Inference and Representation
Lecture 12

Joan Bruna
Courant Institute, NYU

Approximate Posterior Inference

- For most models, the posterior is analytically intractable:

$$p(z \mid x) = \frac{p(x \mid z)p(z)}{\int p(x \mid z')p(z')dz'}$$

- ***Variational Bayesian Inference:*** consider a parametric family of approximations $q(z \mid \beta)$ and optimize variational lower bound with respect to the variational parameters β .

Mean Field Variational Bayes

- Joint likelihood of observed and latent variables:

$$p(X, Z \mid \theta) \qquad \theta\text{: generative model parameters}$$

- Let us consider a posterior approximation $q(z|\beta)$of the form

$$q(z \mid \beta) = \prod_i q_i(z_i \mid \beta_i) \qquad \beta\text{: Variational parameters}$$

  - Mean-field approximation: we model hidden variables as being independent.

- Corresponding lower-bound is given by

$$\log p(X \mid \theta) \geq \int q(z \mid \beta) \log \frac{p(x, z \mid \theta)}{q(z \mid \beta)} dz = \mathbb{E}_{q(z|\beta)}\{\log(p(X, Z \mid \theta)\} + H(q(z \mid \beta))$$

Mean Field Variational Bayes

- **Goal**: optimize lower-bound with respect to variational parameters.
- As we have seen, this is equivalent to minimizing the divergence between true and approximate posterior:

$$\log p(X \mid \theta) = \widetilde{\mathcal{L}}(\theta, \beta) + D_{KL}(q_\beta(z)||p(z|x, \theta))$$

- If $q(z \mid \beta)$ is a factorial distribution, the entropy term is tractable:

$$H(q(z|\beta)) = \sum_i H(q_i(z_i|\beta_i))$$

- Problematic term: $\nabla_\beta \mathbb{E}_{q(z|\beta)} \log p(X, Z|\theta)$

Mean Field Variational Bayes

[Paiskey, Blei, Jordan,'12]

- Denote $f(Z) = \log p(X, Z|\theta)$
- Then

$$\nabla_\beta \mathbb{E}_{q(z|\beta)} f(Z) = \nabla_\beta \int f(z) q(z|\beta) dz$$

$$= \int f(z) \nabla_\beta q(z|\beta) dz$$

$$= \int f(z) q(z|\beta) \nabla_\beta \log q(z|\beta) dz$$

$$= \mathbb{E}_q \{ f(Z) \nabla_\beta \log q(z|\beta) \}$$

- Stochastic approximation of $\nabla_\beta \mathbb{E}_{q(z|\beta)} f(Z)$:

$$\nabla_\beta \mathbb{E}_{q(z|\beta)} f(Z) \approx \frac{1}{S} \sum_{s \leq S, z^{(s)} \sim q(z|\beta)} f(z^{(s)}) \nabla_\beta \log q(z^{(s)}|\beta)$$

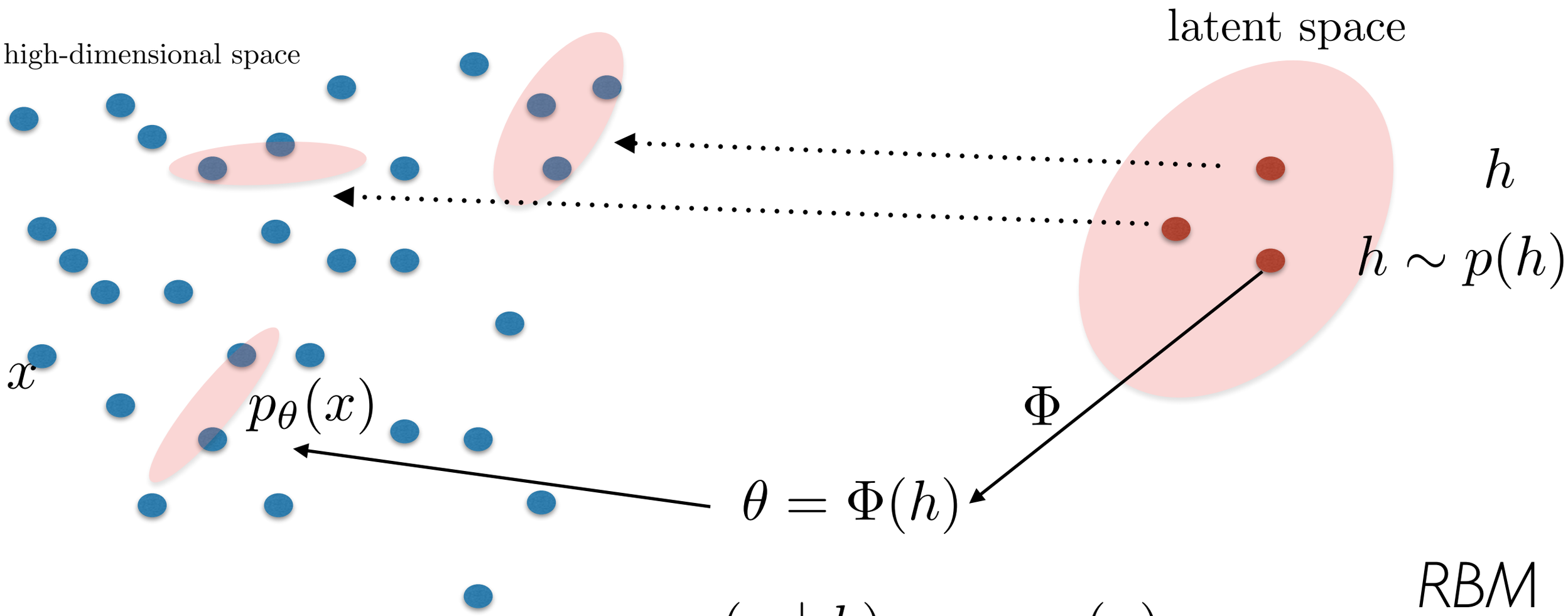
Mean Field Variational Bayes

- The estimator of the gradient is unbiased, but it may suffer from large variance.
  - We may need a large number S of samples to stabilize the descent.
  - This estimator is also the basis of policy gradients in RL.

- Faster alternative?

Latent Graphical Models

- Latent Graphical Models or *Mixtures*.



$$p(x \mid h) = p_{\Phi(h)}(x)$$

$$p(x) = \int p(x, h) dh = \int p(x \mid h) p(h) dh$$

RBM
DBN
DBM
VAE
...

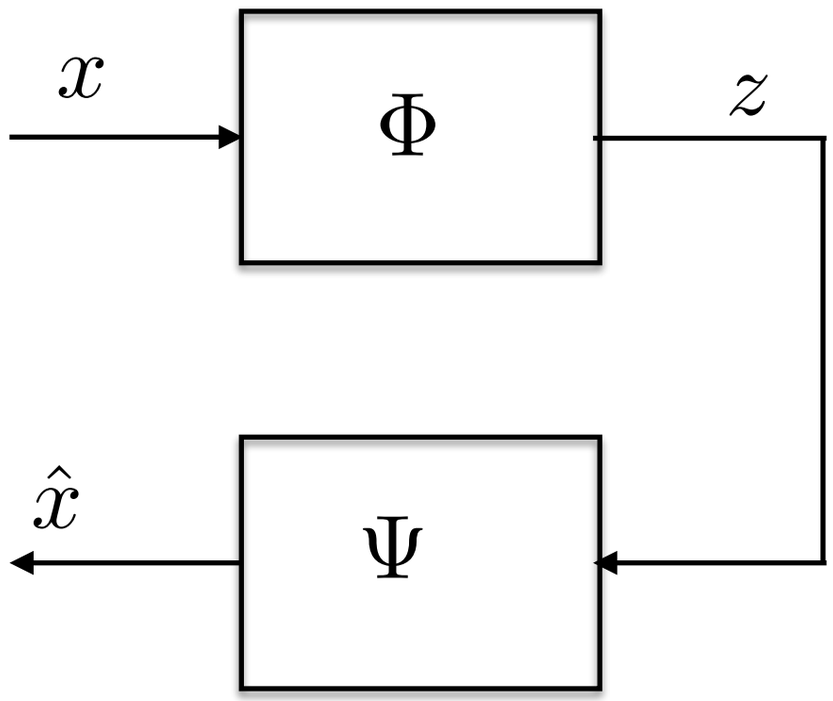
Model: additive combination of simple parametric models

Objectives lectures 12 and 13

- Auto encoders and manifold learning.
- Variational Autoencoders
- Variational Flows
- Generative Adversarial Networks
- Autoregressive Models

Auto encoders

- Goal: given data $X = \{x_i\}$ learn a reparametrization $z_i = \Phi(x_i)$ that approximates X well with minimal capacity.



- The model contains an encoder Φ and a decoder Ψ.
- It introduces an information bottleneck to characterize input data from ambient space.

Auto encoders

- Motivations
  - Dimensionality reduction:

$$x_i \in \mathbb{R}^d \ , \ \Phi : \mathbb{R}^d \rightarrow \mathbb{R}^{\tilde{d}} \ , \tilde{d} \ll d \ .$$

  - Metric learning (in sequential datasets):

$$z_t \approx \frac{1}{2}(z_{t-1} + z_{t+1})$$

linearization in transformed domain

Slow Feature Analysis

  - Unsupervised Pre-training (less popular nowadays): provide initial.

- Q: How to limit the reconstruction capacity?

Auto encoders

- Optimization set-up:

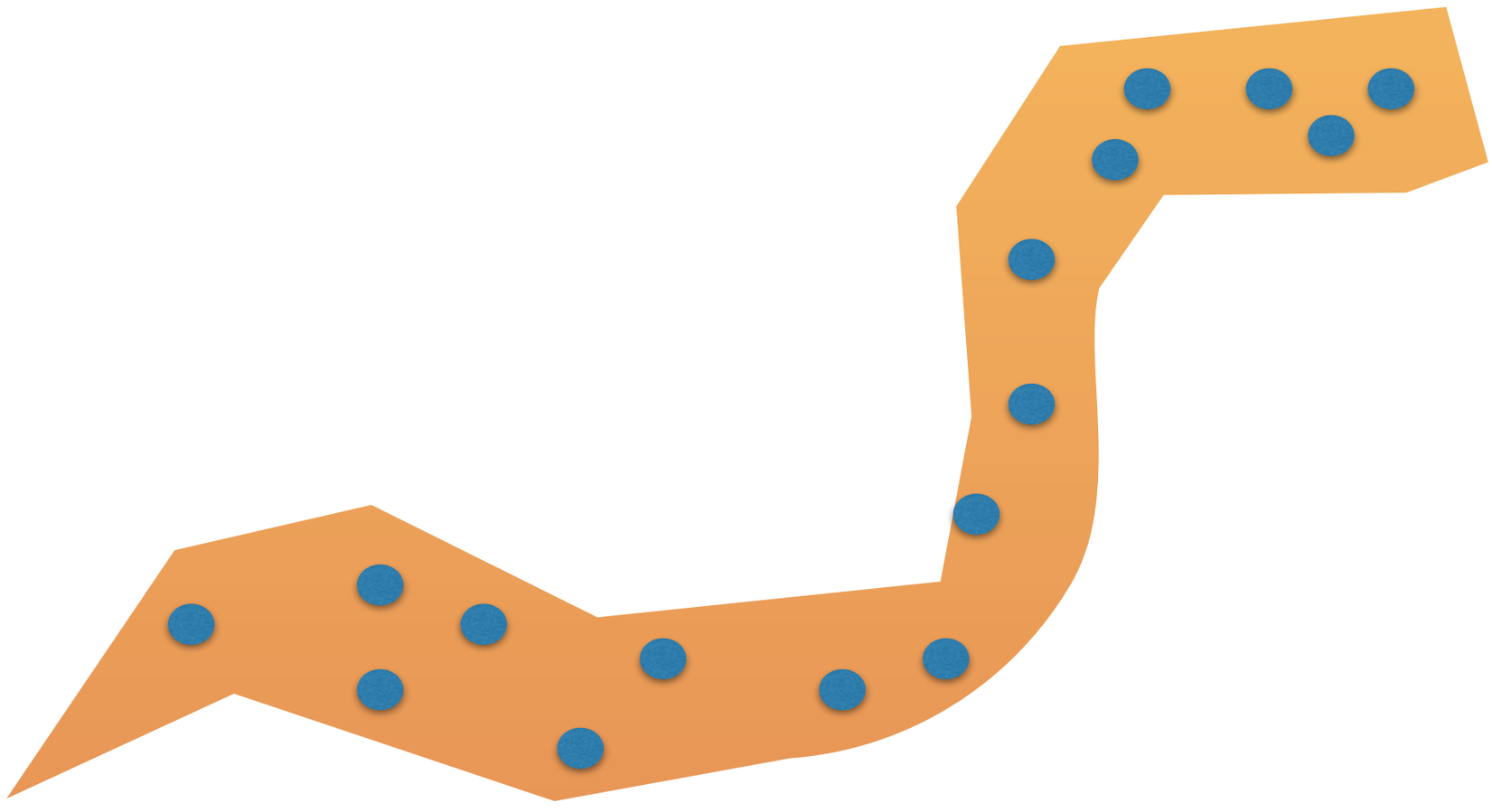
$$\min_{\Phi,\Psi} \frac{1}{n} \sum_{i \leq n} \ell\left(x_i, \Psi(\Phi(x_i))\right) + \mathcal{R}(\Phi(X))$$

$\ell(x, x')$: Reconstruction loss $\qquad$ $\mathcal{R}$: Regularization term

- Choice of models
  - Ψ Linear / Non-linear.
  - $\mathcal{R}(Z) = \|Z\|_1$ (or $\|Z\|_0$) leads to sparse auto-encoders (capacity can be measured by Gaussian Mean Width)
  - $\mathcal{R}(\Phi(x)) = \|\nabla \Phi(x)\|^2$ leads to contractive autoencoders.

Auto encoders: Geometric Interpretation

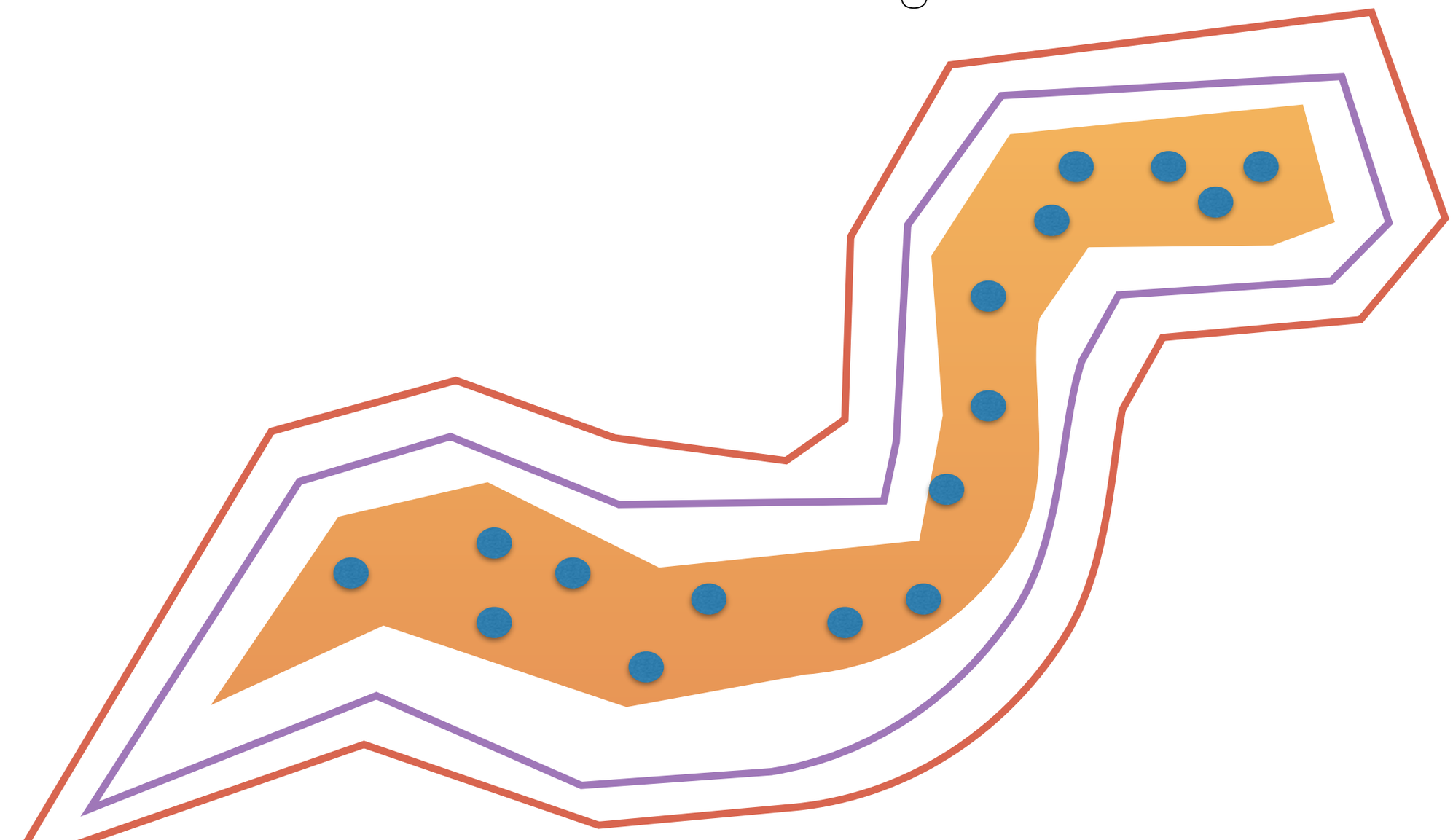
- The reconstruction error approximates a distance to a covering manifold of X



$$\Omega(\epsilon) = \{x \ s.t. \ \|\Psi(\Phi(x)) - x\| \leq \epsilon\}$$

Auto encoders: Geometric Interpretation

- The reconstruction error approximates a distance to a covering manifold of X.
- Intrinsic manifold coordinates "disentangle" factors.



$$\Omega(\epsilon) = \{x \ s.t. \ \|\Psi(\Phi(x)) - x\| \leq \epsilon\}$$

Examples

- Both encoder and decoder are linear
  - PCA

- Linear decoder, one-hot encoder
  - K-Means

- Linear decoder, sparse regularization
  - Dictionary Learning

More Examples

- Sparse Coding approximations

  - Predictive Sparse Decomposition (PSD) [Kavockoglu et al.,'08] considers an Augmented Lagrangian of the Sparse Autoencoder:

$$\min_{D,Z,\Phi} \|X - DZ\|^2 + \lambda\|Z\|_1 + \alpha\|Z - \Phi(X)\|^2$$

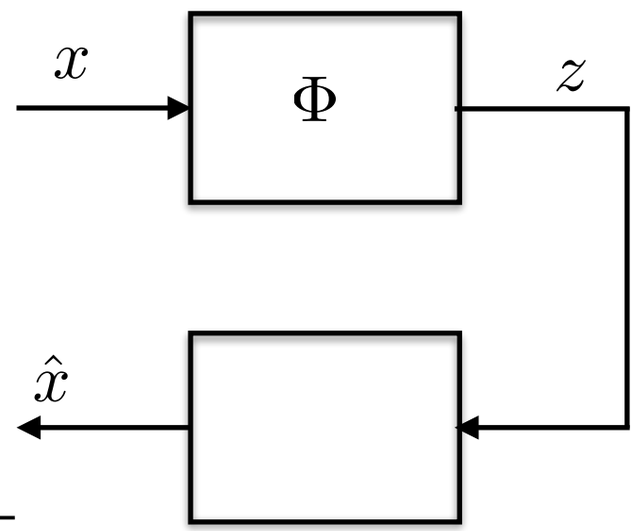
$$\Phi(X) = \mathrm{diag}(\beta)\tanh(WX + b)$$

  - LISTA [Gregor et al,'10]: Deeper Encoder using Recurrent weights.

Auto encoders: Probabilistic Interpretation

- We can also interpret z as latent variables of an underlying generative model for X:

$$p(x) = \int p(z)p(x \mid z)dz$$



- Rather than evaluating the true posterior

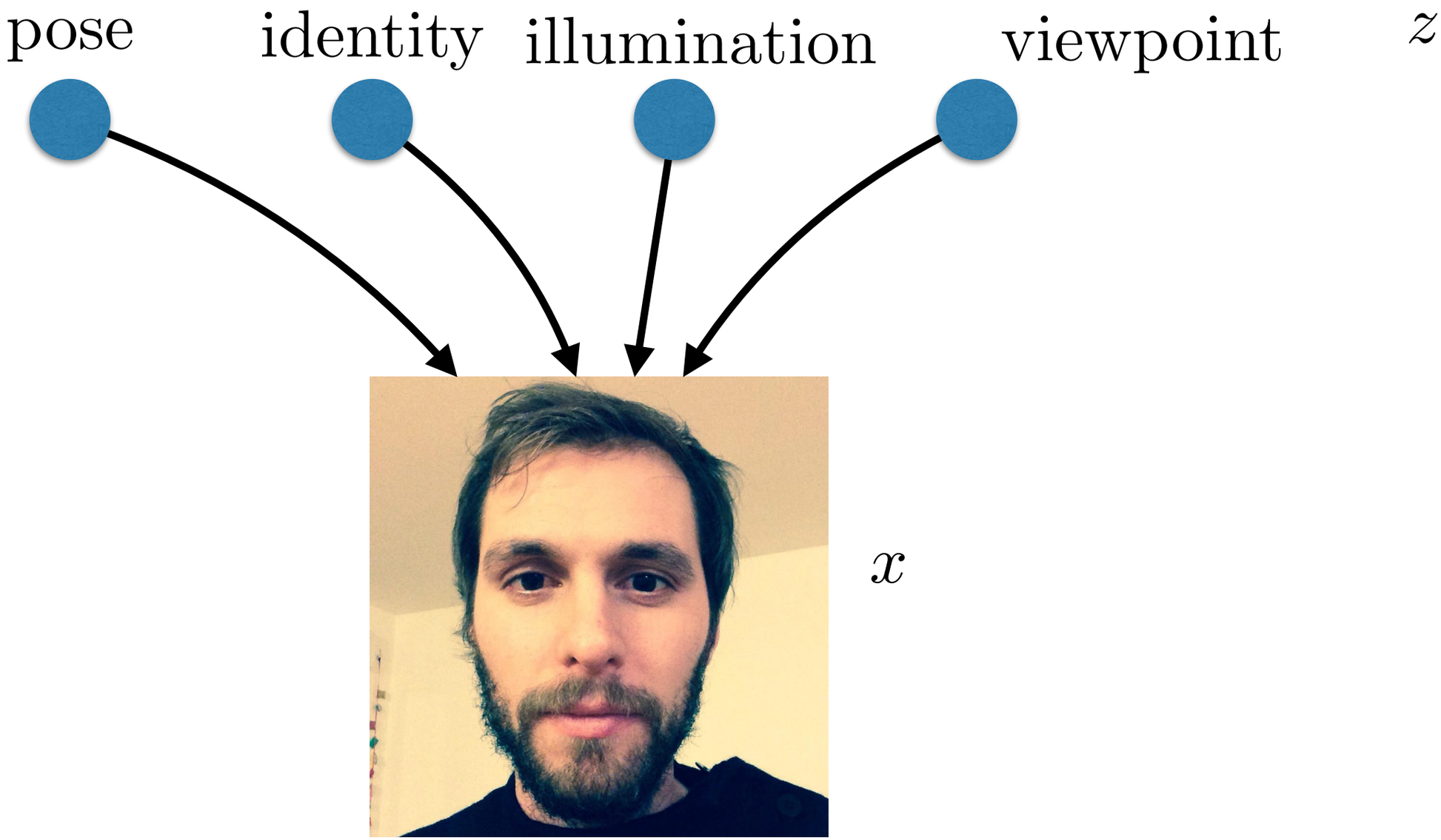
$$p(z \mid x) = \frac{p(z)p(x|z)}{\int p(z')p(x|z')dz'}$$

we consider a point estimate $p(z \mid x) = \delta(z - \Phi(x))$

- Q: How to perform "correct" posterior inference? or a better approximation?

Approximate Posterior Inference

- In latent graphical models, we can interpret latent variables as factors:



How to infer z given x ?

Variational Autoencoders

[Kingma & Welling'14, Rezende et al.'14]

- Recall the variational lower bound:

$$\log p(X \mid \theta) = \mathbb{E}_{q(z|\beta)}\{\log(p(X, Z \mid \theta)\} + H(q(z \mid \beta)) + D_{KL}(q(z|\beta)||p(z|x, \theta)$$

$$\log p(X \mid \theta) = \mathcal{L}(\theta, \beta, X) + D_{KL}(q(z|\beta)||p(z|X, \theta))$$

Variational Autoencoders

[Kingma & Welling'14, Rezende et al.'14]

- Recall the variational lower bound:

$$\log p(X \mid \theta) = \mathbb{E}_{q(z|\beta)}\{\log(p(X, Z \mid \theta)\} + H(q(z \mid \beta)) + D_{KL}(q(z|\beta)||p(z|x, \theta)$$

$$\log p(X \mid \theta) = \mathcal{L}(\theta, \beta, X) + D_{KL}(q(z|\beta)||p(z|X, \theta))$$

- Can we optimize jointly both generative and variational parameters efficiently?

Variational Autoencoders

[Kingma & Welling'14, Rezende et al.'14]

- Recall the variational lower bound:

$$\log p(X \mid \theta) = \mathbb{E}_{q(z|\beta)}\{\log(p(X, Z \mid \theta)\} + H(q(z \mid \beta)) + D_{KL}(q(z|\beta)||p(z|x, \theta)$$

$$\log p(X \mid \theta) = \mathcal{L}(\theta, \beta, X) + D_{KL}(q(z|\beta)||p(z|X, \theta))$$

- Can we optimize jointly both generative and variational parameters efficiently?
- For appropriate posterior approximations, we can reparametrize samples as

$$Z \sim q(z|x, \beta) \Rightarrow Z \stackrel{d}{=} g_\beta(\epsilon, x) \ , \ \epsilon \sim p_0$$

$$\left(\text{e.g. } \ q(z|x, \beta) = \mathcal{N}(z; \mu(x), \Sigma(x)) \leftrightarrow z = \mu(x) + \Sigma(x)^{1/2}\epsilon \ , \ \epsilon \sim \mathcal{N}(0, \mathbf{1})\right)$$

Variational Autoencoders

- It results that

$$\mathcal{L}(\theta, \beta, X) = -D_{KL}(q_\beta(z|X)||p_\theta(z)) + \mathbb{E}_{q_\beta(z|X)}\{\log p(X|z, \theta)\}$$

can be estimated via Monte-Carlo by

$$\widehat{\mathcal{L}(\theta, \beta, X)} = -D_{KL}(q_\beta(z|X)||p_\theta(z)) + \frac{1}{S}\sum_{s \leq S} \log p(X|z^{(s)}, \theta)$$

$$z^{(s)} = g_\beta(X, \epsilon^{(s)}) \text{ and } \epsilon^{(s)} \sim p_0 \ .$$

Variational Autoencoders

- It results that

$$\mathcal{L}(\theta, \beta, X) = -D_{KL}(q_\beta(z|X)||p_\theta(z)) + \mathbb{E}_{q_\beta(z|X)}\{\log p(X|z, \theta)\}$$

can be estimated via Monte-Carlo by

$$\widehat{\mathcal{L}(\theta, \beta, X)} = -D_{KL}(q_\beta(z|X)||p_\theta(z)) + \frac{1}{S}\sum_{s \leq S} \log p(X|z^{(s)}, \theta)$$

$$z^{(s)} = g_\beta(X, \epsilon^{(s)}) \text{ and } \epsilon^{(s)} \sim p_0 \ .$$

- First term acts as a regularizer: limits the capacity of the encoder
- Second term is a reconstruction error.

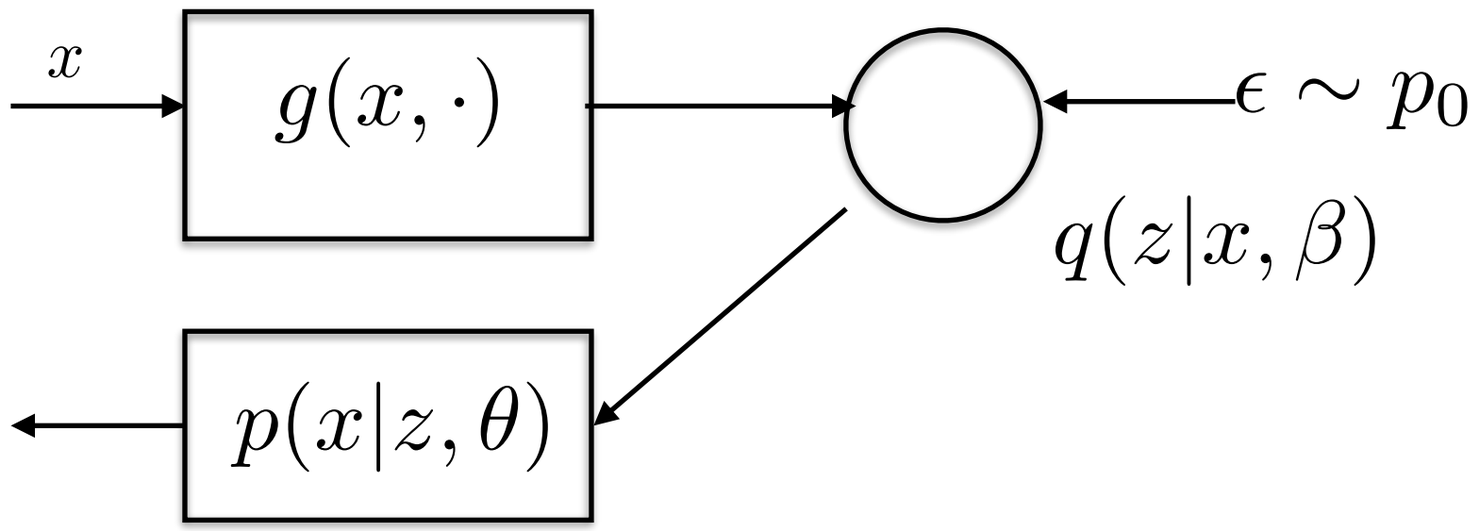
Variational Autoencoders

- How to model $x \mapsto g_\beta(x, \cdot)$ and $z \mapsto p_\theta(\cdot, z)$?

Variational Autoencoders

- How to model $x \mapsto g_\beta(x, \cdot)$ and $z \mapsto p_\theta(\cdot, z)$?

- VAE idea: use neural networks to approximate variational and generative parameters.



Variational Autoencoder

- Example: Let the prior over latent variables be Gaussian isotropic:

$$p(z) = \mathcal{N}(z; 0, \mathbf{I})$$

Variational Autoencoder

- Example: Let the prior over latent variables be Gaussian isotropic:

$$p(z) = \mathcal{N}(z; 0, \mathbf{I})$$

- Let the conditional likelihood be also Gaussian:

$$p(x|z) = (x; \mu(z), \Sigma(z)) \qquad \mu(z), \Sigma(z) \ : \ \text{Neural networks}$$

Variational Autoencoder

- Example: Let the prior over latent variables be Gaussian isotropic:

$$p(z) = \mathcal{N}(z; 0, \mathbf{I})$$

- Let the conditional likelihood be also Gaussian:

$$p(x|z) = (x; \mu(z), \Sigma(z)) \qquad \mu(z), \Sigma(z) \ : \ \text{Neural networks}$$

- Variational approximate posterior also Gaussian:

$$q_\beta(z|x) = \mathcal{N}(z; \overline{\mu}(x), \overline{\Sigma}(x))$$

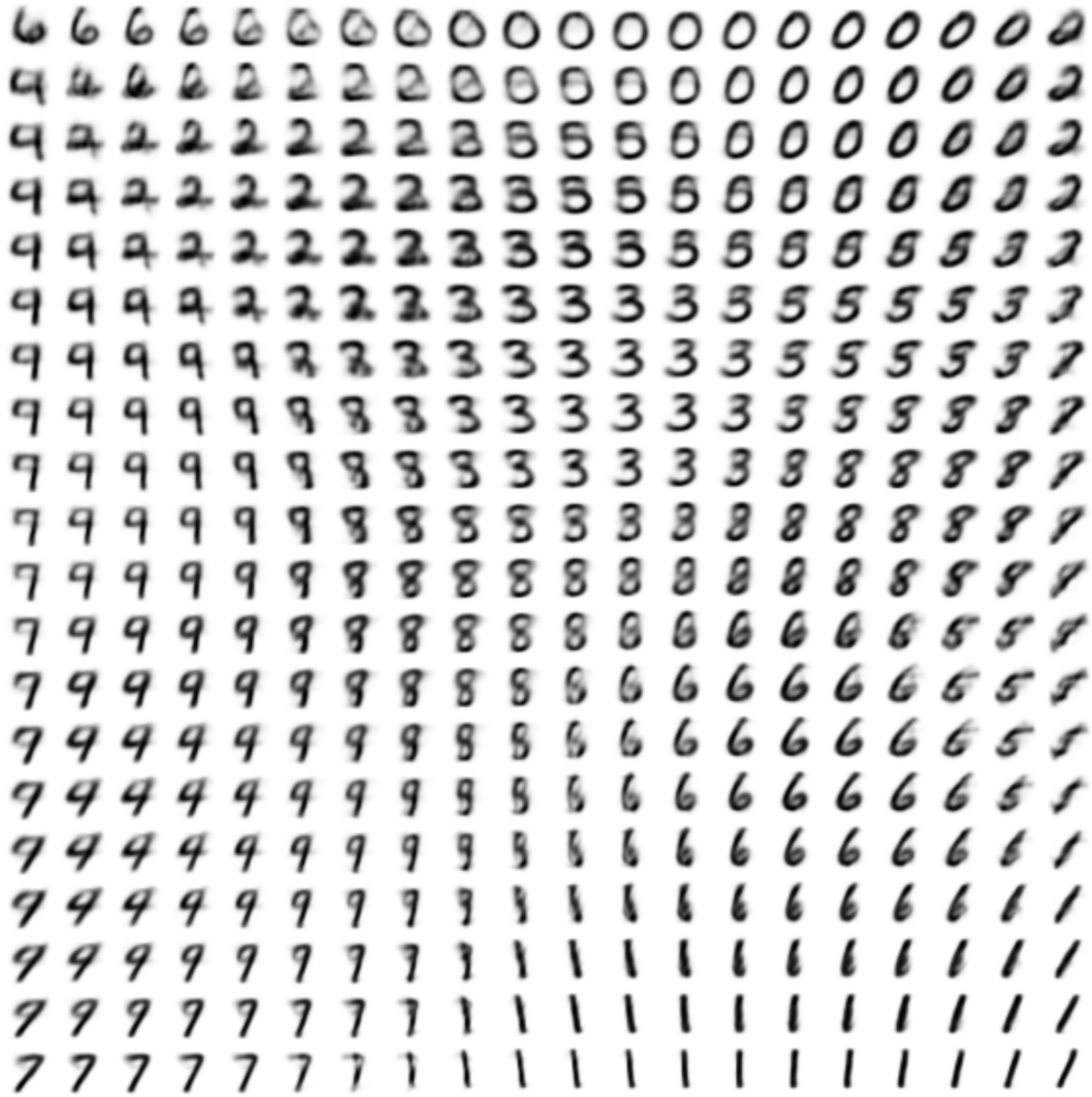
$$\overline{\mu}(z), \overline{\Sigma}(z) \ : \ \text{Neural networks}, (\overline{\Sigma} \text{ diagonal})$$

$$Z \sim q_\beta(z|x) \Leftrightarrow Z = \overline{\mu}(x) + \overline{\Sigma}(x)^{1/2}\epsilon \ , \ \epsilon \sim \mathcal{N}(0, 1)$$

Variational Autoencoder

- Examples using a two-dimensional latent space:

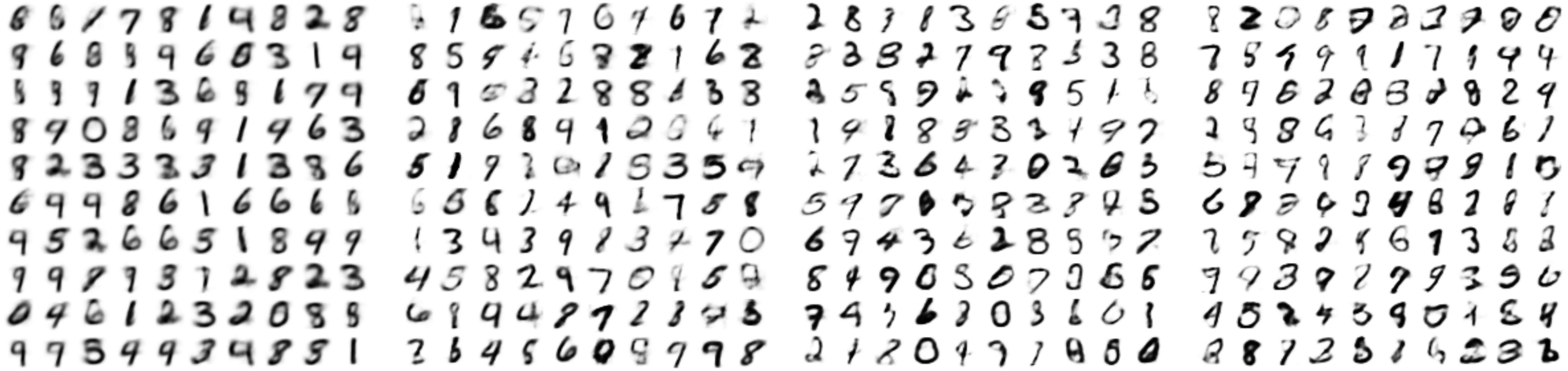
(a) Learned Frey Face manifold



(b) Learned MNIST manifold

Examples

- Increasing latent dimensionality:



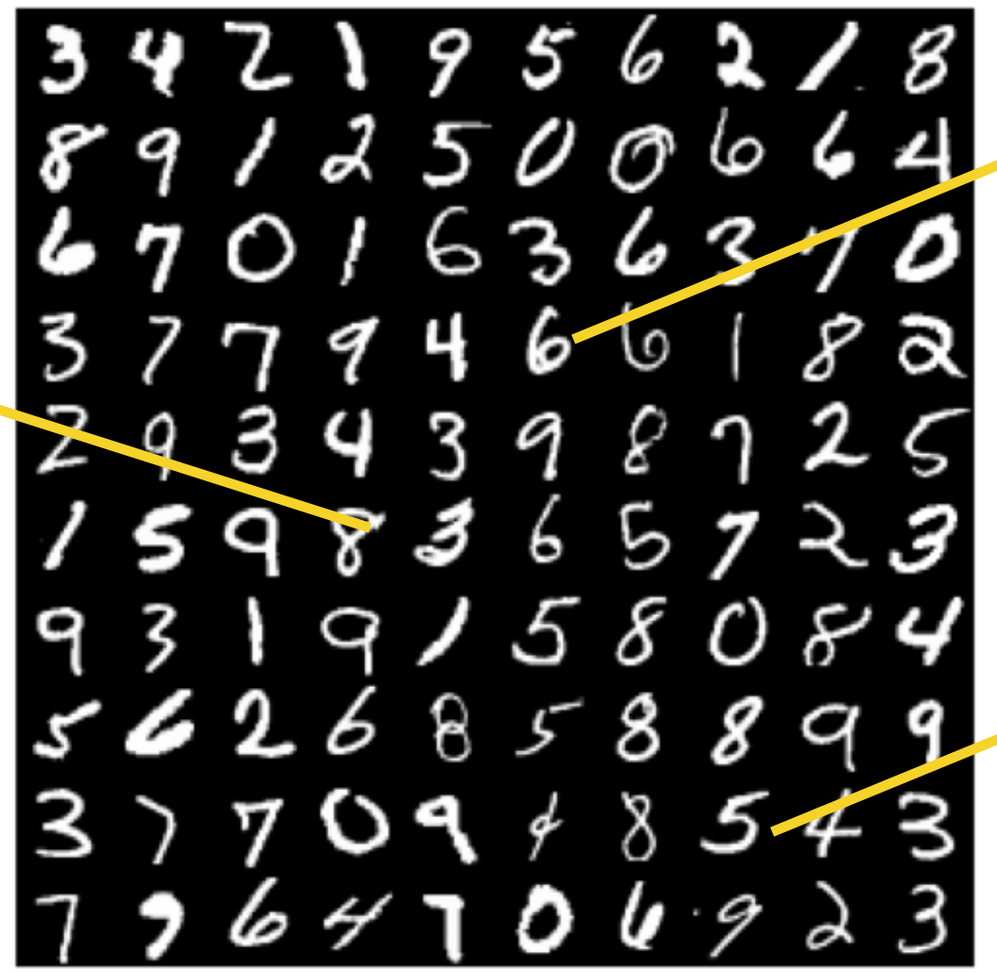
(a) 2-D latent space (b) 5-D latent space (c) 10-D latent space (d) 20-D latent space

Extensions to semi-supervised learning

- Semi-supervised learning:

We observe $\{x_i\}_{i \leq L_1}$ and $\{x_j, y_j\}_{j \leq L_2}$, with $x_i \sim p(x)$, $x_j \sim p(x)$.

$L_1 \gg L_2$

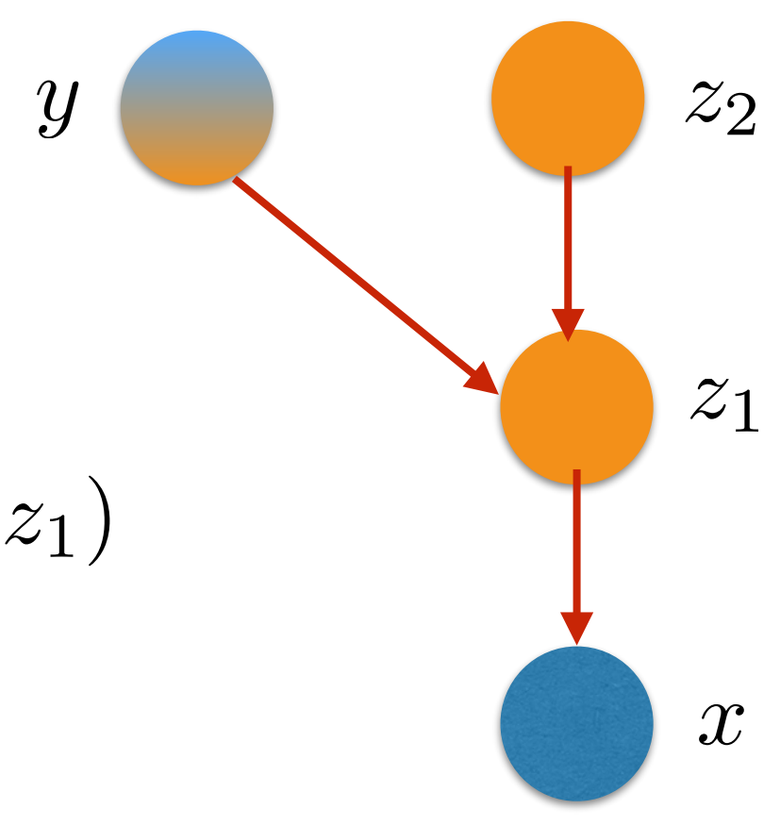


Extension to Semi-Supervised Learning

- "Semi-supervised Learning with Deep Generative Networks", Kingma et al,'14.

- Labels are treated as either observed or hidden.

$$p(x, y, z_1, z_2) = p(y)p(z_2)p(z_1|z_2, y)p(x|z_1)$$



Extension to Semi-Supervised Learning

- *"Semi-supervised Learning with Deep Generative Networks", Kingma et al,'14.*

- For datapoint with labels:

$$\log p_\theta(x, y) \geq \mathbb{E}_{q_\beta(z|x,y)} \left(\log p_\theta(x|y, z) + \log p_\theta(y) + \log p(z) - \log q_\beta(z|x, y) \right)$$

- For datapoint with no labels:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\beta(y,z|x)} \left(\log p_\theta(x|y, z) + \log p_\theta(y) + \log p(z) - \log q_\beta(z, y|x) \right)$$

Extension to Semi-Supervised Learning

- *"Semi-supervised Learning with Deep Generative Networks", Kingma et al,'14.*
- Classification results on MNIST:

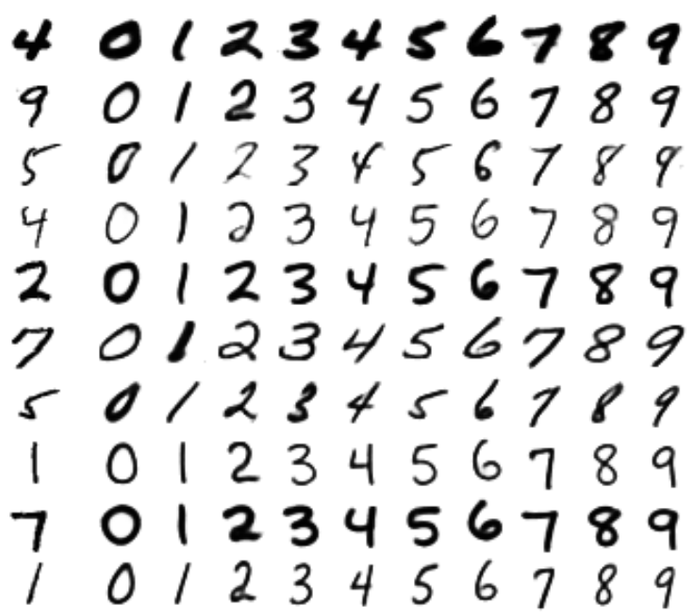
Table 1: Benchmark results of semi-supervised classification on MNIST with few labels.

N	NN	CNN	TSVM	CAE	MTC	AtlasRBF	M1+TSVM	M2	M1+M2
100	25.81	22.98	16.81	13.47	12.03	8.10 ($\pm$ 0.95)	11.82 ($\pm$ 0.25)	11.97 ($\pm$ 1.71)	**3.33** ($\pm$ 0.14)
600	11.44	7.68	6.16	6.3	5.13	–	5.72 ($\pm$ 0.049)	4.94 ($\pm$ 0.13)	**2.59** ($\pm$ 0.05)
1000	10.7	6.45	5.38	4.77	3.64	3.68 ($\pm$ 0.12)	4.24 ($\pm$ 0.07)	3.60 ($\pm$ 0.56)	**2.40** ($\pm$ 0.02)
3000	6.04	3.35	3.45	3.22	2.57	–	3.49 ($\pm$ 0.04)	3.92 ($\pm$ 0.63)	**2.18** ($\pm$ 0.04)

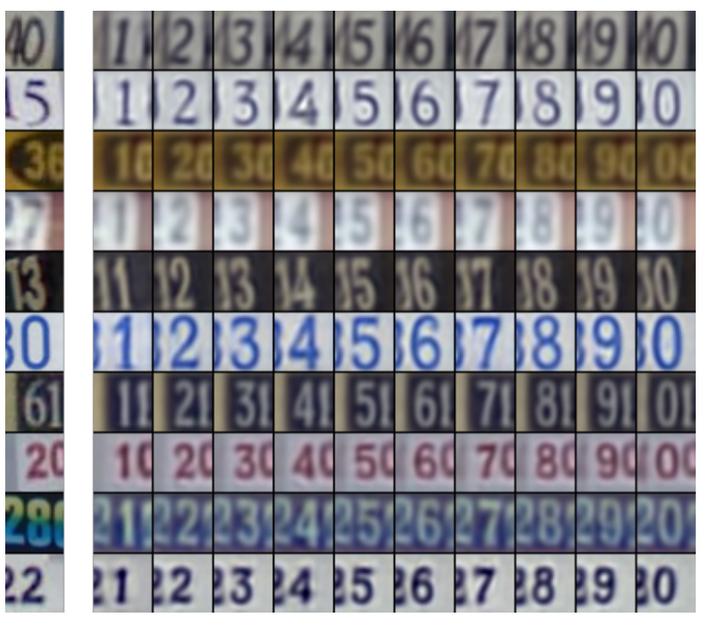
Extension to Semi-Supervised Learning

- *"Semi-supervised Learning with Deep Generative Networks", Kingma et al,'14.*
- Disentangling label and "style":

(a) Handwriting styles for MNIST obtained by fixing the class label and varying the 2D latent variable **z**



(b) MNIST analogies



(c) SVHN analogies

Incorporate MCMC to posterior approx.

"Markov Chain Monte Carlo and Variational Inference: Bridging the Gap", Salimans et al'15

- We saw in Lecture 7 how to use Markov Chains to approximate intractable posteriors.

$$p(z \mid x) \stackrel{d}{=} \lim_{T \to \infty} q_0(z_0 \mid x) \prod_{t \leq T} q(z_t \mid z_{t-1}, x) \ .$$

Incorporate MCMC to posterior approx.

"Markov Chain Monte Carlo and Variational Inference: Bridging the Gap", Salimans et al'15

- We saw in Lecture 7 how to use Markov Chains to approximate intractable posteriors.

$$p(z \mid x) \stackrel{d}{=} \lim_{T \to \infty} q_0(z_0 \mid x) \prod_{t \leq T} q(z_t \mid z_{t-1}, x) \ .$$

- For fixed T, this can be seen as another variational approximation, by considering $y = z_1, \ldots, z_{T-1}$ as extra hidden variables.

Incorporate MCMC to posterior approx.

"Markov Chain Monte Carlo and Variational Inference: Bridging the Gap", Salimans et al'15

- We saw in Lecture 7 how to use Markov Chains to approximate intractable posteriors.

$$p(z \mid x) \stackrel{d}{=} \lim_{T \to \infty} q_0(z_0 \mid x) \prod_{t \leq T} q(z_t \mid z_{t-1}, x) \ .$$

- For fixed T, this can be seen as another variational approximation, by considering $y = z_1, \ldots, z_{T-1}$ as extra hidden variables.
- The resulting Variational Lower bound becomes

$$\begin{aligned} \mathcal{L}_{MCMC} &= \mathcal{L} - \mathbb{E}_{q(z_T|x)}\{D_{KL}(r(y|z_T, x)||q(y \mid z_T, x)\} \\ &\leq \mathcal{L} \leq \log p(x) \ . \end{aligned}$$

$r(y|x, z_T)$: auxiliary variational approximation

Incorporate MCMC to posterior approx.

"Markov Chain Monte Carlo and Variational Inference: Bridging the Gap", Salimans et al'15

- We saw in Lecture 7 how to use Markov Chains to approximate intractable posteriors.

$$p(z \mid x) \stackrel{d}{=} \lim_{T \to \infty} q_0(z_0 \mid x) \prod_{t \leq T} q(z_t \mid z_{t-1}, x) \ .$$

- For fixed T, this can be seen as another variational approximation, by considering $y = z_1, \ldots, z_{T-1}$ as extra hidden variables.
- The resulting Variational Lower bound becomes

$$\begin{aligned} \mathcal{L}_{MCMC} &= \mathcal{L} - \mathbb{E}_{q(z_T|x)}\{D_{KL}(r(y|z_T, x)||q(y \mid z_T, x)\} \\ &\leq \mathcal{L} \leq \log p(x) \ . \end{aligned}$$

$r(y|x, z_T)$: auxiliary variational approximation

- If we choose r to be an inverse Markov chain, we obtain

$$\mathcal{L}_{aux} = \mathbb{E}_q \left\{\log p(x, z_T) - \log q(z_0|x)\right\} + \sum_{t=1}^{T} \left(\log r_t(z_{t-1}|x, z_t) - \log q_t(z_t|x, z_{t-1})\right)$$

Incorporate MCMC to posterior approx.

"Markov Chain Monte Carlo and Variational Inference: Bridging the Gap", Salimans et al'15

$$\mathcal{L}_{aux} = \mathbb{E}_q \left\{ \log p(x, z_T) - \log q(z_0|x) \right\} + \sum_{t=1}^{T} \left(\log r_t(z_{t-1}|x, z_t) - \log q_t(z_t|x, z_{t-1}) \right)$$

- The authors consider Hamilton Monte-Carlo as MCMC choice, resulting in Hamiltonian Variational Inference.
- It provides a flexible (albeit more computationally demanding) variational approximation that can be adjusted with the number T of MCMC steps.

Variational inference with Importance Sampling

"Importance Weighted Autoencoders"
Burda et al'16

- Another mechanism to improve the variational lower bound is to use importance sampling.
- For each k, we define

$$\mathcal{L}_k(x) = \mathbb{E}_{z_1,\ldots,z_k \sim q(z|x)} \left[\log \frac{1}{k} \sum_{i=1}^{k} \frac{p(x, z_i)}{q(z_i|x)} \right] .$$

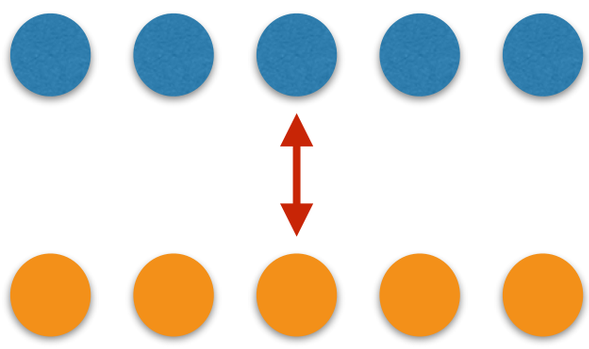
- It results that

$$\forall\ k\ ,\ \ \log p(x) \geq \mathcal{L}_{k+1}(x) \geq \mathcal{L}_k(x)\ ,\quad \text{and}$$

$$\lim_{k \to \infty} \mathcal{L}_k(x) = \log p(x) \text{ if } \frac{p(x,z)}{q(z|x)} \text{ is bounded } .$$

Other directed models

- Restricted Boltzmann Machines [Smolenski'86,Hinton,'02] are undirected graphical models with binary variables

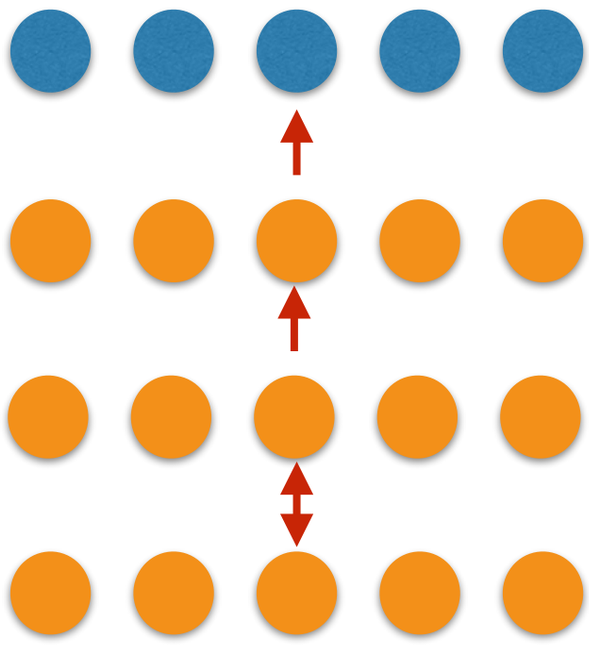


x (visible)

z (hidden)

$$p(x,z) = \exp\left(\langle \theta_1, xz^T \rangle + \langle \theta_2, x \rangle + \langle \theta_3, z \rangle - \log A(\theta)\right)$$

- Deep Belief Networks [Hinton et al'02]



x (visible)

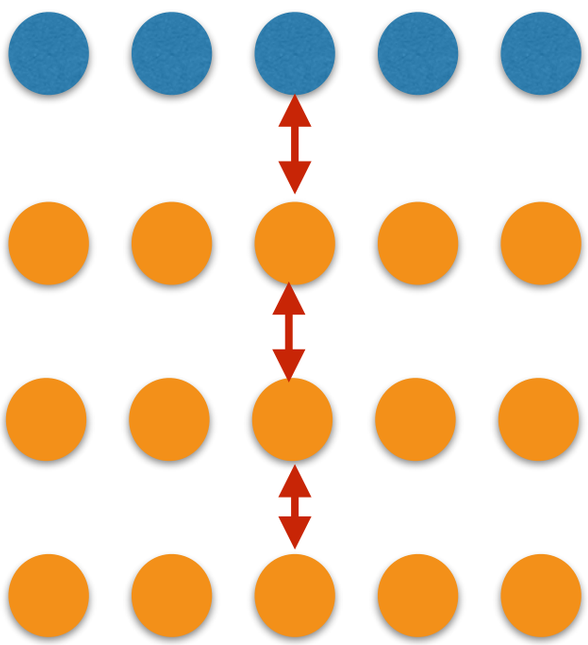
z_1 (hidden)

z_2 (hidden)

z_L (hidden)

Other directed models

- Deep Boltzmann Machines [Saladutnikov & Hinton,'09]



Can be trained greedily layer-wise

- See also:
  - Wake-Sleep [Hinton et al'95]
  - Generative Stochastic Networks [Bengio,'13].
  - ...

Limits of Mixture Models

- Inference can be computationally expensive for large models.
- The modeling $p(x)$ is reduced to the task of modeling $p(x|z)$

- Q: How to account for image variability?
  - $p(x|z) = \mathcal{N}(\Phi(z), \Sigma(z))$ corresponds to a model of *additive variability*:

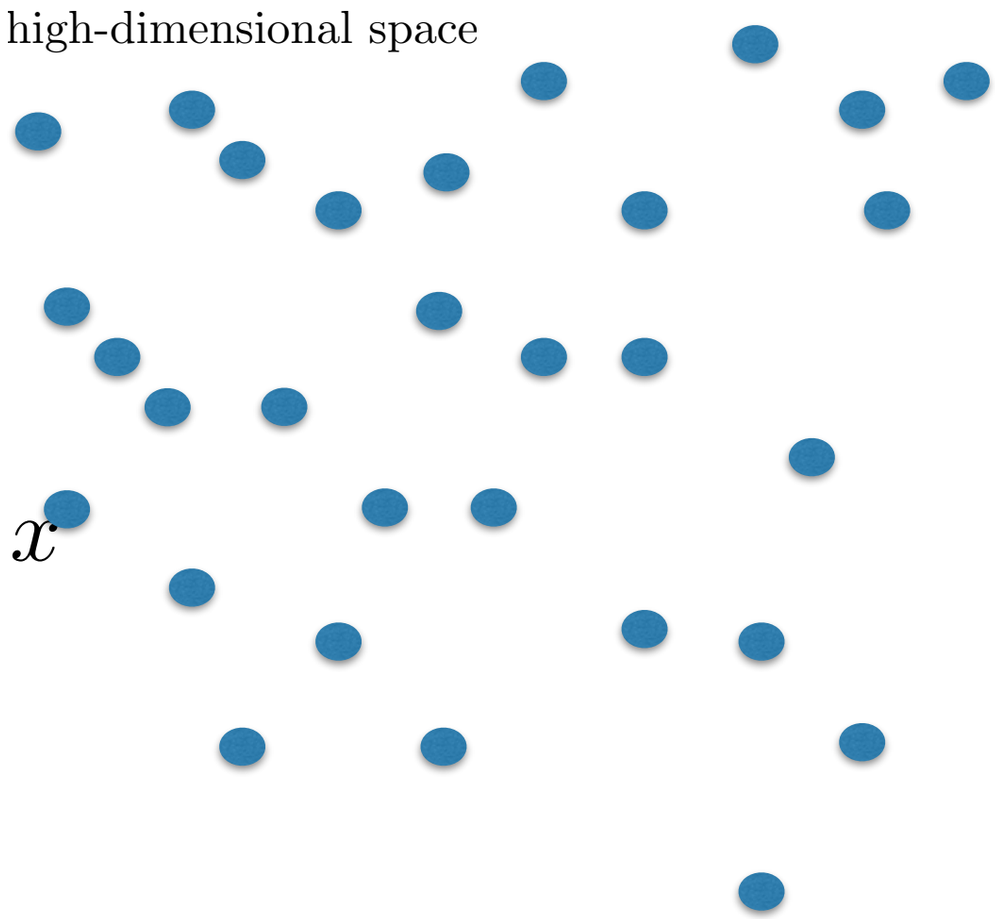
$$x = \Phi(z) + \epsilon \ , \ \epsilon \sim \mathcal{N}(0, \Sigma(z))$$

$$-\log p(x|z) \propto \|\Sigma(z)^{-1/2}(x - \Phi(z))\|^2$$

  - In particular, can we guarantee that $|p(x_\tau) - p(x)| \lesssim \|\tau\|$ with a mixture model?
  - Gaussian likelihoods tend to suffer from regression to the mean.

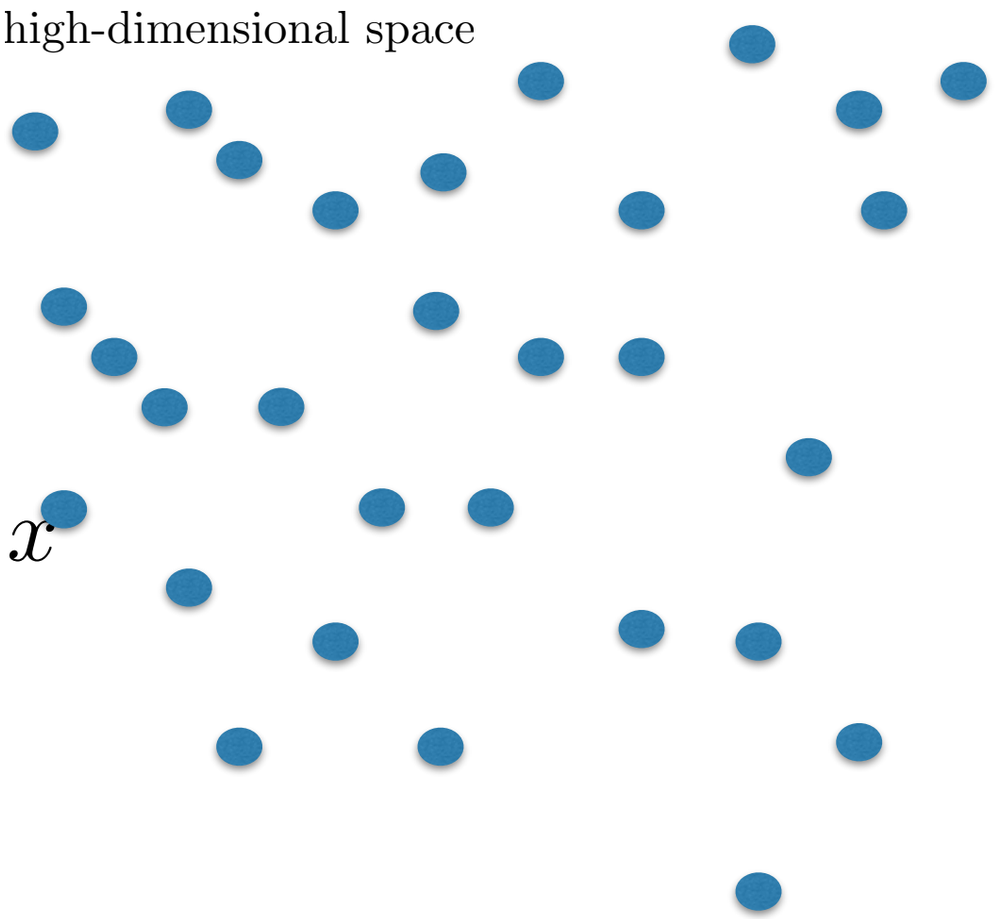
Generative Models of Complex data

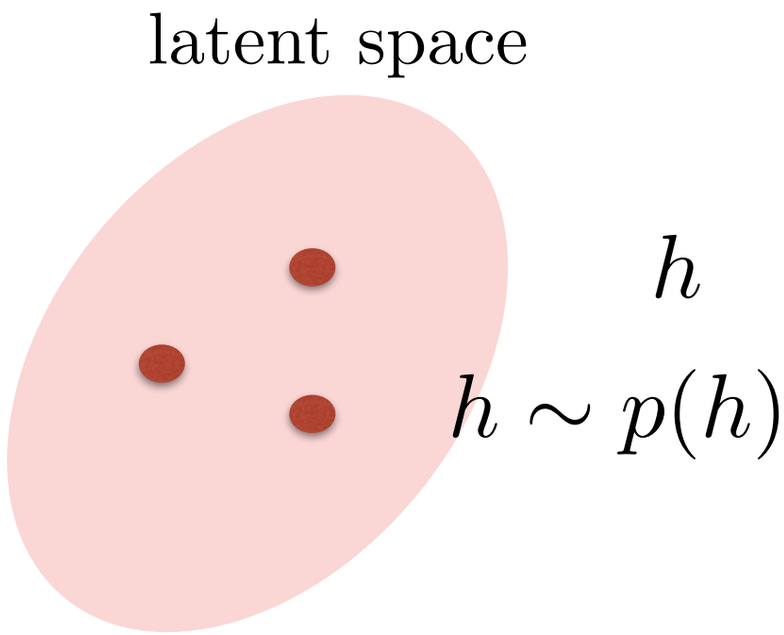
- Flows or Transports of Measure:



Generative Models of Complex data

- Flows or Transports of Measure:

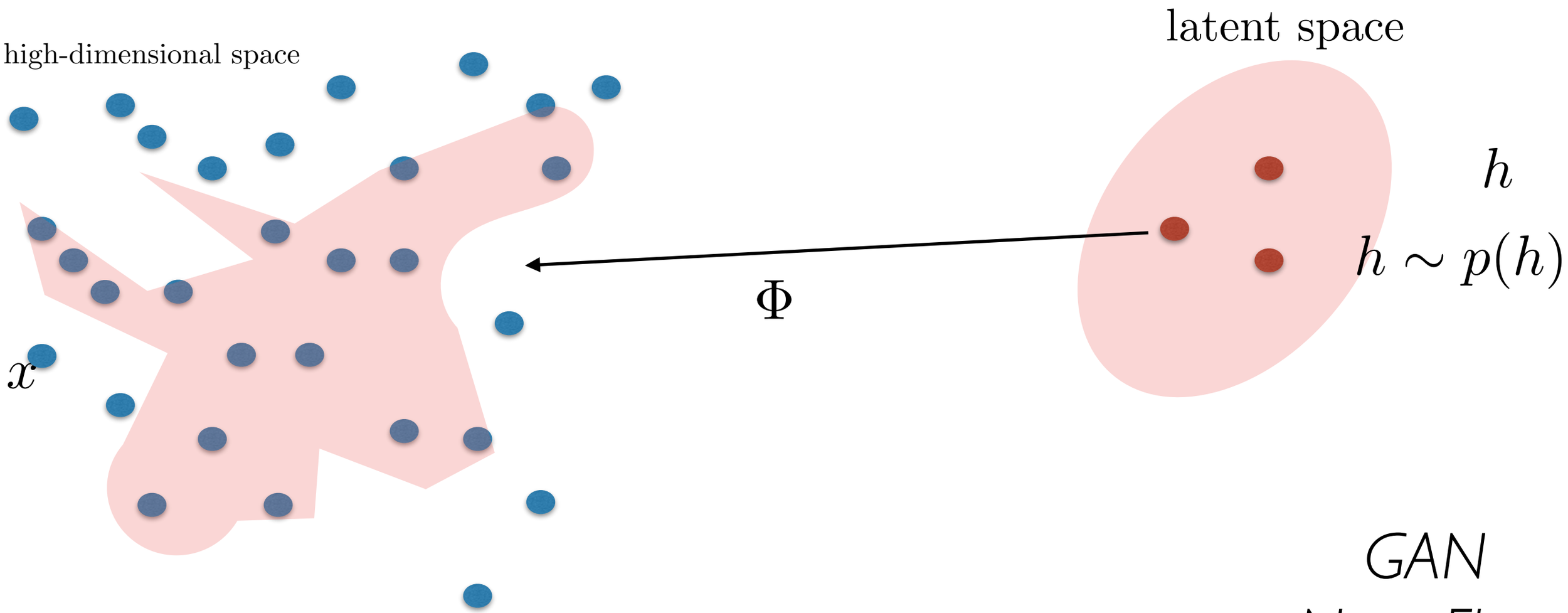




GAN
NormFlow
...

Generative Models of Complex data

- Flows or Transports of Measure



GAN
NormFlow
...

$p(x)$ defined implicitly with

$$\int f(x)p(x)dx = \int f(\Phi(h))p(h)dh \ , \ \forall \ f \text{ measurable}$$

Measure Transports

- How to train the transport Φ?

- We will see two methods:
  - Directly by optimizing data log-likelihood [Normalizing Flows]
  - Using a Discriminative Model [Generative Adversarial Networks]

Normalizing Flows

[Variational Inference with Normalizing Flows, Rezende & Mohamed'15]
[Tabak et al.'10]

- Consider a diffeomorphism $\Phi : \mathbb{R}^N \to \mathbb{R}^N$.
- If $z \in \mathbb{R}^N$ is a random variable with density $q(z)$, what is the density of $z' = \Phi(z)$?
- We have, for any measurable f,

$$\mathbb{E}_{z\sim q}(f(z')) = \int f(z')q(z)dz$$

$$= \int f(\Phi(z))q(z)dz = \int f(z)q(\Phi^{-1}(z))|\det(\nabla\Phi^{-1}(z))|dz$$

$$= \int f(z)\tilde{q}(z)dz = \mathbb{E}_{z'\sim\tilde{q}}(f(z'))\ \text{, with}$$

$$\tilde{q}(z') = q(z)\,|\det\nabla\Phi(z)|^{-1}\ \ ,\ \ z = \Phi^{-1}(z')\ .$$

Normalizing Flows

- The density $q_K(z)$ obtained by transporting a base measure q_0 through a cascade of K diffeomorphisms $\Phi_1, \ldots, \Phi_K$ is

$$z_K = \Phi_K \circ \ldots \Phi_1(z_0) \ , \ \text{with } z_0 \sim q_0(z)$$

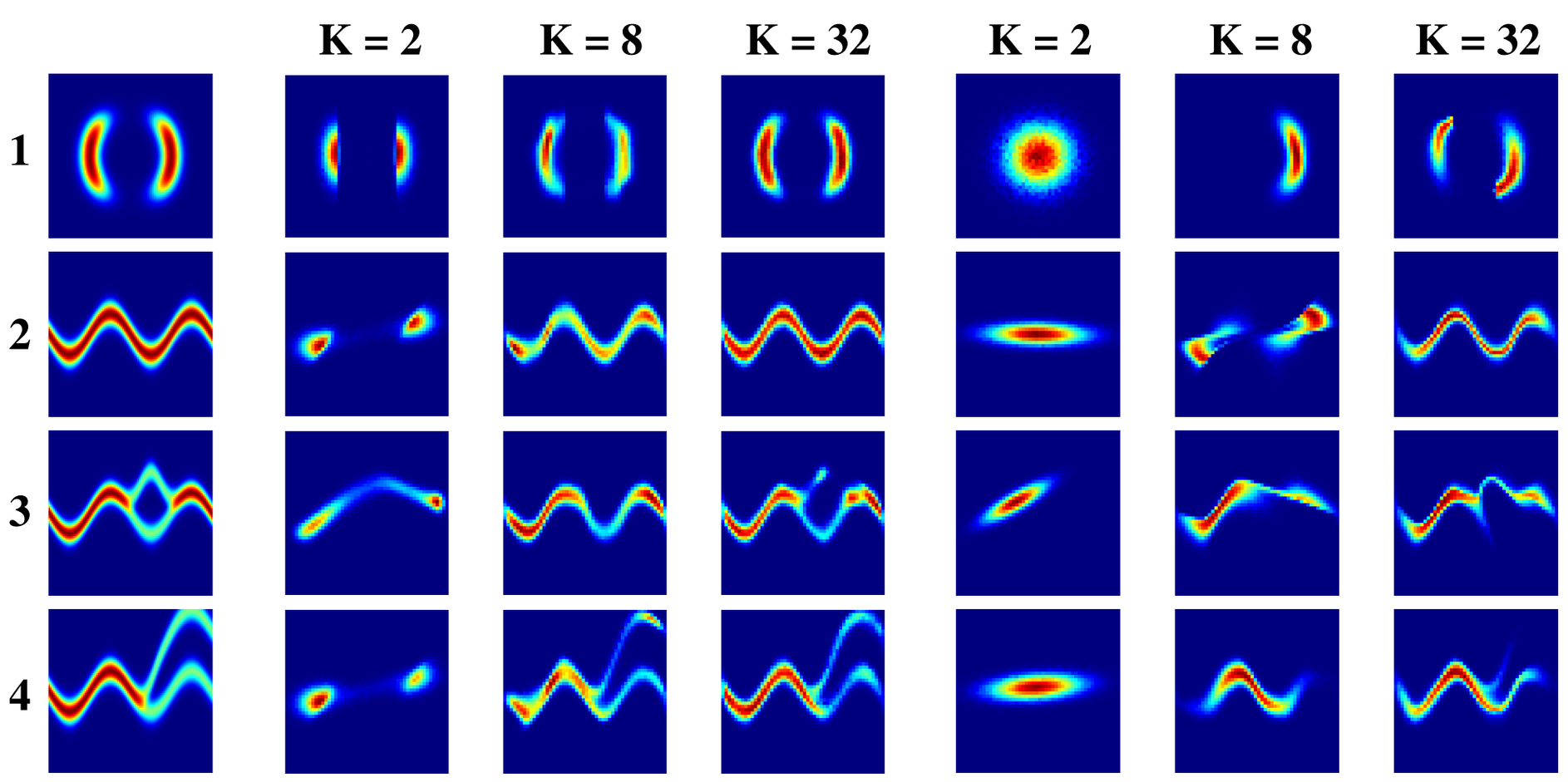
$$\log q_K(z) = \log q_0(z_0) - \sum_{k<K} \log \left| \det \nabla_{z_k} \Phi_k \right| \ .$$

- One can parametrize invertible flows and use them within the variational inference to improve the variational approximation. [Rezende et al.'15]
- Also considered in ["NICE", Dinh et al'15].
- Special case of *Autoregressive Flows* (i.e. Jacobian triangular) explored in "Variational Inference with Inverse Autoregressive Flows", by [Kingma, Salimans & Welling, NIPS'16].

Normalizing Flows

- Some low-dimensional transport results:

[Rezende et al.'15]



(a) (b) Norm. Flow (c) NICE

Diffusion and Non-equilibrium Thermodynamics

[Sohl-Dickstein et al.'15]

- We can also consider *infinitesimal* flows:

$$\frac{\partial q_t(z)}{\partial t} = \mathcal{F}(q_t(z)) \ , \ q_0(z) = p_0(z) \ .$$

$\mathcal{F}$ describes the dynamics.

Diffusion and Non-equilibrium Thermodynamics

[Sohl-Dickstein et al.'15]

- We can also consider *infinitesimal* flows:

$$\frac{\partial q_t(z)}{\partial t} = \mathcal{F}(q_t(z)) \ , \ q_0(z) = p_0(z) \ .$$

$\mathcal{F}$ describes the dynamics.

- For $\mathcal{F} = -\Delta$ we have Gaussian diffusion.

  It defines a Markov diffusion kernel that successively transforms data distribution $p_0(x)$ into a tractable distribution $\pi(x)$:

$$\pi(x) = \int T_\pi(x|x')\pi(x')dx'$$

$q(x^{(t+1)}|x^{(t)}) = T_\pi(x^{(t+1)}|x^{(t)}, \beta_t)$ $\qquad$ β_t: diffusion rate.

Diffusion and Non-equilibrium Thermodynamics

[Sohl-Dickstein et al.'15]

- The "forward" trajectory diffuses the data distribution into a tractable distribution, eg Gaussian.

Diffusion and Non-equilibrium Thermodynamics

[Sohl-Dickstein et al.'15]

- The "forward" trajectory diffuses the data distribution into a tractable distribution, eg Gaussian.
- The generative model learns how to reverse the diffusion:

$$p(x^{(0\ldots T)}) = p(x^{(T)}) \prod_{t \leq T} p(x^{(t-1)}|x^{(t)}) \ .$$

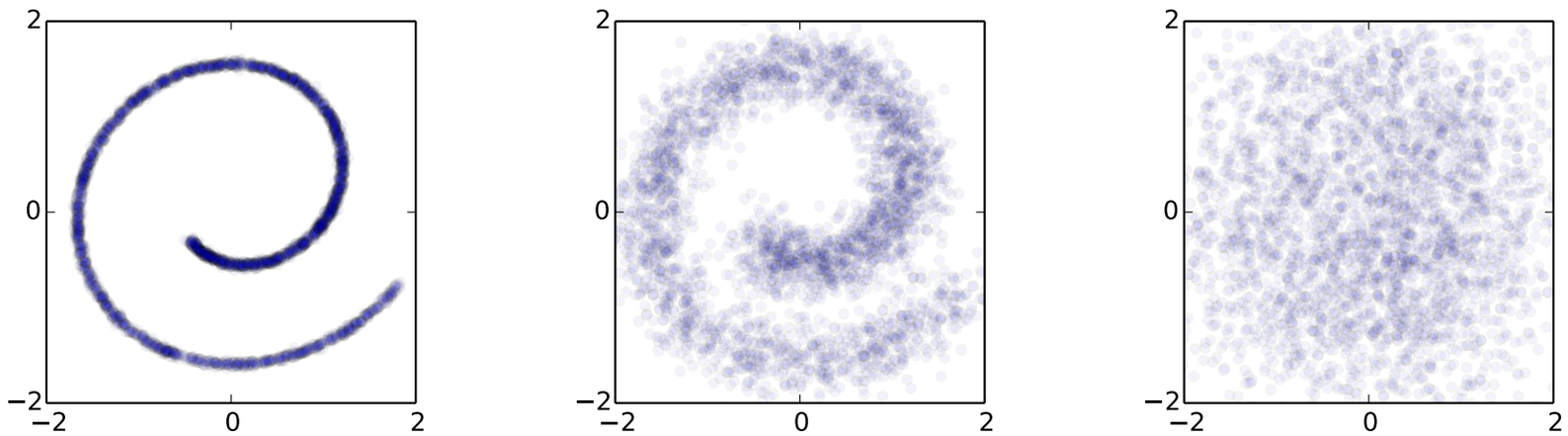
  - in the limit of infinitesimal difusion, the forward and backward kernel have the same functional form (Gaussian).
  - The parameters of the model are $\{\mu(x^{(t)}, t), \Sigma(x^{(t)}, t)\}_{t \leq T}$
- The data likelihood admits lower bound that can be evaluated efficiently using annealed importance sampling.

Diffusion and Non-equilibrium Thermodynamics

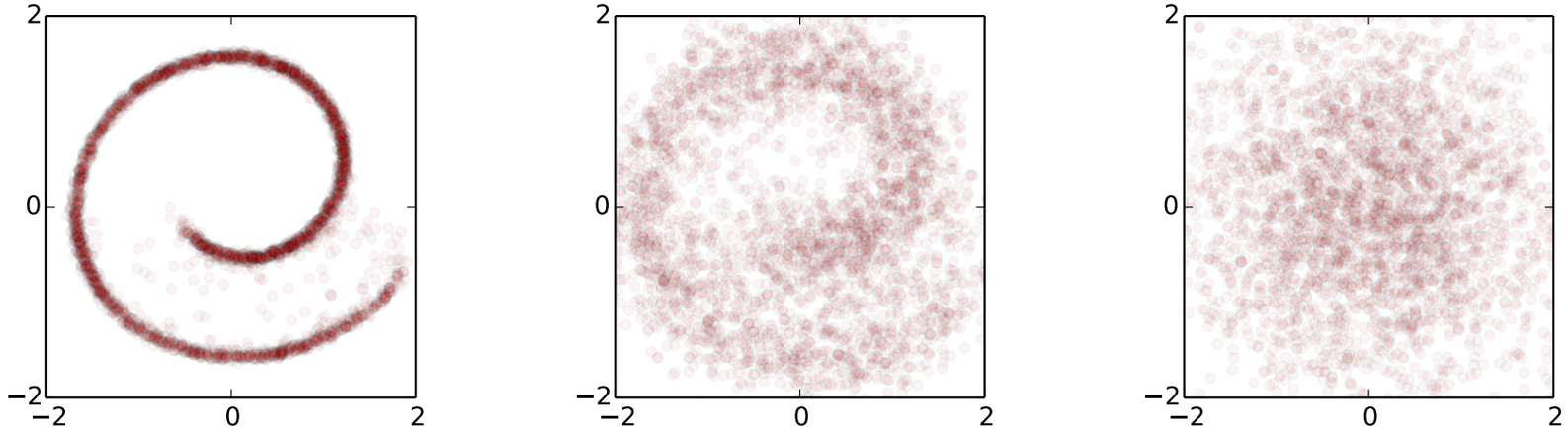
[Sohl-Dickstein et al.'15]

$t = 0$ $\quad t = \frac{T}{2}$ $\quad t = T$

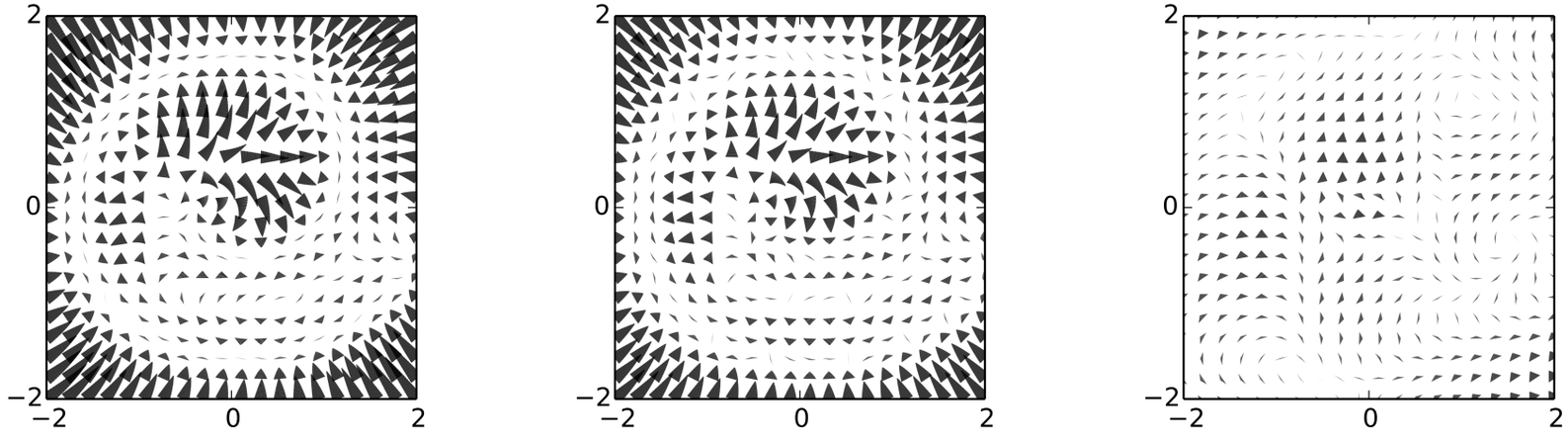
$q\left(\mathbf{x}^{(0 \cdots T)}\right)$



$p\left(\mathbf{x}^{(0 \cdots T)}\right)$

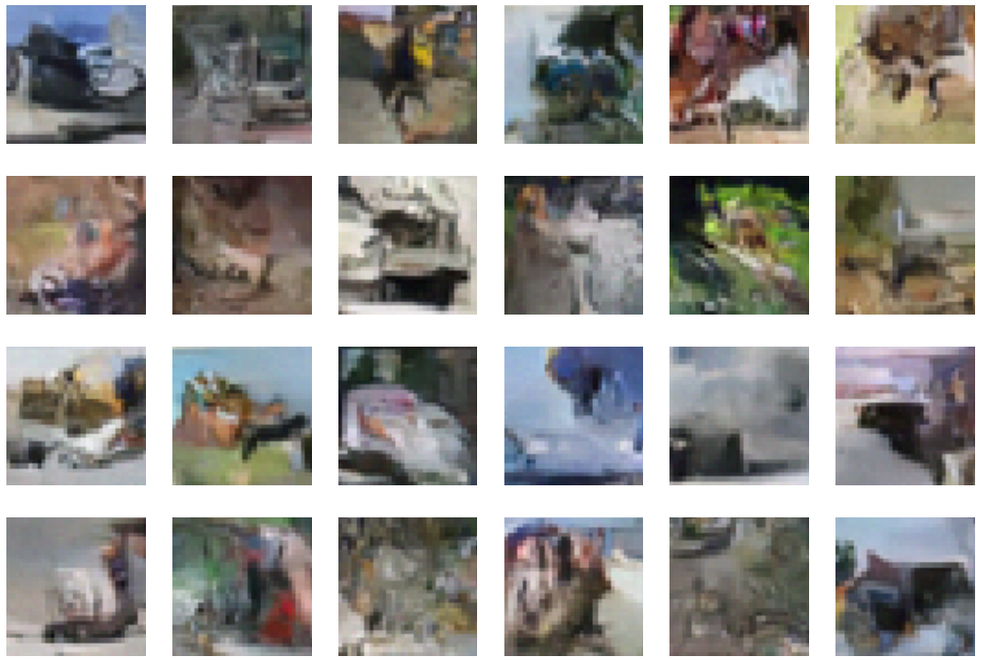


$\mathbf{f}_\mu\left(\mathbf{x}^{(t)}, t\right) - \mathbf{x}^{(t)}$

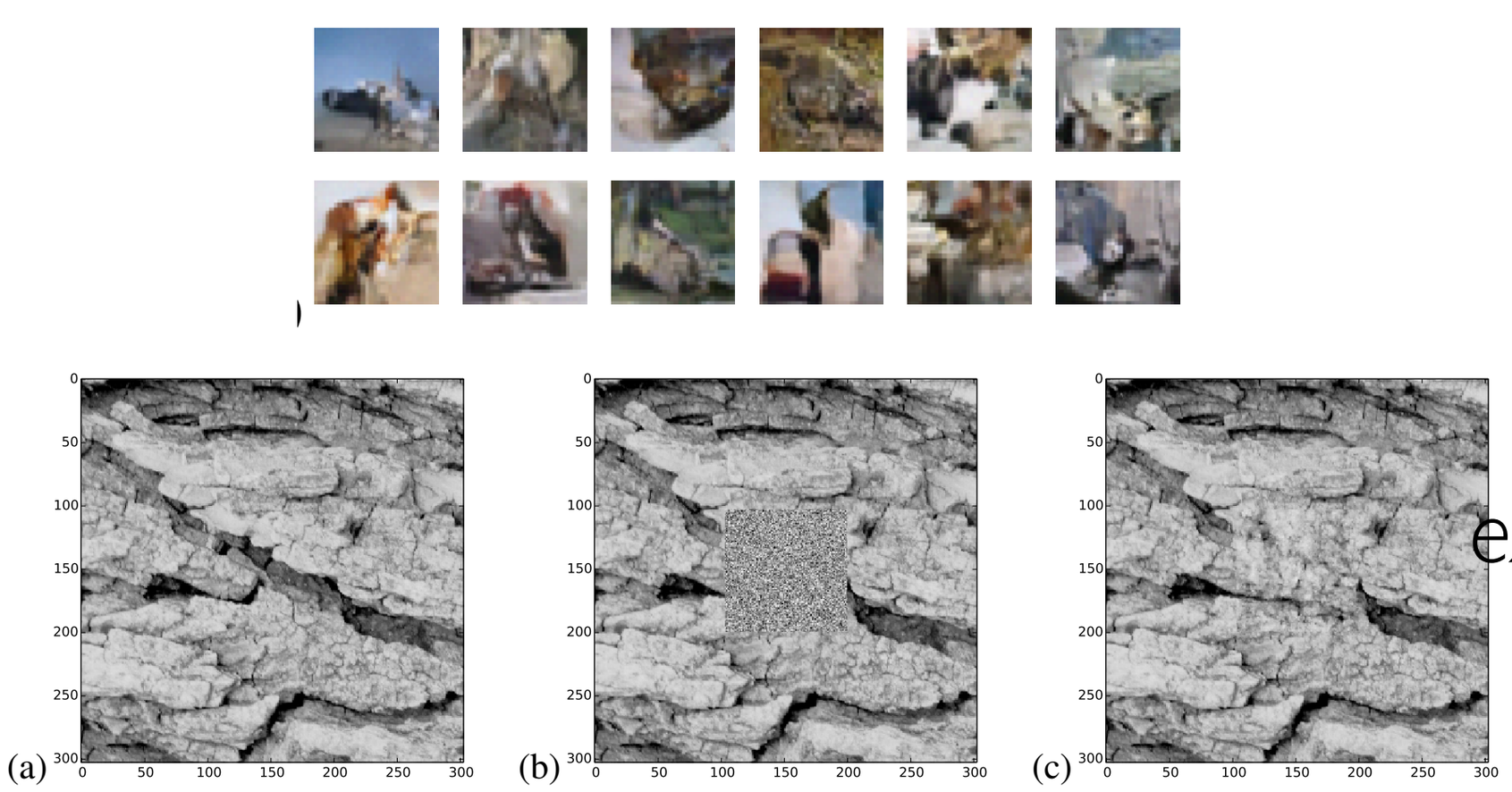


Diffusion and Non-equilibrium Thermodynamics

[Sohl-Dickstein et al.'15]



samples from the model trained on CIFAR-10



inpainting experiments