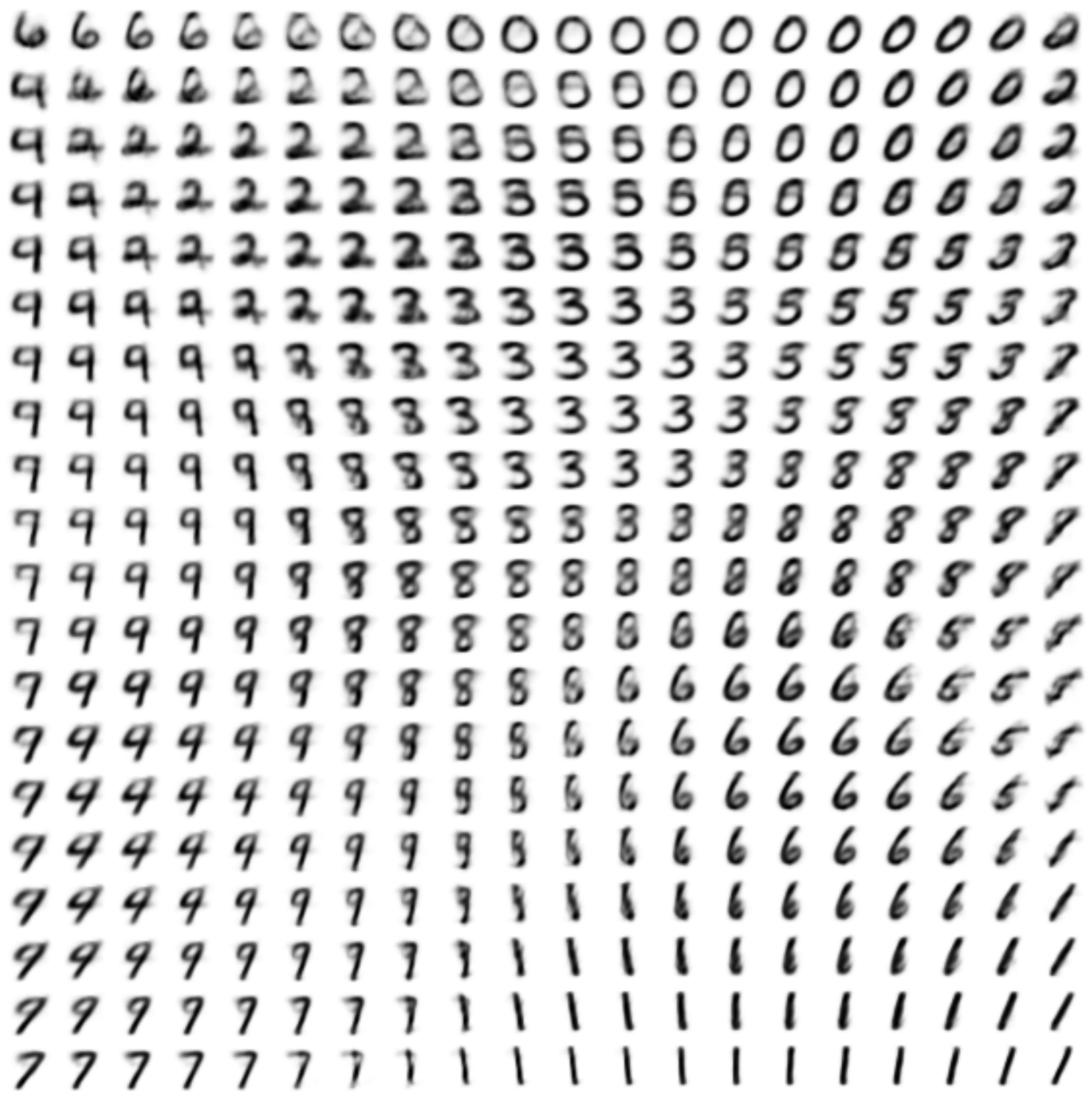
(Erdogan et al., 2016)

Variational Auto-Encoder (VAE)

- Vanilla VAE

$$p(z) = \mathcal{N}(0, \mathrm{I}); \quad p(x|z) = \prod_j p(x_j|z); \quad q(z|x) = N(\mu(x), \mathrm{diag}(\sigma^2(x)))$$





(Shafkat et al. 2016)

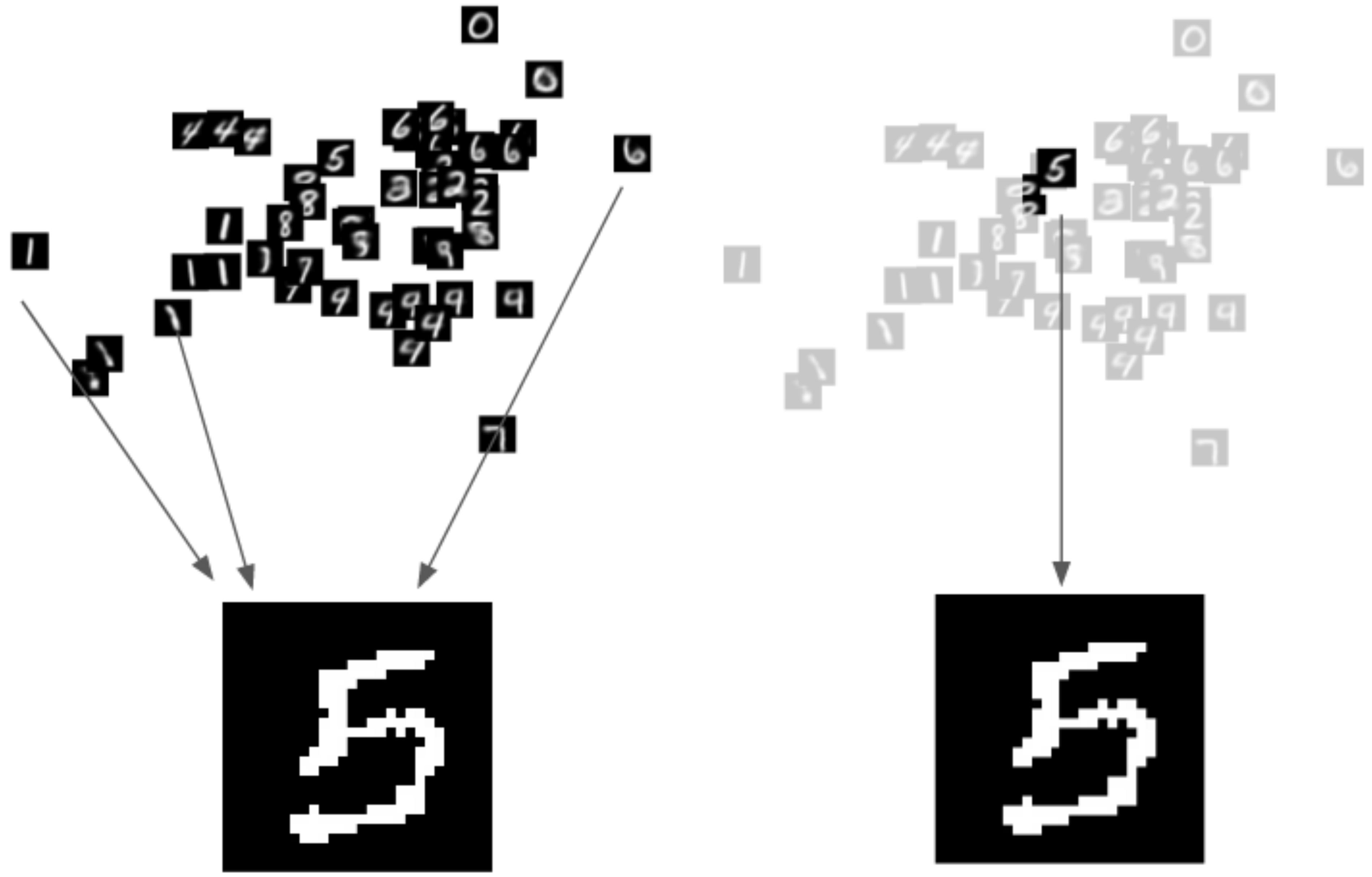
(Kingma et al. 2014)

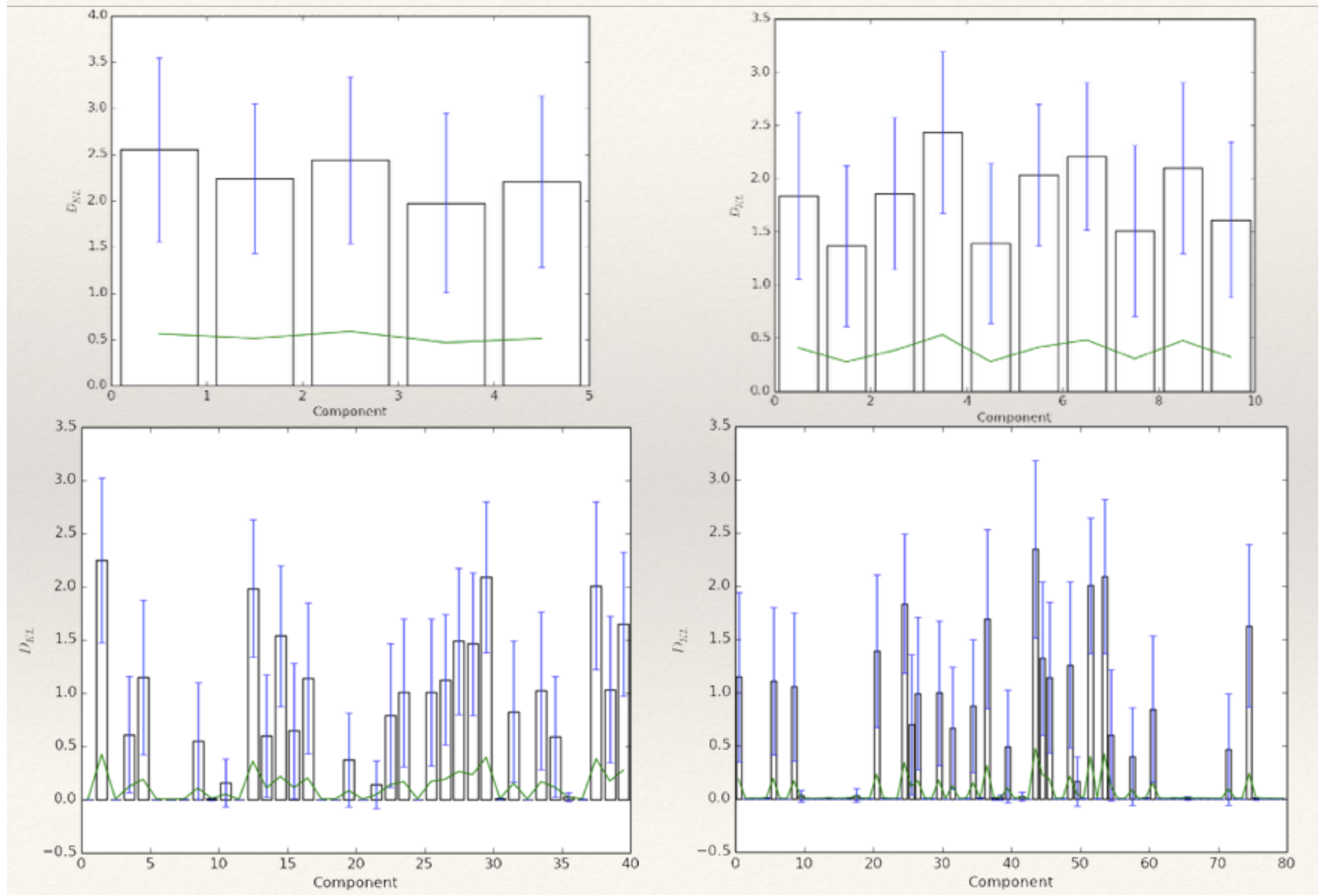
Caveats, Information Theoretic POV

- Regularized Autoencoder?

the smaller the better?

$$\underbrace{\mathbb{E}_{z\sim q_\phi(z|x_i)}\left[\log p_\theta(x_i|z)\right]}_{\text{reconstruction loss}} - \underbrace{D_{\mathrm{KL}}\left(\log q_\phi(z|x_i)\|p(z)\right)}_{\text{regularization term}}$$





(Dinh & Dumoulin 2014)

Evaluation

- Importance Sampling (stochastic lower bound)

$$p(x) \approx \frac{1}{M} \sum_{m=1}^{M} \frac{p(x, z_m)}{q(z_m|x)}$$

$$\log p(x) \approx \underset{m=1}{\overset{M}{\mathrm{L\Sigma E}}} \left[\log p(x, z_m) - \log q(z_m|x)\right]$$

- Further reading: Annealed Importance Sampling by Radford Neal

(Neal, 1998)

What Assumptions have we made so far?

MILA

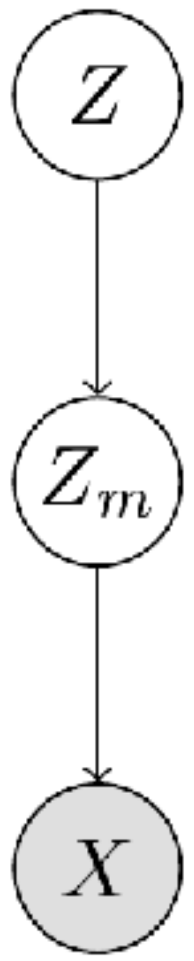
- What can we change? (Non-Vanilla VAE)
  - fixing $p(z) = \mathcal{N}(0, \mathrm{I})$
  - fixing $p_\theta(x|z) = \prod_j p_\theta(x_j|z)$; $\Theta : \theta \in \Theta$
  - fixing $\mathcal{Q} : q \in \mathcal{Q}$
  - using $q_\phi(z|x_i)$ instead of $q_i(z)$; fixing $\Phi : \phi \in \Phi$

Better Model Assumptions

What likelihood p(x|z) to use?

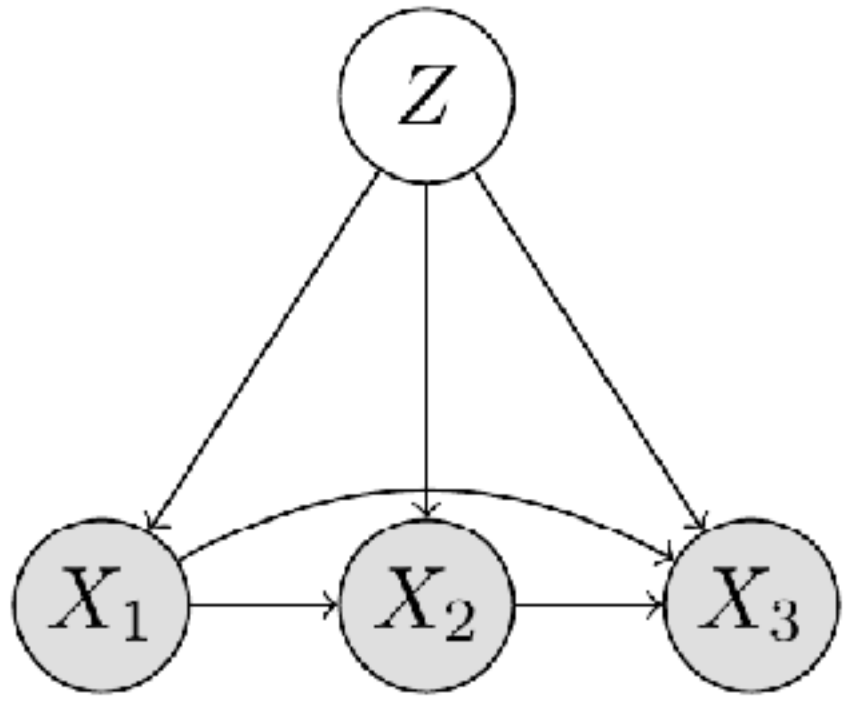
- Hierarchy?

$$p(x|z) = \int_{z_{mid}} p(x, z_{mid}|z) \mathrm{d}z_{mid}$$



- Autoregressive model?

$$p(x|z) = \prod_j p(x_j | x_{j'<j}, z)$$

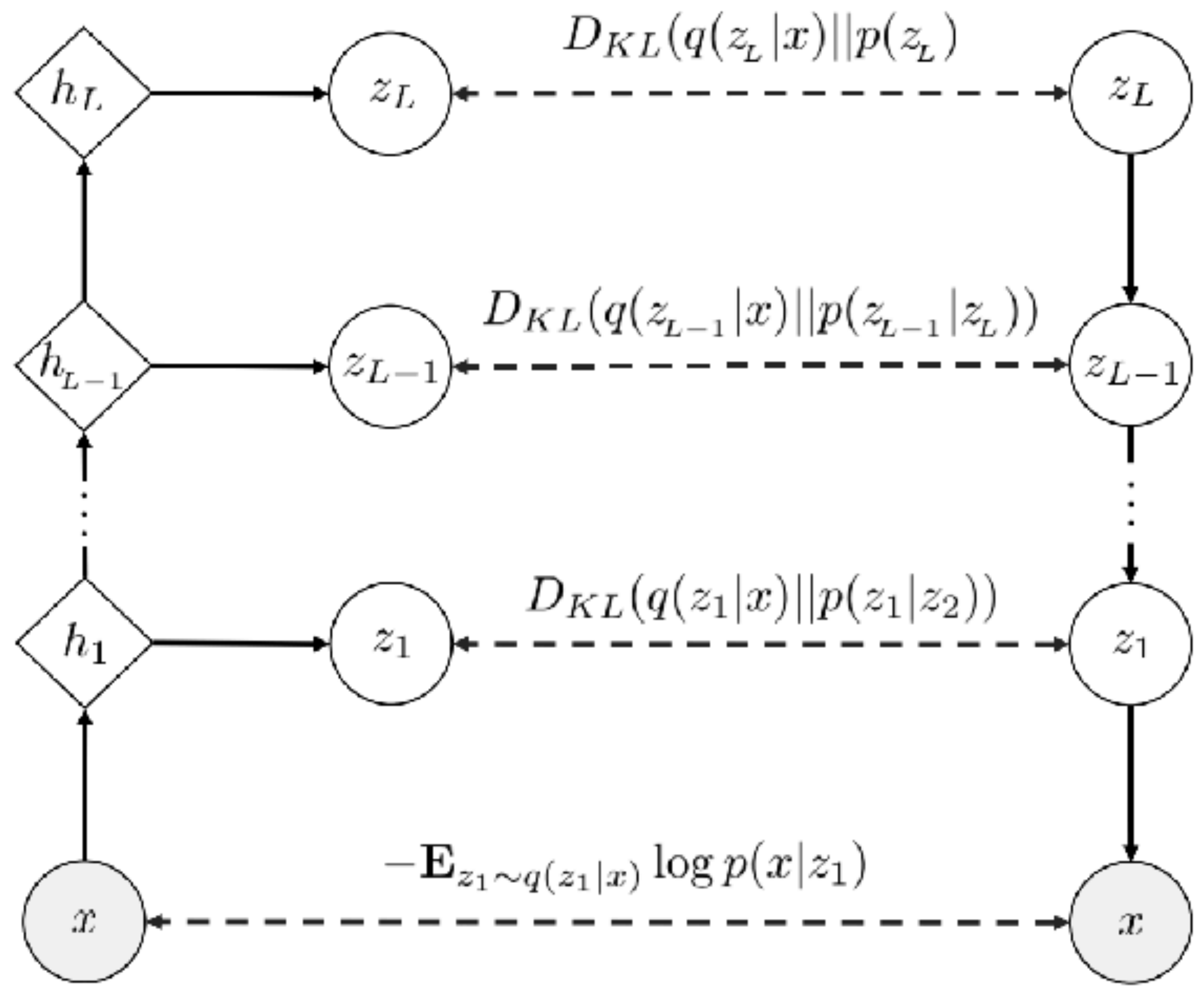


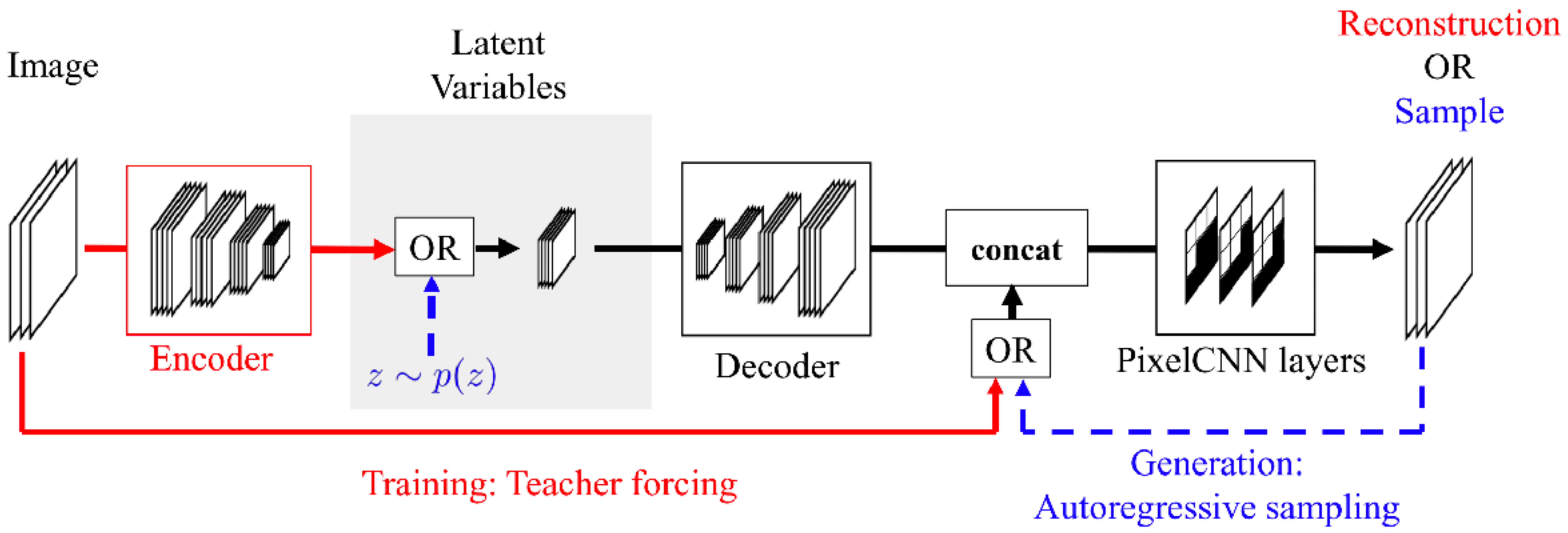
PixelVAE

$$-L(x,q,p) = -E_{z_1\sim q(z_1|x)}\log p(x|z_1) + D_{KL}(q(z_1,\cdots z_L|x)||p(z_1,\cdots,z_L))$$

$$\vdots$$

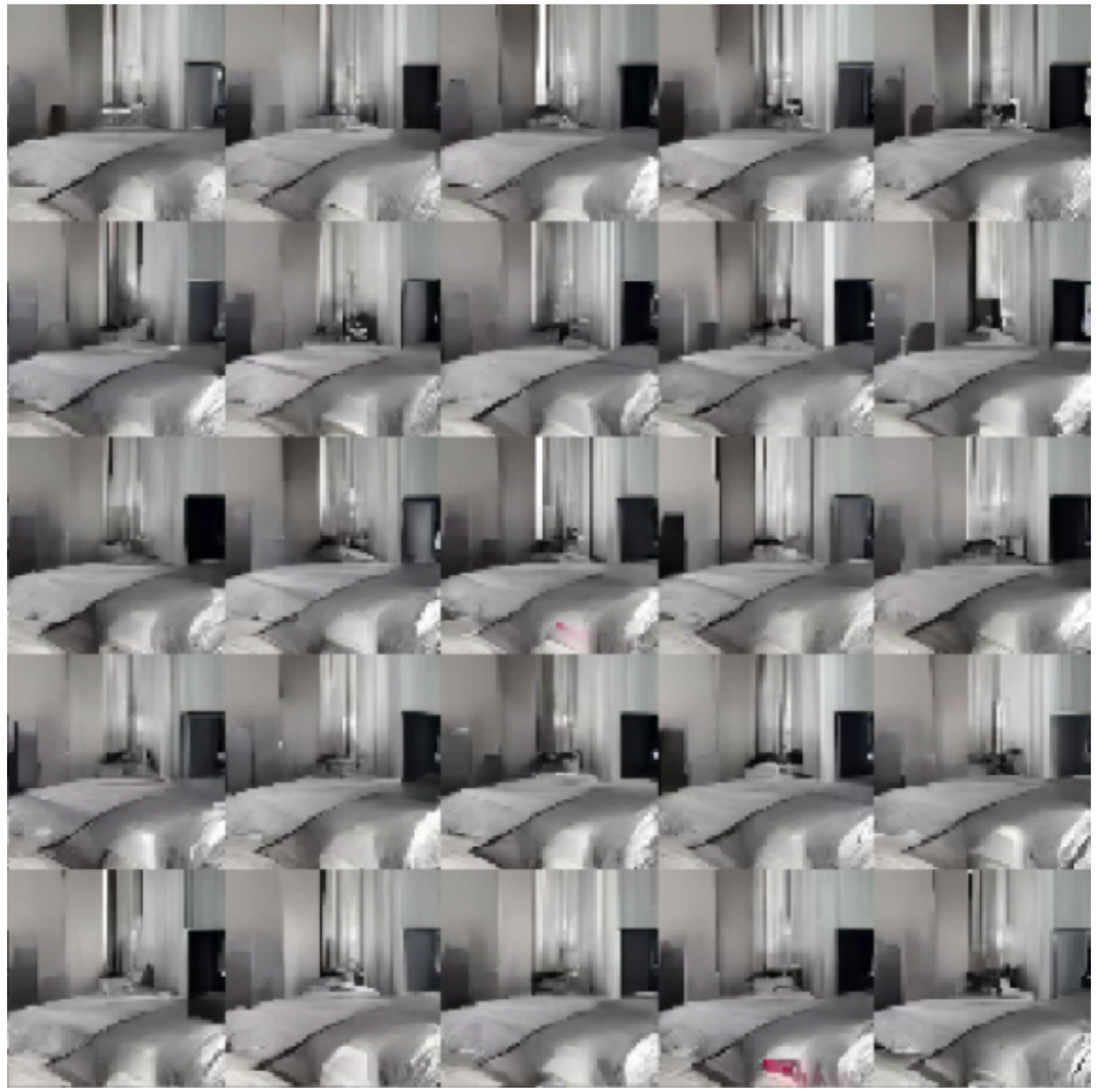
$$= -E_{z_1\sim q(z_1|x)}\log p(x|z_1) + \sum_{i=1}^{L}\mathbf{E}_{z_{i+1}\sim q(z_{i+1}|x)}\left[D_{KL}(q(z_i|x)||p(z_i|z_{i+1}))\right] \qquad p(x|z_1)=\prod_j p(x_j|x_{j'<j},z_1)$$

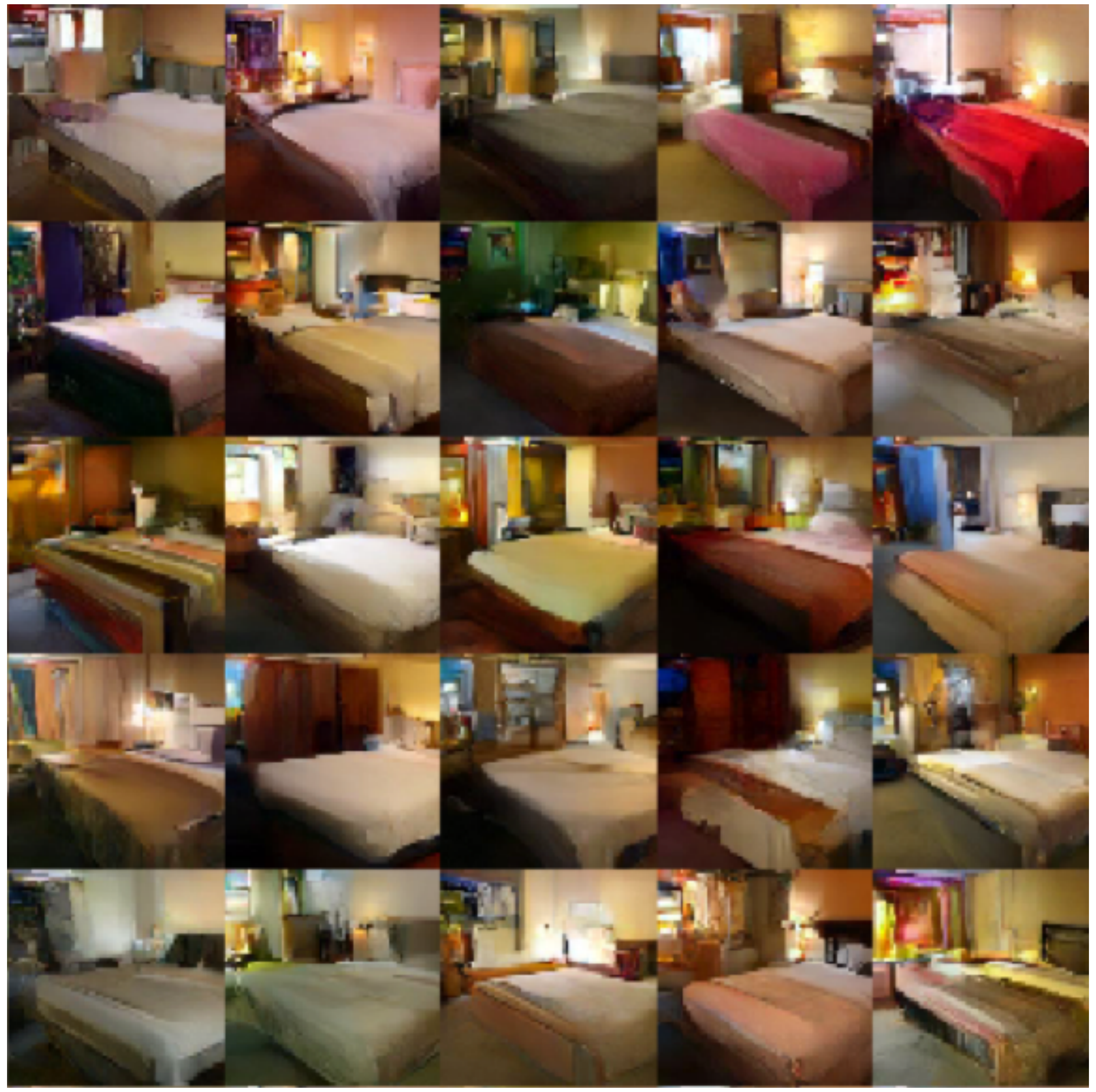




(Gulrajani et al., 2017)

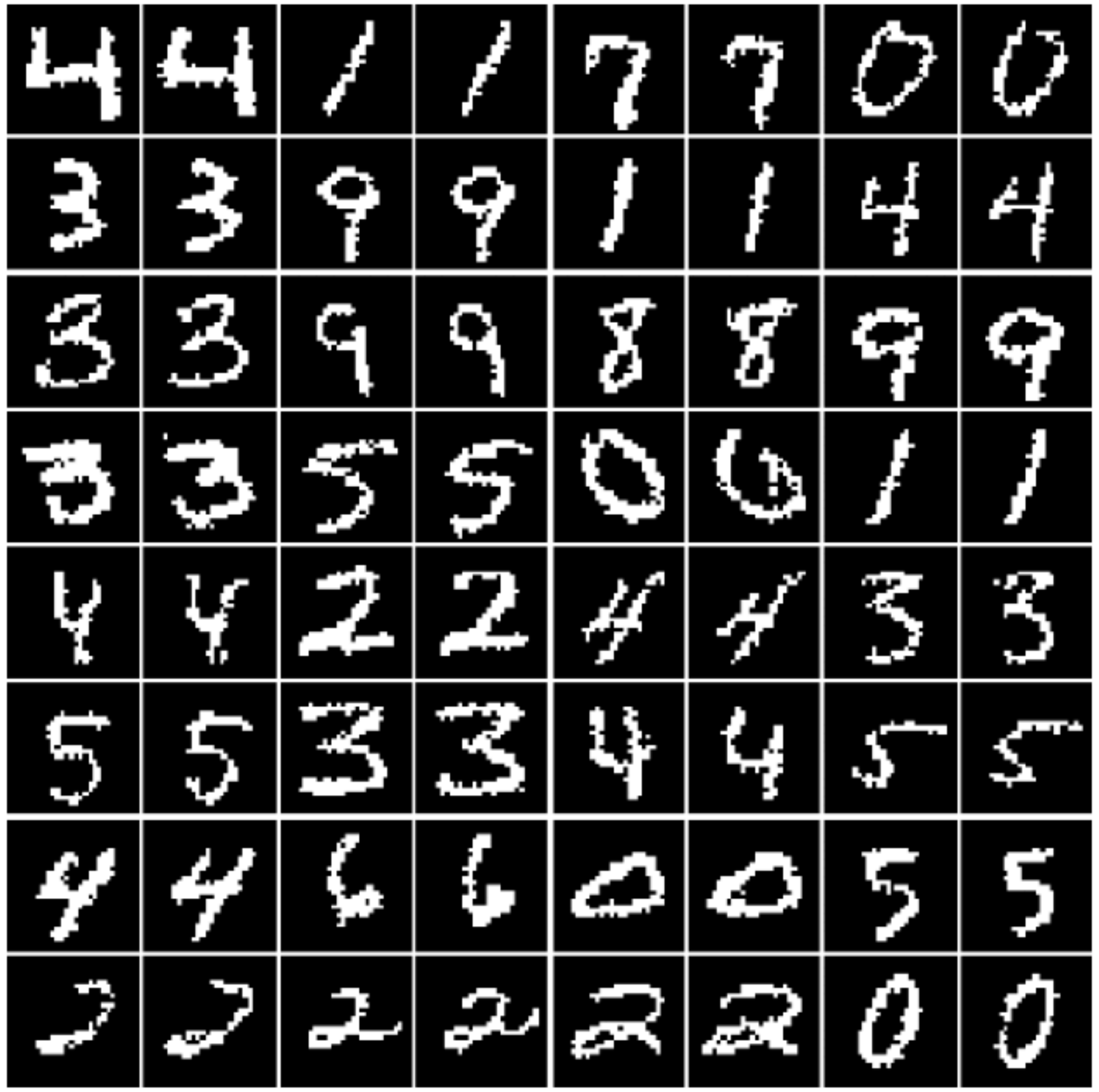
PixelVAE samples

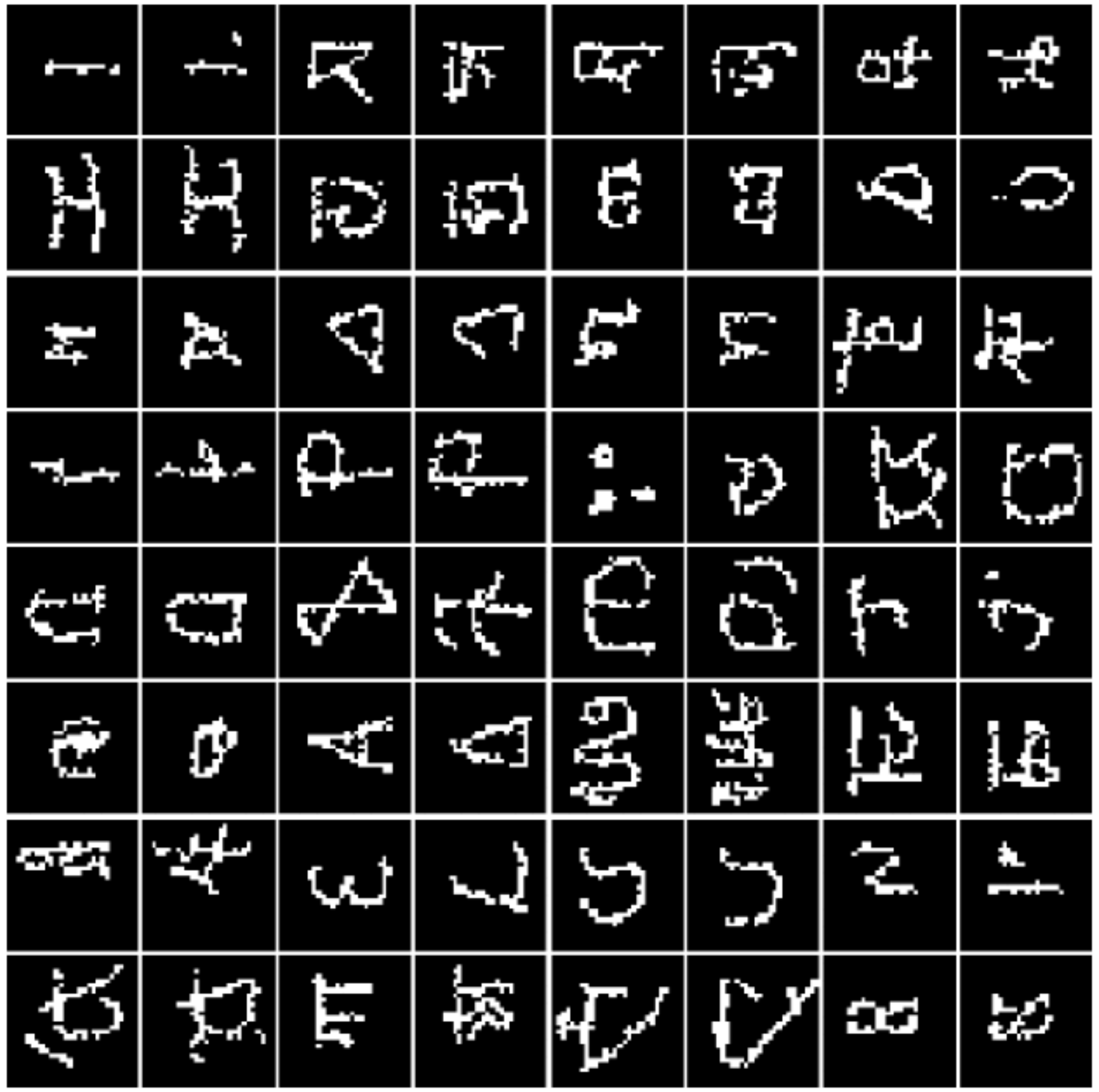




(Gulrajani et al., 2017)

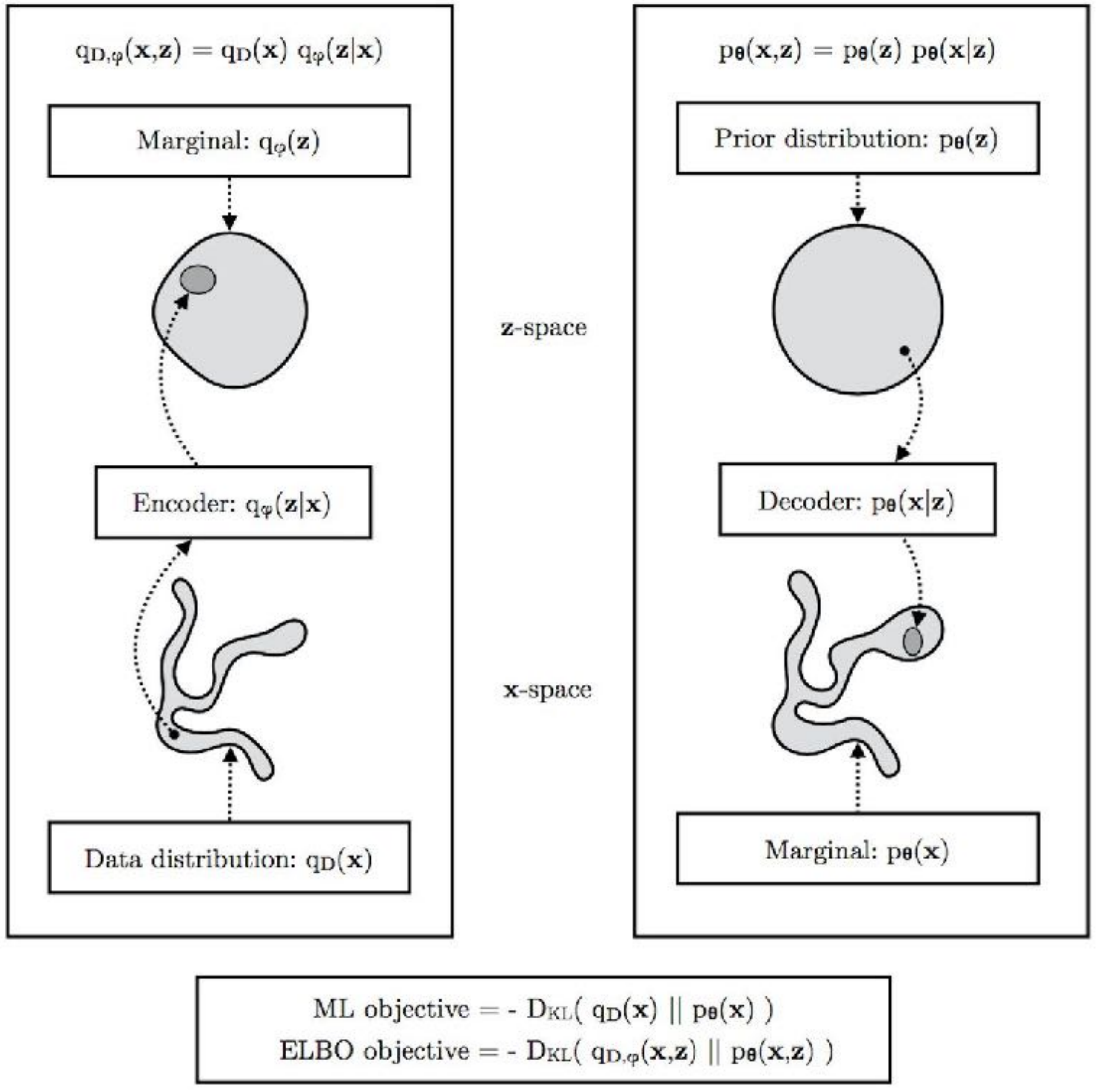
Lossy Encoding

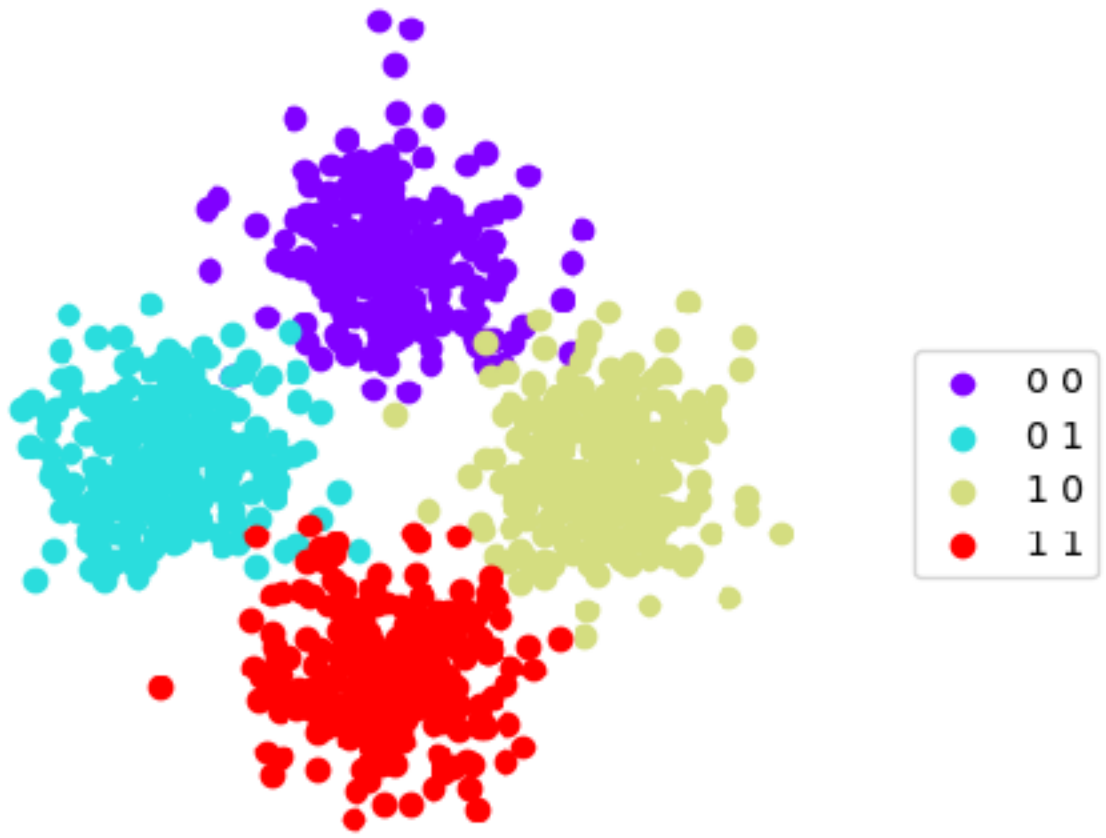


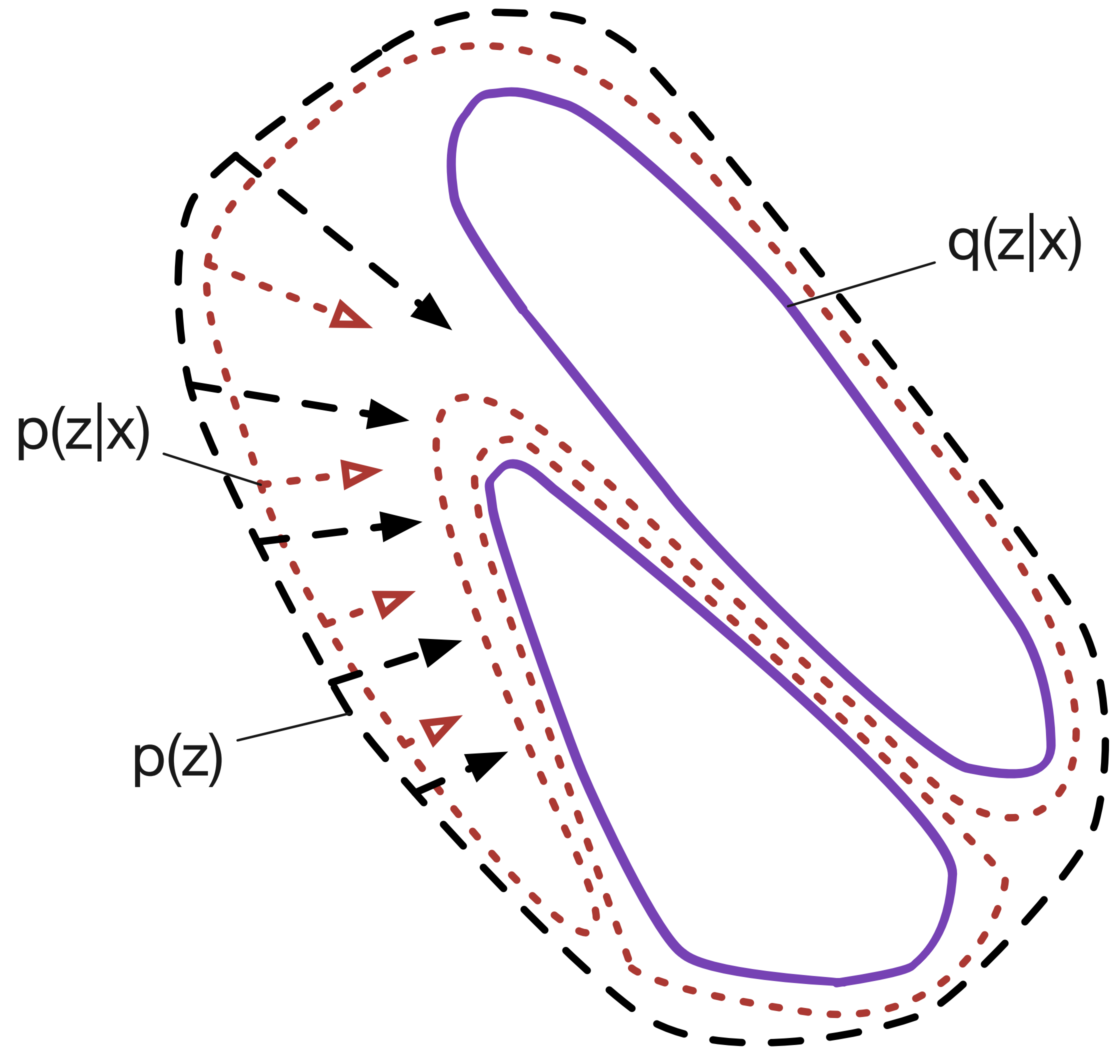


(Chen et al., 2017)

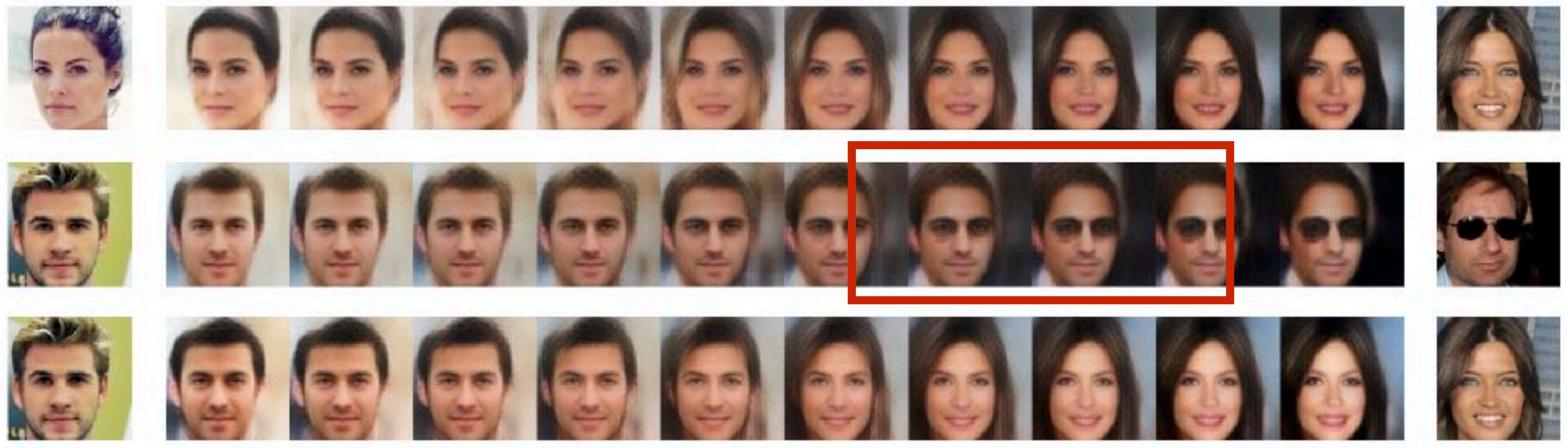
What prior p(z) to use?







$$q(z) = \sum_x q_D(x) q_\phi(z|x) = \mathbb{E}_{q_D(x)}\left[q(z|x)\right]$$



(Kingma et al. 2017)

(Huang et al., 2017; Hou et al., 2016)

Prior Can Be Learned!

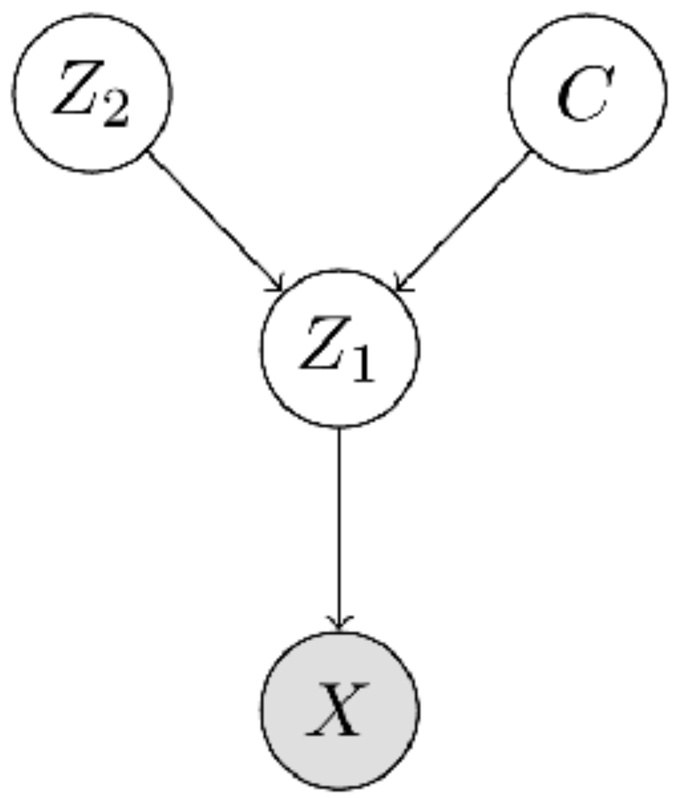
if DLGMs find it difficult to produce unimodal Gaussian marginal posteriors, then perhaps we should investigate multimodal priors that can meet q(z) halfway.

— Matt Hoffman, 2016

NIPS Workshop on Advances in Approximate Bayesian Inference Workshop

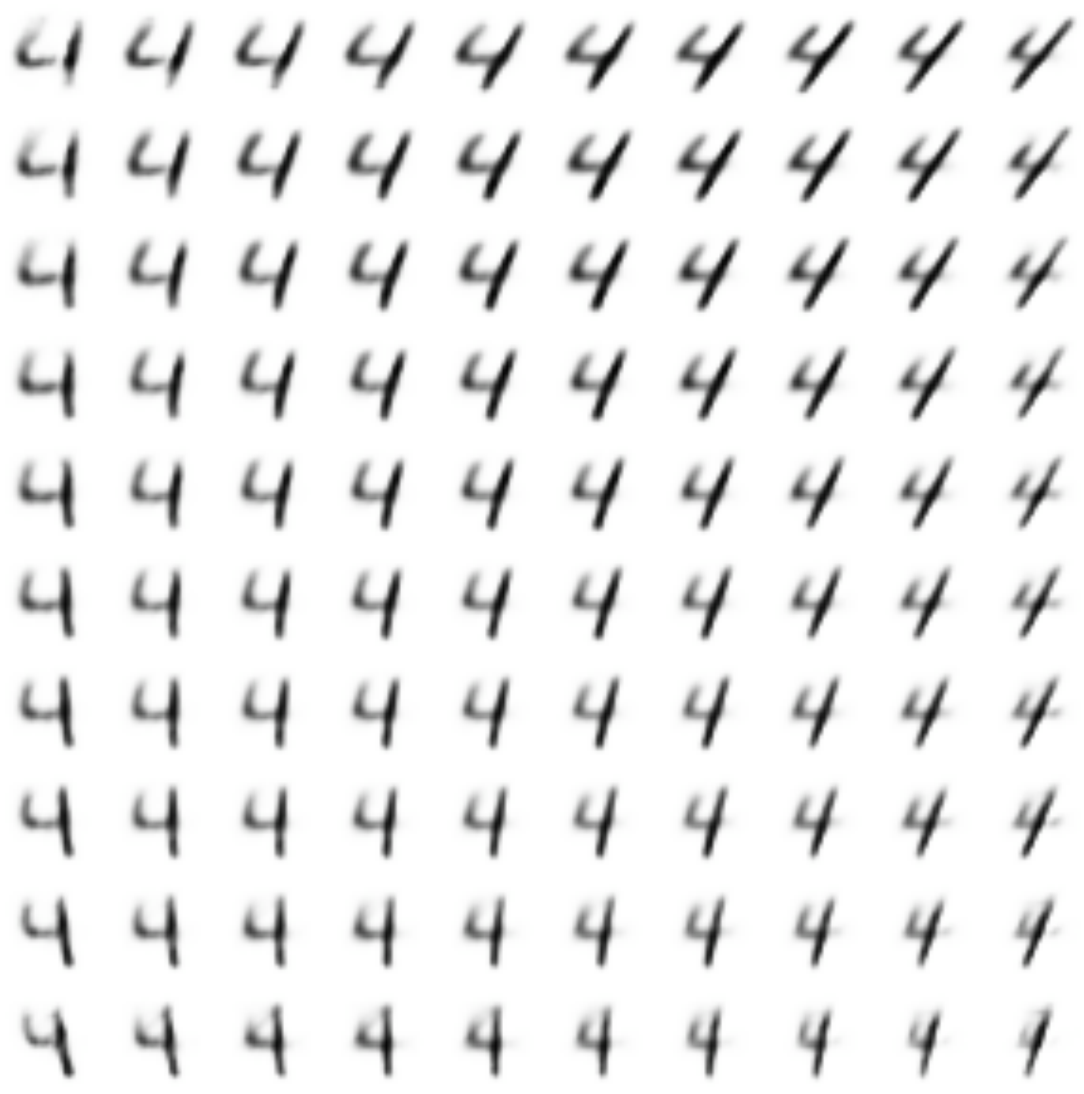
(Hoffman & Johnson, 2016)

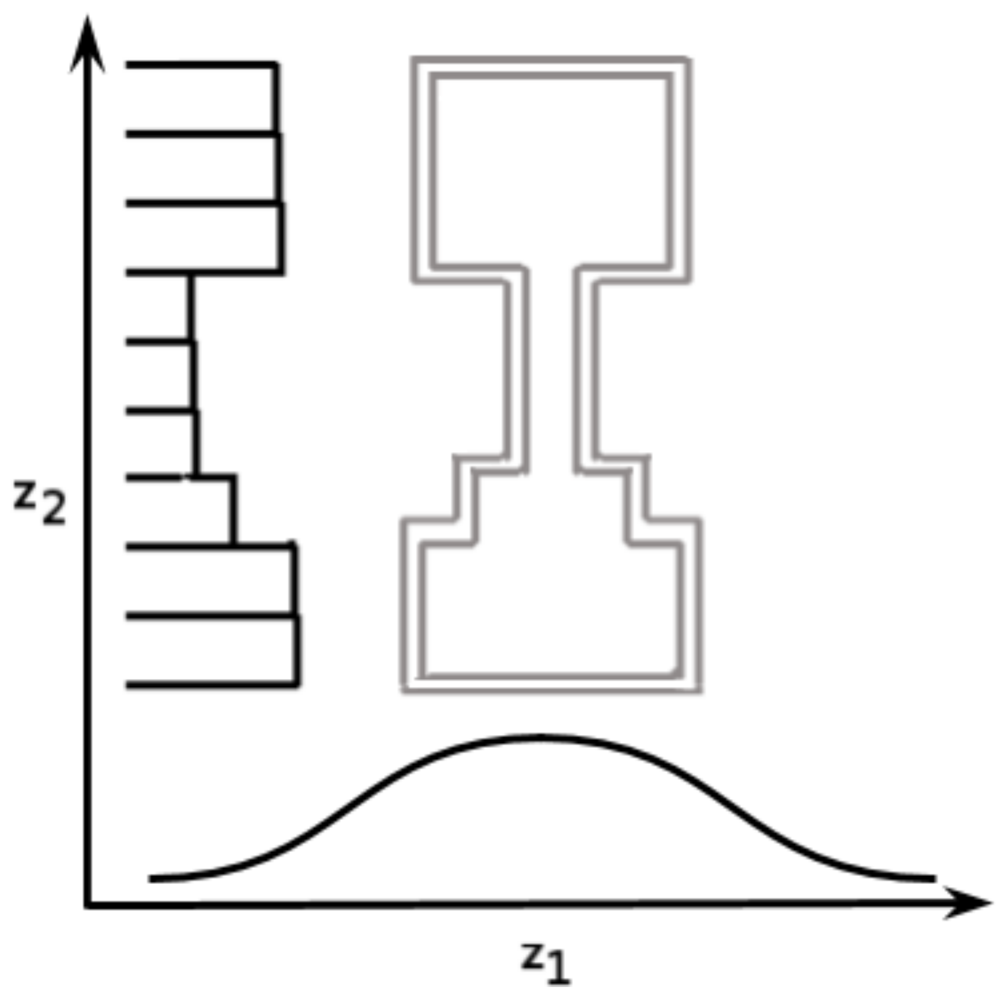
Multimodal Prior



$$p(z_1|z_2) = \sum_c p(z_1|z_2, c)p(c)$$

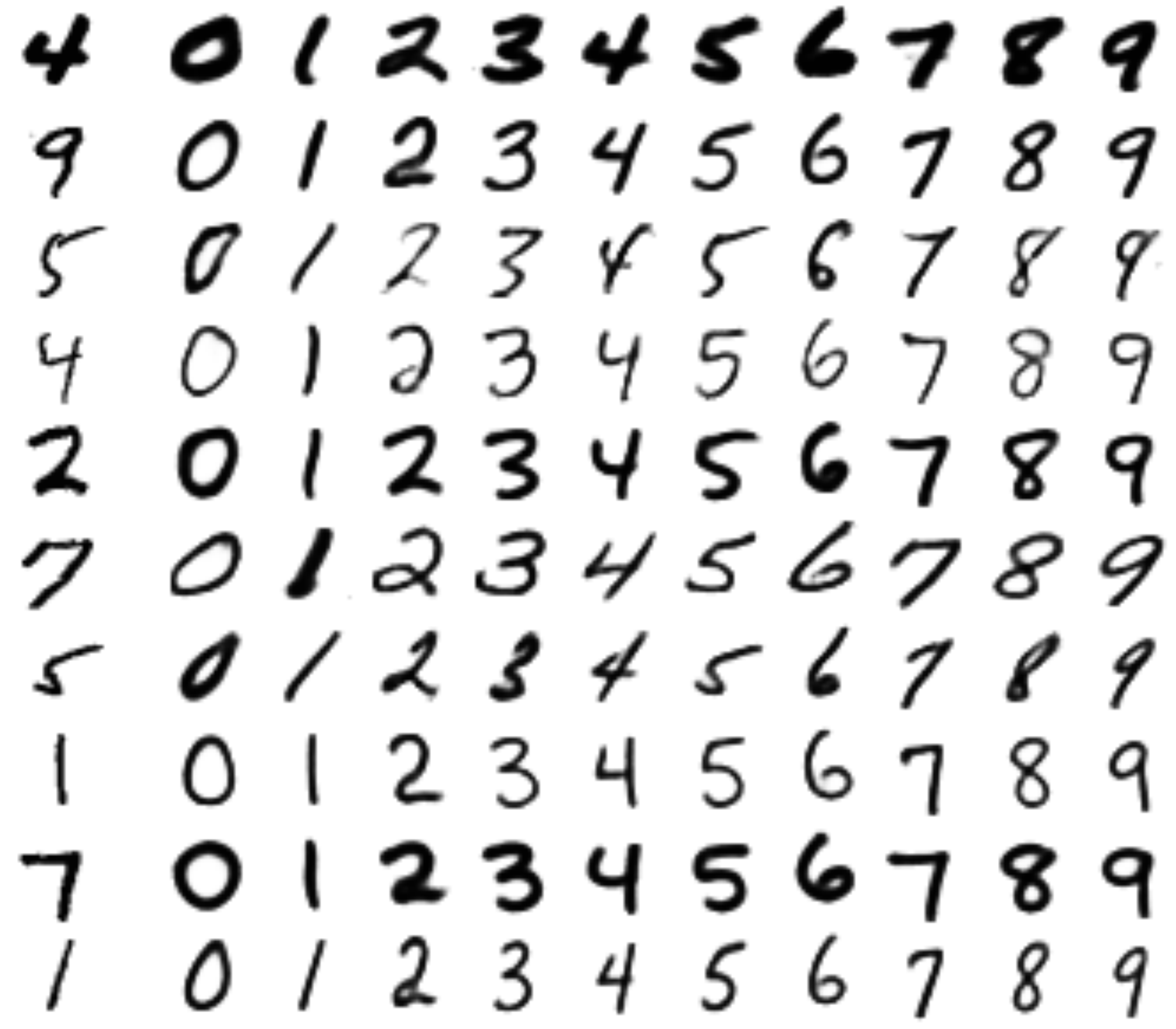
mixture (of Gaussians) prior





$$P(z) = \frac{1}{K} \sum_{i=1}^{n} 1_{\left(\frac{i-1}{n} \leq z \leq \frac{i}{n}\right)}^{a_i}$$

$$K = \sum_{i=1}^{n} K_i, \quad \text{where } K_0 := 0, K_i := \frac{a_i}{n} \text{ for } i = 1, \ldots, n.$$



(Kingma et al., 2014; Dilokthanakul et al. 2016)

(Serban et al., 2016)

Better Inference

Suboptimality in Variational Inference

MILA

$$\log p_\theta(x) - \mathcal{L}(\theta, \phi; x) = \underbrace{D_{\mathrm{KL}}\left(q^* \| p\right)}_{\text{Gap 1}} + \underbrace{\left[D_{\mathrm{KL}}\left(q^*_\theta \| p\right) - D_{\mathrm{KL}}\left(q^* \| p\right)\right]}_{\text{Gap 2}} + \underbrace{\left[D_{\mathrm{KL}}\left(q_\theta \| p\right) - D_{\mathrm{KL}}\left(q^*_\theta \| p\right)\right]}_{\text{Gap 3}}$$

Dataset	MNIST				Fashion MNIST			
Variational Family	q_{FFG}	q_{FFG}	q_{Flow}	q_{Flow}	q_{FFG}	q_{FFG}	q_{Flow}	q_{Flow}
Encoder Capacity	Regular	Large	Regular	Large	Regular	Large	Regular	Large
$\log \hat{p}(x)$	-89.85	-89.09	-88.82	-88.59	-97.78	-94.55	-97.35	-96.16
$\mathcal{L}_{\mathrm{VAE}}[q^*_{Flow}]$	-90.83	-90.05	-90.40	-90.26	-98.03	-95.93	-97.74	-96.87
$\mathcal{L}_{\mathrm{VAE}}[q^*_{FFG}]$	-91.19	-90.34	-102.88	-103.53	-99.03	-98.09	-130.90	-129.24
$\mathcal{L}_{\mathrm{VAE}}[q]$	-92.76	-91.12	-91.42	-91.25	-103.20	-101.28	-102.19	-100.60
Approximation	1.34	1.25	1.58	1.67	1.25	3.54	0.39	0.71
Amortization	1.57	0.78	1.02	0.99	4.17	3.19	4.45	3.73
Inference Gap	2.91	2.03	2.60	2.66	5.42	6.73	4.84	4.44

(Cremer et al. 2017)

(Krishnan et al., 2017; Marino et al., 2017; Kim et al., 2018)

Reducing Variational Gap

Methods that aim at reducing the variational gap (nonexhaustive)

	Gap 1	Gap 2	Gap 3
Better optimization			✗
Better encoder [2]		✗	
Meta learning [10, 12, 8]		(✗)	(✗)
Normalizing flows [16, 9, 6]	✗		
Hierarchical VI [15, 11]	✗		
Implicit VI [13, 7]	✗		
Variational boosting [14]	✗		
MCMC as part of q_ϕ [17]	✗	✗	✗
MCMC on top of q_ϕ [3, 5]	(✗)	(✗)	(✗)
Gradient flow [4]	✗	✗	✗
Importance sampling [1]	(✗)	(✗)	(✗)

Reducing Variational Gap Ref

[1] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[2] C. Cremer, X. Li, and D. Duvenaud. Inference suboptimality in variational autoencoders. 2017.

[3] Nando De Freitas, Pedro Højen-Sørensen, Michael I Jordan, and Stuart Russell. Variational mcmc. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 120–127. Morgan Kaufmann Publishers Inc., 2001.

[4] David Duvenaud, Dougal Maclaurin, and Ryan Adams. Early stopping as nonparametric variational inference. In *Artificial Intelligence and Statistics*, pages 1070–1077, 2016.

[5] Matthew D Hoffman. Learning deep latent gaussian models with markov chain monte carlo. In *International Conference on Machine Learning*, pages 1510–1519, 2017.

[6] C.-W. Huang, D. Krueger, and A. Courville. Facilitating multimodality in normalizing flows. 2017.

[7] Ferenc Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

[8] Yoon Kim, Sam Wiseman, Andrew C Miller, David Sontag, and Alexander M Rush. Semi-amortized variational autoencoders. *arXiv preprint arXiv:1802.02550*, 2018.

[9] Diederik P Kingma, Tim Salimans, and Max Welling. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, 2016.

[10] Rahul G Krishnan, Dawen Liang, and Matthew Hoffman. On the challenges of learning with inference networks on sparse, high-dimensional data. *arXiv preprint arXiv:1710.06085*, 2017.

[11] Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. *arXiv preprint arXiv:1602.05473*, 2016.

[12] Joseph Marino, Yisong Yue, and Stephan Mandt. Iterative inference models.

[13] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv preprint arXiv:1701.04722*, 2017.

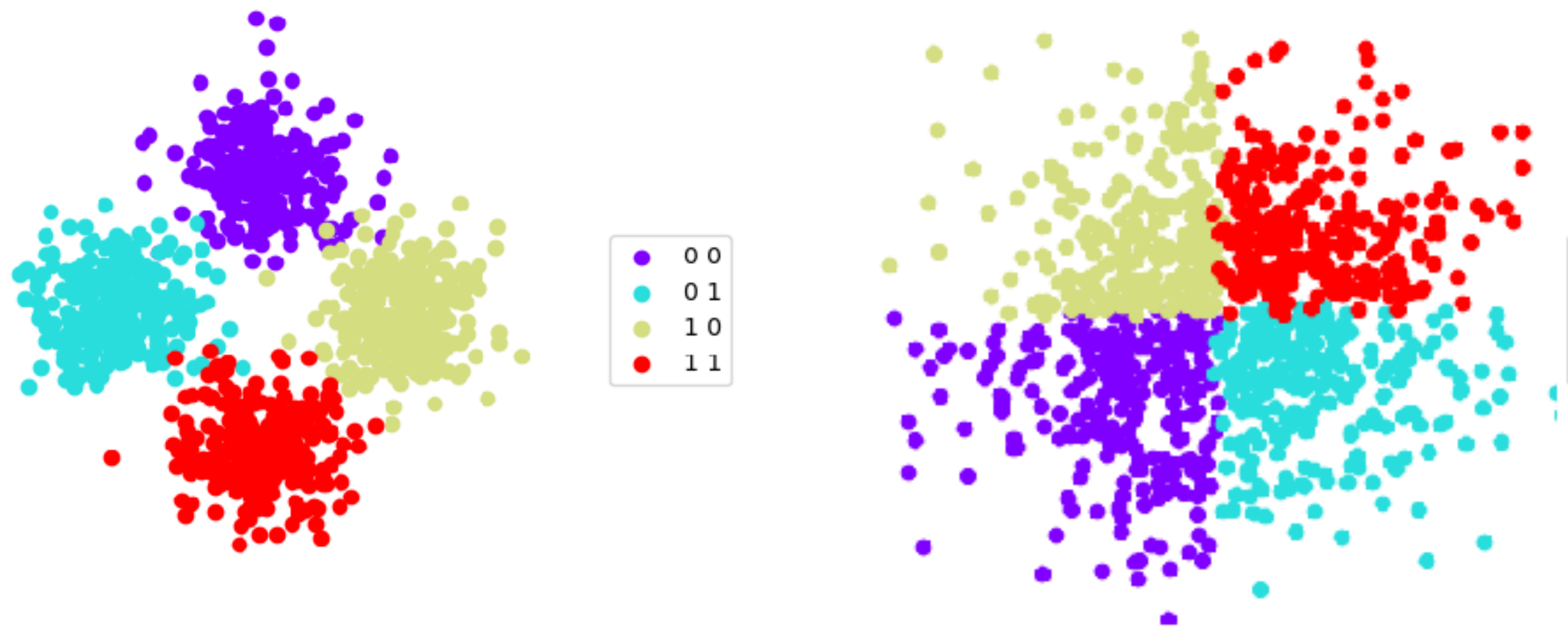
[14] Andrew C Miller, Nicholas Foti, and Ryan P Adams. Variational boosting: Iteratively refining posterior approximations. *arXiv preprint arXiv:1611.06585*, 2016.

[15] Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333, 2016.

[16] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[17] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1218–1226, 2015.

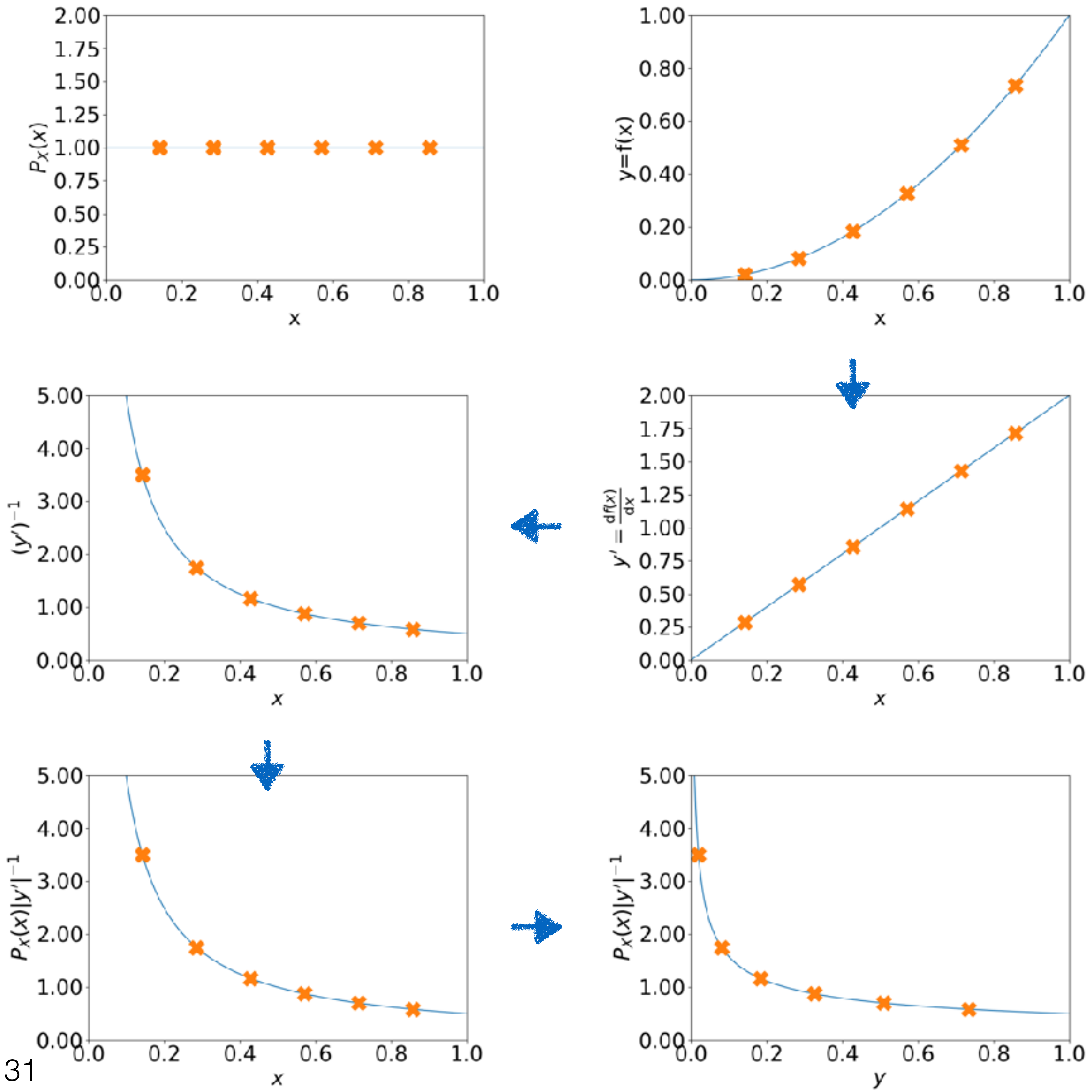
Transformation Methods: Normalizing Flows



- Change of variable distribution

$$x \sim P_X(x); \quad y = f(x)$$

$$P_Y(y) = P_X(x) \left| \frac{\mathrm{d}f(x)}{\mathrm{d}x} \right|^{-1}$$



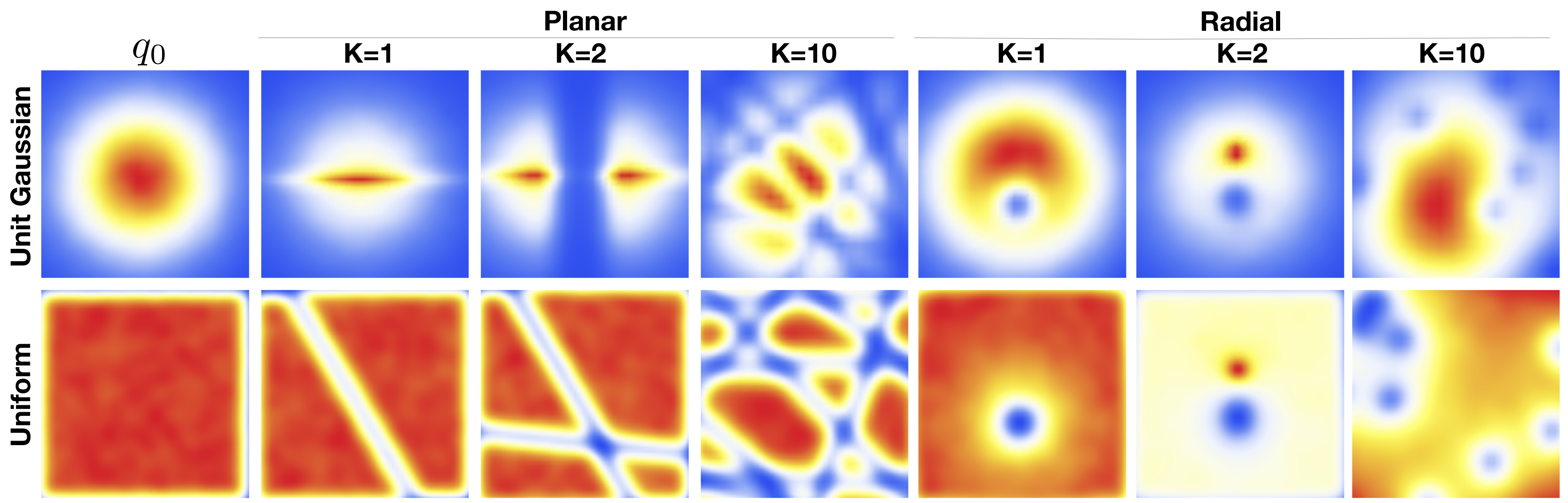
(Rezende & Mohamed 2016)

Variational Inference with Normalizing Flows

- The density $q_K(z)$ obtained by successively transforming a random variable z_0 with distribution q_0 through a chain of K transformations f_k is

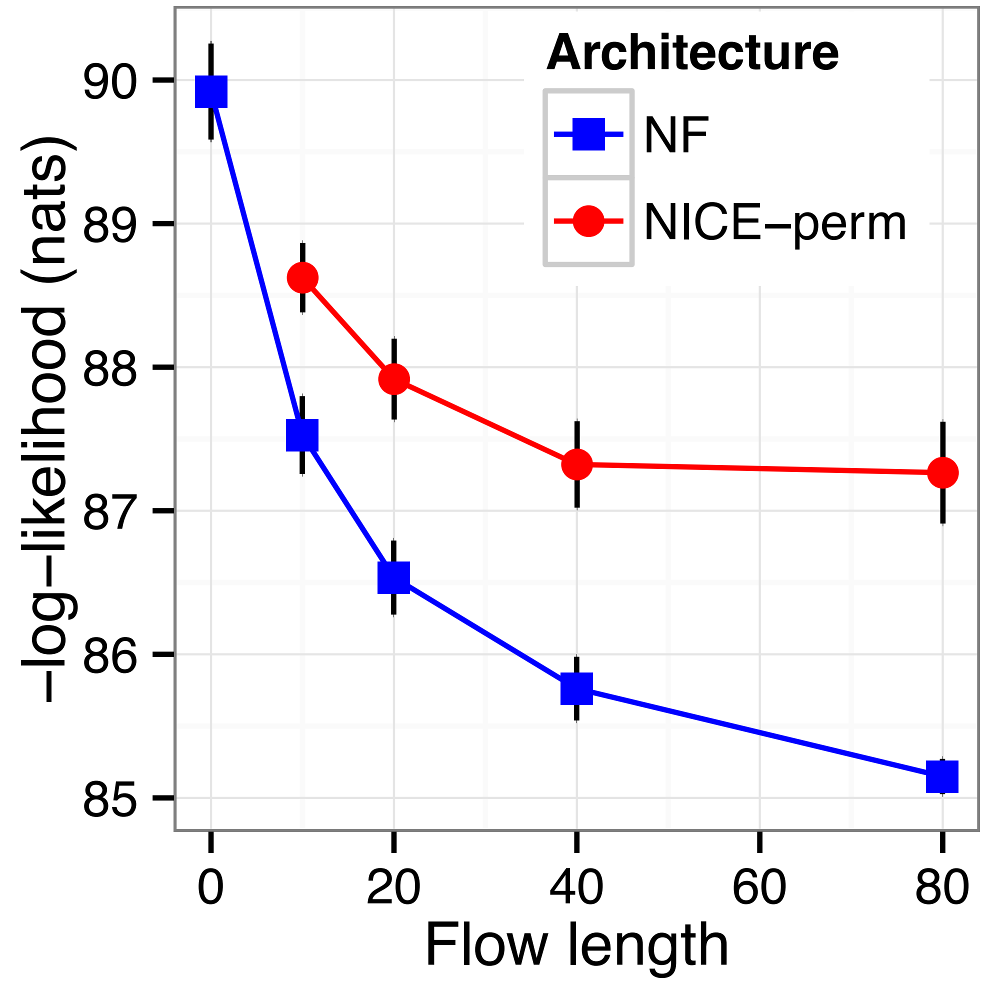
$$z_K = f_K \circ \ldots \circ f_2 \circ f_1(z_0)$$

$$q_K(z_K) = \log q_0(z_0) - \sum_{k=1}^{K} \log \left| \det \frac{\partial f_k}{\partial z_{k-1}} \right|$$



- Law of unconscious statistician (LOTUS)

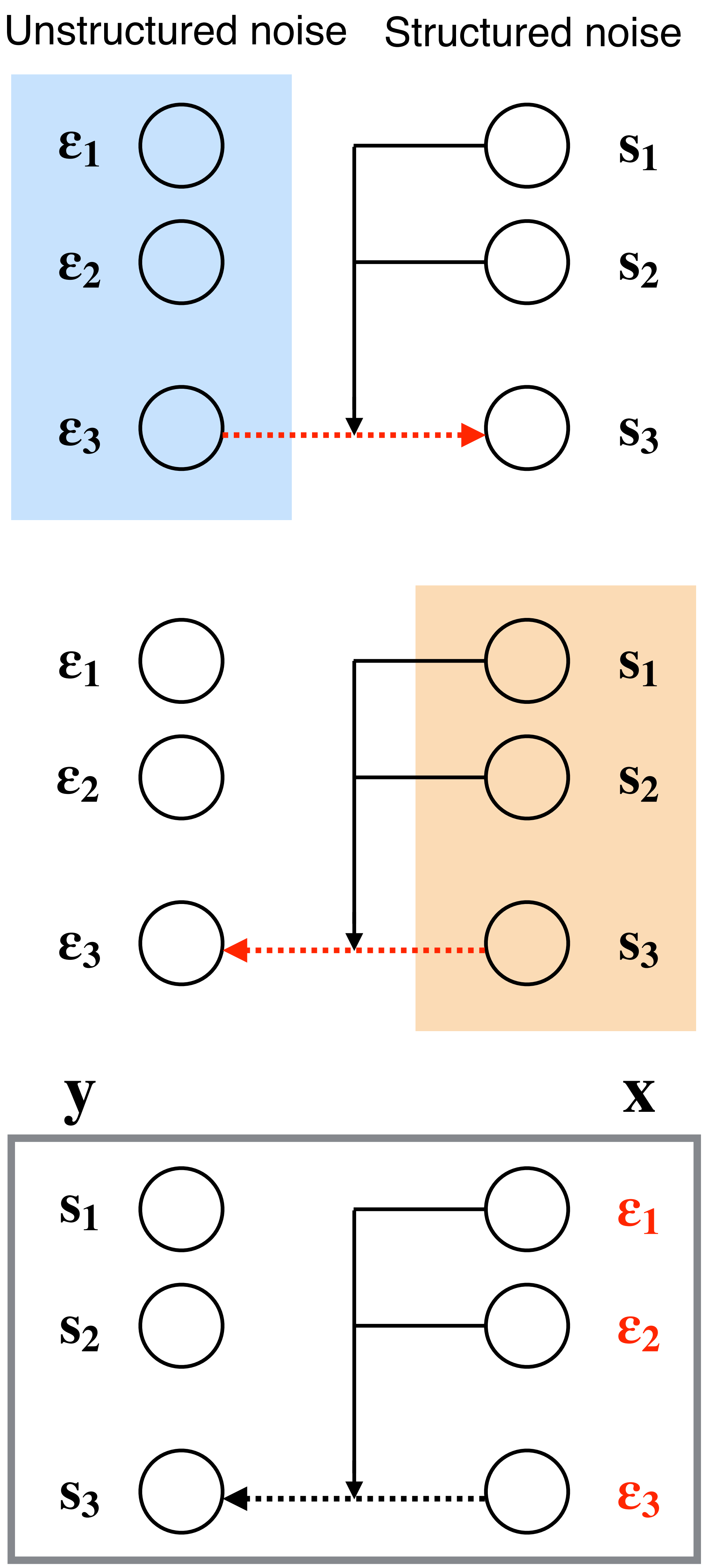
$$\mathbb{E}_{z_K}\left[h(z_K)\right] = \mathbb{E}_{z_0}\left[h(f_K \circ \ldots \circ f_2 \circ f_1(z_0))\right]$$



How to rewrite the ELBO objective?

(Rezende & Mohamed 2016)

Inverse Autoregressive Flows



$$y_1 = \mu_1(\emptyset) + \sigma_1(\emptyset) \cdot \epsilon_1$$

$$y_j = \mu_j(y_{1:j-1}) + \sigma_j(y_{1:j-1}) \cdot \epsilon_j$$

$$\epsilon_1 = \frac{y_1 - \mu_1(\emptyset)}{\sigma_1(\emptyset)}$$

$$\epsilon_j = \frac{y_j - \mu_j(y_{1:j-1})}{\sigma_j(y_{1:j-1})}$$

- Change of variable distribution

$$x \sim P_X(x); \quad y = f(x)$$

$$P_Y(y) = P_X(x) \left| \frac{\mathrm{d}f(x)}{\mathrm{d}x} \right|^{-1}$$

Autoregressive model

$$p(y_{1:P}) = p(y_1) \prod_{j=2}^{P} p(y_j | y_{1:j-1})$$

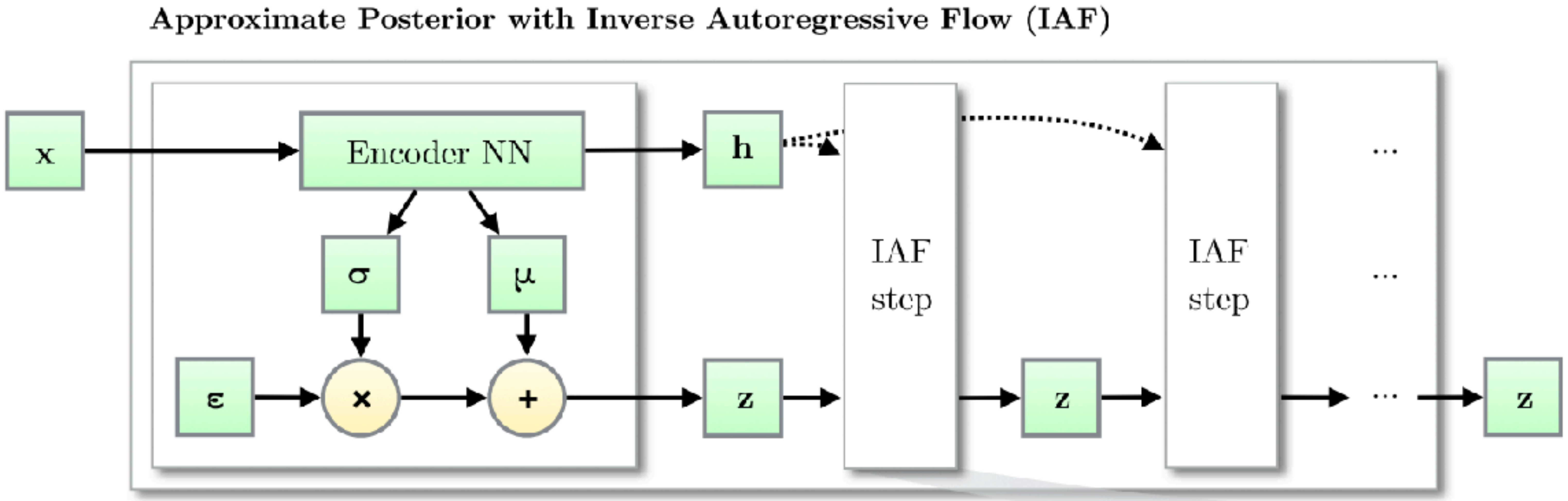
$$= \mathcal{N}(y_1; \mu_1, \sigma_1^2) \prod_{j=2}^{P} \mathcal{N}(y_j | \mu_j(y_{1:j-1}), \sigma_j^2(y_{1:j-1}))$$

- *Scales well to high-dimensional space in terms of complexity (no rank-one restriction)*
- *Linear complexity in computing log determinant of Jacobian (upper triangular Jacobian)*
- *No dependency in sampling (non-sequential)*
- *More expressive than pure autoregressive model (composable)*

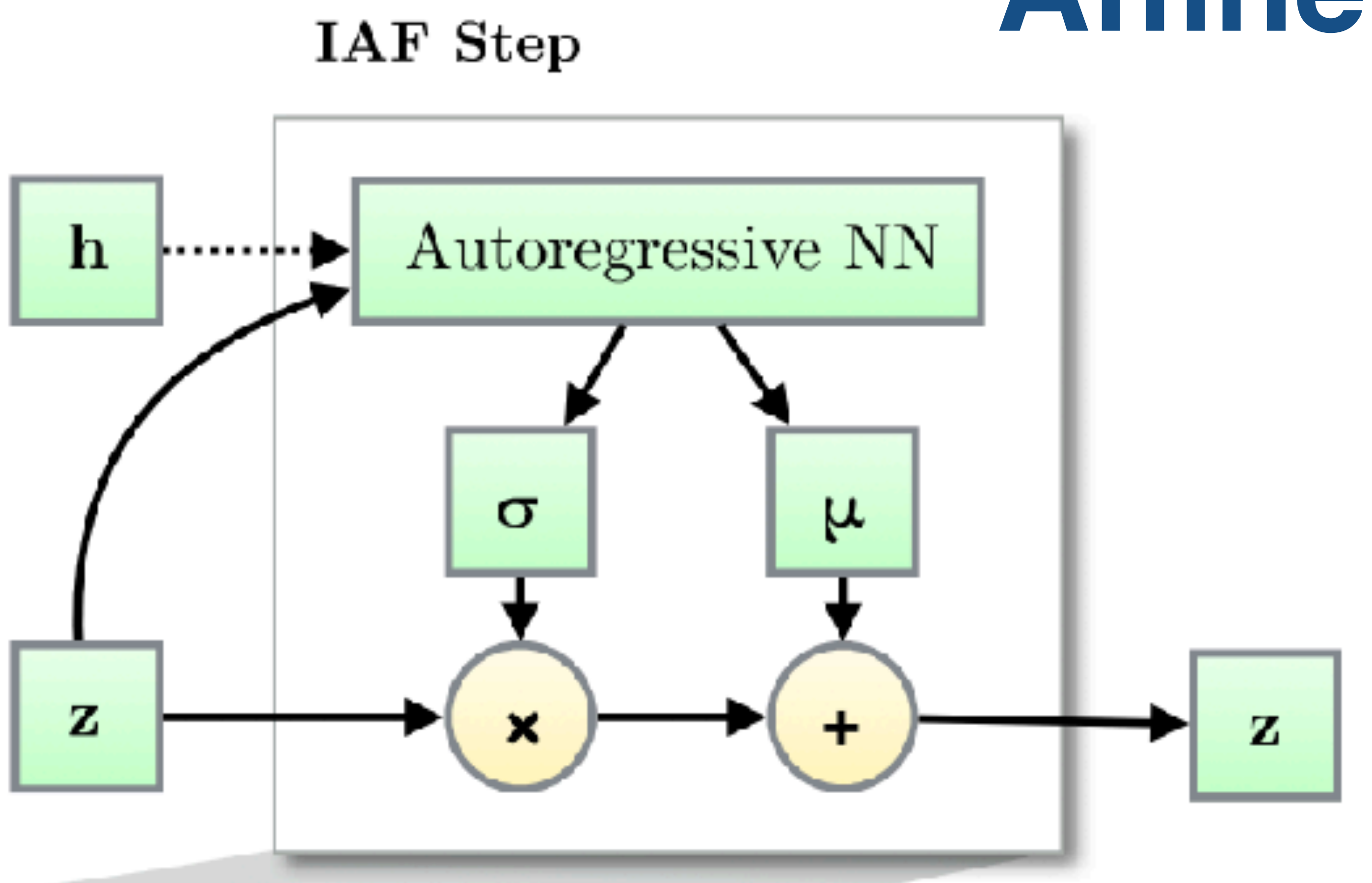
(Kingma et al., 2016; Huang et al., 2018)

Improving Variational Inference with IAF

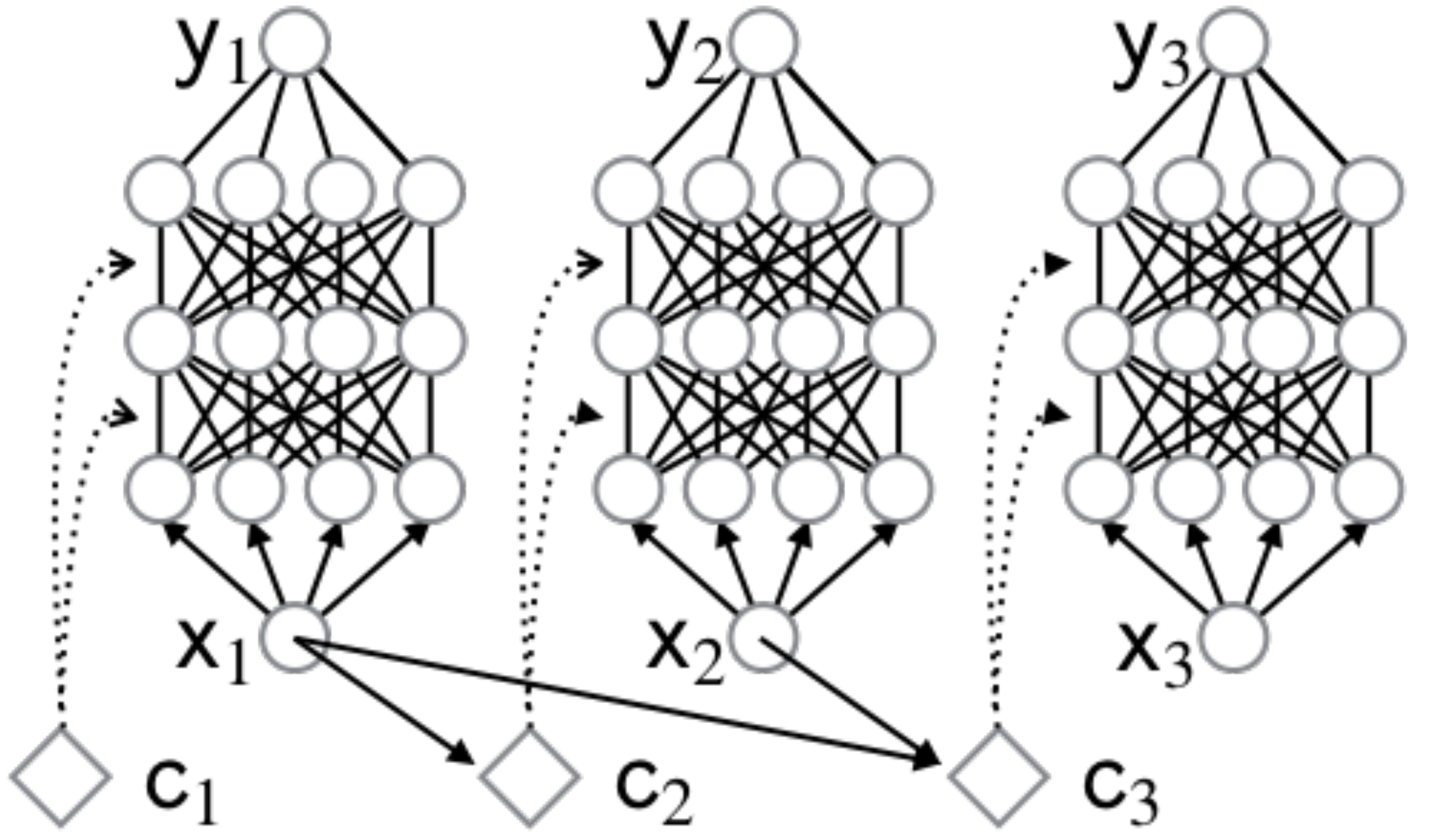
MILA



Affine



Non-affine



$$\epsilon \sim \mathcal{N}(0, \mathrm{I})$$

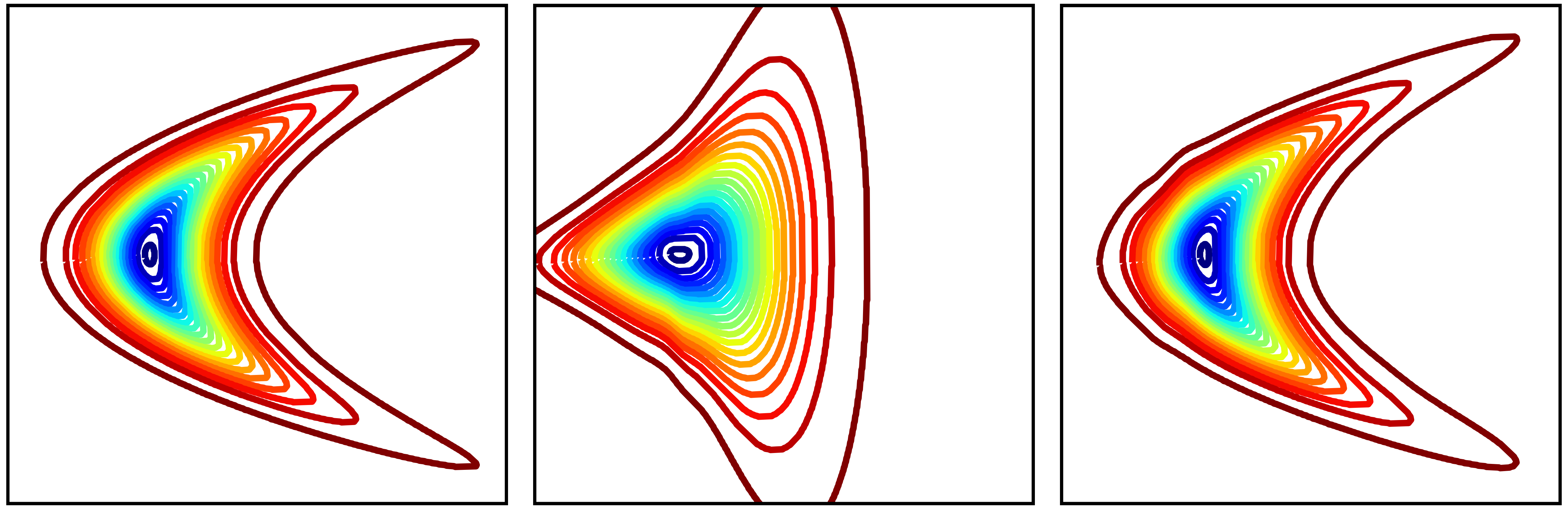
$$z_0 = \mu_0(x) + \sigma_0(x) \odot \epsilon$$

$$z_k = \mu_k(z_{k-1}) + \sigma_k(z_{k-1}) \odot z_{k-1}$$

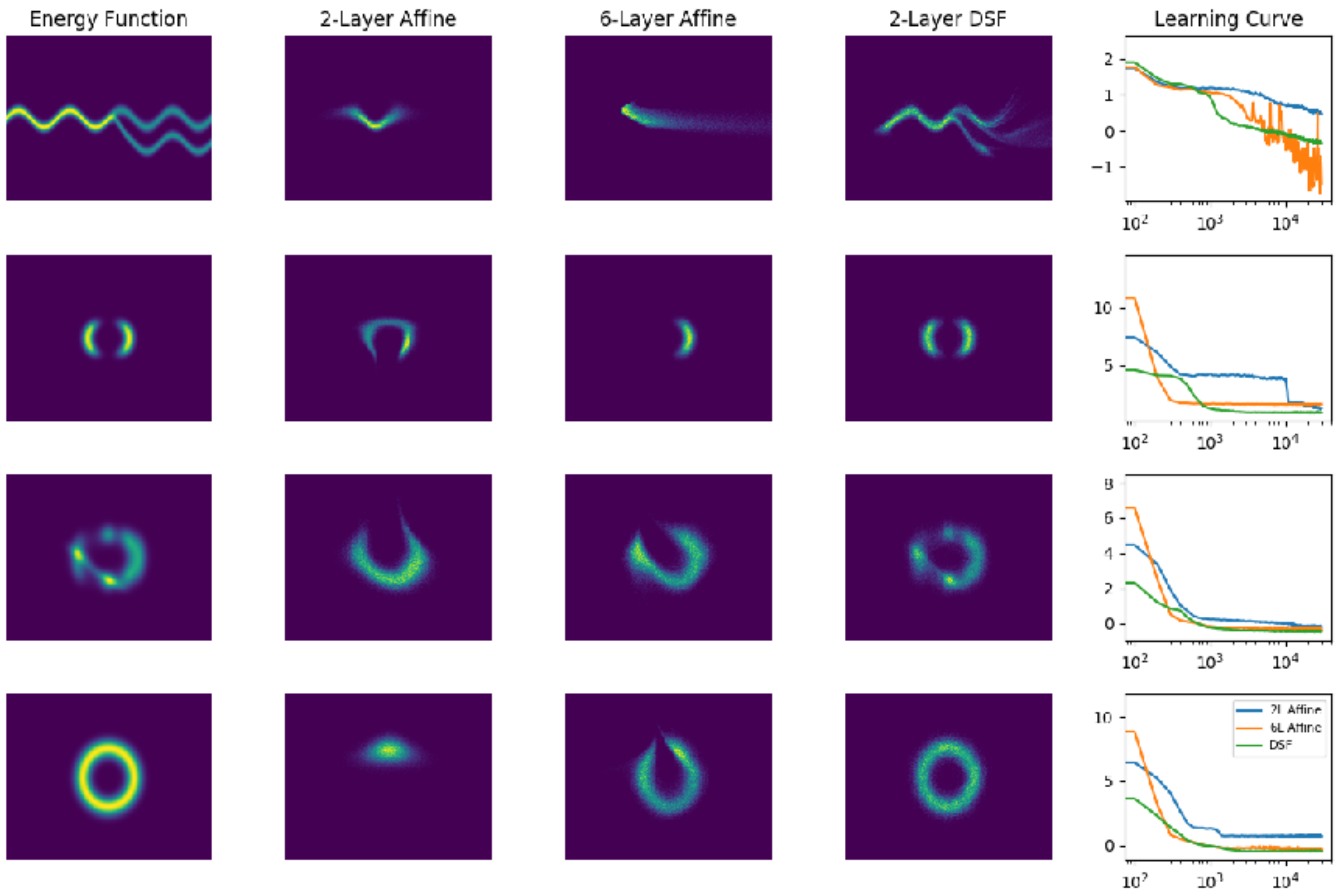
Method	bits/dim $\leq$
Multivariate Gaussian (van den Oord et al., 2016b)	4.70
NICE (Dinh et al., 2014)	4.48
Deep GMMs (van den Oord and Schrauwen, 2014)	4.00
Real NVP (Dinh et al., 2016)	3.49
PixelRNN (van den Oord et al., 2016b)	**3.00**
Gated PixelCNN (van den Oord et al., 2016c)	**3.03**
Results with variationally trained latent-variable models:	
Deep Diffusion (Sohl-Dickstein et al., 2015)	5.40
Convolutional DRAW (Gregor et al., 2016)	3.58
ResNet VAE with IAF (Ours)	**3.11**

(Kingma et al., 2016; Huang et al., 2018)

Expressiveness of Autoregressive FLows



Target MADE 5 MADEs



(Papamakarios, 2017)

(Huang et al., 2017-(b))

Importance Weighted Autoencoders

$$\mathcal{L}_k(x) = \mathbb{E}_{z_{1:k}\sim q(z|x)} \log \frac{1}{k} \sum_{i=1}^{k} \underbrace{\frac{p(x, z_i)}{q(z_i|x)}}_{w_i}$$

$$\nabla \mathcal{L}_k(x) = \mathbb{E}_{\epsilon_{1:k}} \left[\sum_{i=1}^{k} \tilde{w}_i \nabla \log w_i \right]$$

Theorem 1. *For all k, the lower bounds satisfy*

$$\log p(\mathbf{x}) \geq \mathcal{L}_{k+1} \geq \mathcal{L}_k.$$

Moreover, if $p(\mathbf{h}, \mathbf{x})/q(\mathbf{h}|\mathbf{x})$ *is bounded, then* $\mathcal{L}_k$ *approaches* $\log p(\mathbf{x})$ *as* k *goes to infinity.*

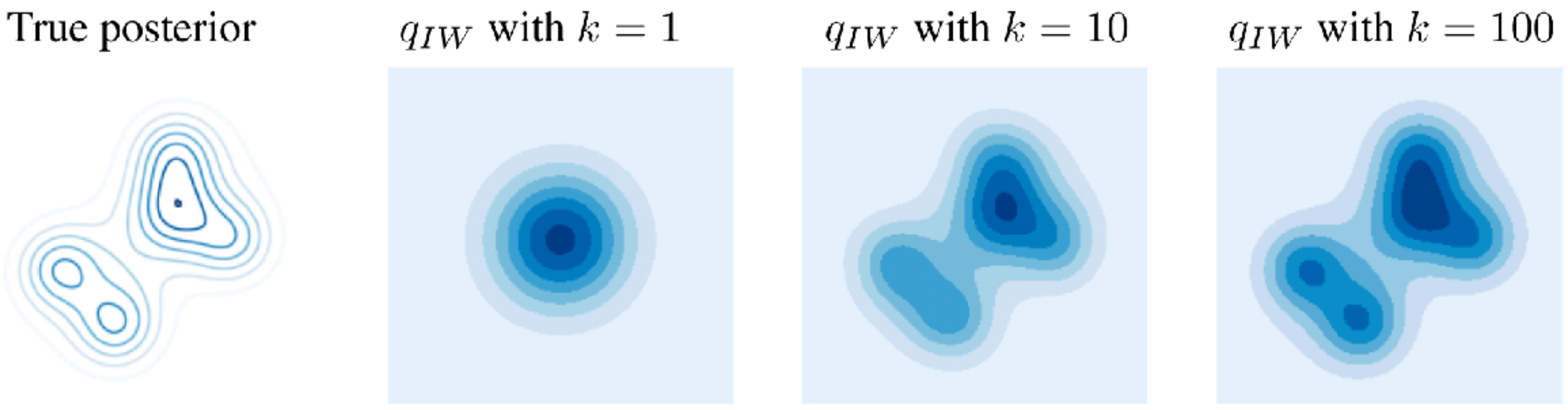
		MNIST			
		VAE		IWAE	
# stoch. layers	k	NLL	active units	NLL	active units
1	1	86.76	19	86.76	19
	5	86.47	20	85.54	22
	50	86.35	20	84.78	25
2	1	85.33	16+5	85.33	16+5
	5	85.01	17+5	83.89	21+5
	50	84.78	17+5	82.90	26+7

(Burda et al. 2016)

Iterative Refinement with Importance Sampling

$$q_{IW}(z|x) = E_{z_2 \ldots z_k \sim q(\cdot|x)} \left[\frac{p(x,z)}{\frac{1}{k}\left(\frac{p(x,z)}{q(z|x)} + \sum_{j=2}^{k} \frac{p(x,z_j)}{q(z_j|x)}\right)} \right]$$

$$\mathcal{L}_{VAE}[q_{IW}] = \cdots = E_{z_1 \ldots z_k \sim q(z|x)} \left[log \left(\frac{1}{k} \sum_{j=1}^{k} \frac{p(x,z_j)}{q(z_j|x)} \right) \right] = \mathcal{L}_{IWAE}[q]$$



(Cremer et al. 2017-(b))

Summary

We've talked about …

- Using latent variable models (deep latent gaussian models) as generative models;
  - Variational Bayes, ELBO
- Using an encoder to amortize inference
- Using reparameterization trick to reduce variance in gradient estimate; to make stochastic backward propagation possible
- How to improve model assumption using non fully factorized gaussian distribution as prior or likelihood
- How to improve inference using better inference network (e.g. larger capacity or meta learning)
- How to improve inference using transformation methods, or sampling methods

Reference

Reference

- Julien Despois, 2017. *Latent Space Visualization — Deep Learning Bits* (blog post).
- Michael Jordan et al., 1999. *An Introduction to Variational Methods for Graphical Models*.
- Martin Wainwright and Michael Jordan, 2008. *Graphical Models, Exponential Families, and Variational Inference*.
- David Blei et al., 2016. *Variational Inference: A Review for Statisticians*.
- Christ Cremer et al., 2017. *Inference Suboptimality in Variational Autoencoders.*
- Miriam Shiffman, 2016. *Under the Hood of the Variational Autoencoder* (blog post).
- Samuel Gershman and Noah Goodman, 2014. *Amortized Inference in Probabilistic Reasoning.*
- Diederik Kingma and Max Welling, 2014. *Auto-encoding Variational Bayes.*
- Rezende et al., 2014. *Stochastic Backpropagation and Approximate Inference in Deep Generative Models.*

Reference

- Michelins Titsias and Miguel Lazaro-Gredilla, 2014. *Doubly Stochastic Variational Bayes for non-Conjugate Inference*.
- Goker Erdogan, 2017. *Reparameterization Trick* (blog post).
- Irhum Shafkat. *Intuitively Understanding Variational Autoencoders* (blog post).
- Diederik Kingma, 2017. *Variational Inference and Deep Learning: A New Synthesis* (Ph.D. Thesis)
- Laurent Dinh and Vicent Dumoulin, 2014. *Training Neural Bayesian Nets* (slides).
- Radford Neal, 1998. *Annealed Importance Sampling*.
- Edward Challis and David Barber, 2012. *Affine Independent Variational Inference*.
- Ishaan Gulrajani et al., 2017. *PixelVAE: A Latent Variable Model for Natural Images*.
- Xi Chen et al., 2017. *Variational Lossy Autoencoder*.

Reference

- Chin-Wei Huang et al., 2017. *Learnable Explicit Density for Continuous Latent Space and Variational Inference*.
- Xianxu Hou et al., 2016. *Deep Feature Consistent Variational Autoencoder*.
- Matthew Hoffman and Matthew Johnson, 2016. *ELBO surgery: yet another way to carve up the variational evidence lower bound*.
- Diederik Cinema et al., 2014. *Semi-Supervised Learning with Deep Generative Models*.
- Nat Dilokthanakul et al., 2016. *Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders*.
- Iulian Serban et al., 2016. *Multi-Modal Variational Encoder-Decoders*.
- Rahul Krishnan et al., 2017. *On the challenges of learning with inference networks on sparse, high-dimensional data*.
- Joseph Marino et al., 2017. *Iterative Inference Models (Learning to Infer)*.

Reference

- Yoon Kim et al., 2018. *Semi-Amortized Variational Autoencoders*.
- Yuri Burda et al., 2016. *Importance Weighted Autoencoders*.
- Danilo Rezende and Shakir Mohamed, 2016. *Variational Inference with Normalizing Flows.*
- Diederik Kingma et al., 2016. *Improved Variational Inference with Inverse Autoregressive Flow*.
- Chin-Wei Huang et al., ~2018. *Neural Autoregressive Flows* (not published yet).
- George Papamakarios et al., 2017. *Masked Autoregressive Flow for Density Estimation*.
- Chin-Wei Huang et al., 2017-(b). *Facilitating Multimodality in Normalizing Flows*.
- Yuri Burda et al., 2016. *Importance Weighted Autoencoders*.
- Chris Cremer et al., 2017-(b). *Reinterpreting Importance-Weighted Autoencoders*.