Inference and Representation
Lecture 8

Joan Bruna
Courant Institute, NYU

NYU | COURANT INSTITUTE OF MATHEMATICAL SCIENCES

Lecture 8 Objectives

- Exponential Families
- Variational Inference
- Variational Autoencoders

Variational Inference

- Q: What does *variational* mean?

Variational Inference

- Q: What does *variational* mean?
- In general, it refers to the idea of expressing a quantity of interest θ^* (e.g. a posterior probability) as the solution of an optimization problem:

$$\theta^* = \inf_{\theta \in \mathcal{M}} f(\theta) \ .$$

Variational Inference

- Q: What does *variational* mean?
- In general, it refers to the idea of expressing a quantity of interest θ^* (e.g. a posterior probability) as the solution of an optimization problem:

$$\theta^* = \inf_{\theta \in \mathcal{M}} f(\theta) \ .$$

- Approximating the solution can now be accomplished by
  – Simplifying the domain $\mathcal{M}$.
  – Simplifying the function f .

Variational Inference

- Q: What does *variational* mean?
- In general, it refers to the idea of expressing a quantity of interest θ^* (e.g. a posterior probability) as the solution of an optimization problem:

$$\theta^* = \inf_{\theta \in \mathcal{M}} f(\theta) \ .$$

- Approximating the solution can now be accomplished by
  - Simplifying the domain $\mathcal{M}$.
  - Simplifying the function f .
- Such approximations are particularly powerful in presence of convex structures.
- Let us start with variational inference in the exponential family.

Exponential Families

- Suppose we have iid data $x_1, \dots x_n$ and we consider a collection of *sufficient statistics* $\{\phi_k(X)\}_k$

- The empirical expectations of these statistics are

$$\hat{\mu}_k = \frac{1}{n} \sum_i \phi_k(x_i)$$

- Q: Can we build a distribution $p(x)$ consistent with these empirical moments? i.e.

$$\mathbb{E}_{X \sim p(x)}\{\phi_k(X)\} = \hat{\mu}_k \quad \text{for all } k.$$

Exponential Families

- Suppose we have iid data $x_1, \ldots x_n$ and we consider a collection of *sufficient statistics* $\{\phi_k(X)\}_k$

- The empirical expectations of these statistics are

$$\hat{\mu}_k = \frac{1}{n} \sum_i \phi_k(x_i)$$

- Q: Can we build a distribution $p(x)$ consistent with these empirical moments? i.e.

$$\mathbb{E}_{X \sim p(x)} \{\phi_k(X)\} = \hat{\mu}_k \quad \text{for all } k.$$

- In general, this is an underdetermined problem. How to choose wisely amongst all possible solutions?

Exponential Families and Maximum Entropy

- A reasonable choice is to consider the distribution with *maximum entropy* subject to the empirical moments:

$$p^* = \arg\max_p H(p) \ , \ s.t. \ \mathbb{E}_p\{\phi_k(X)\} = \hat{\mu}_k \ \text{for all} \ k.$$

$$\text{Shannon Entropy:} \ H(p) = -\mathbb{E}\{\log(p)\} \ .$$

Exponential Families and Maximum Entropy

- A reasonable choice is to consider the distribution with *maximum entropy* subject to the empirical moments:

$$p^* = \arg\max_p H(p) \ , \ \ s.t. \ \mathbb{E}_p\{\phi_k(X)\} = \hat{\mu}_k \ \text{for all} \ k.$$

$$\text{Shannon Entropy: } H(p) = -\mathbb{E}\{\log(p)\} \ .$$

- The general form of maximum entropy is

$$p(x) \propto \exp\left\{ \sum_k \lambda_k \phi_k(x) \right\}$$

λ_k: Lagrange multipliers adjusted such that $\mathbb{E}_p \phi_k(X) = \hat{\mu}_k$ for all k.

Exponential Families

- The exponential family associated with ϕ is defined as the parametric family:

$$p_\theta(x) = \exp\{\langle \theta, \phi(x) \rangle - A(\theta)\} \ , \ \text{with}$$

$$A(\theta) = \log \int \exp\{\langle \theta, \phi(x) \rangle\} dx \qquad \text{log-partition function}$$

Exponential Families

- The exponential family associated with ϕ is defined as the parametric family

$$p_\theta(x) = \exp\{\langle \theta, \phi(x) \rangle - A(\theta)\} \ , \ \ \textbf{with}$$

$$A(\theta) = \log \int \exp\{\langle \theta, \phi(x) \}dx \qquad \textbf{log-partition function}$$

- It is well defined for the family of parameters

$$\Omega = \{\theta \ ; \ A(\theta) < \infty\}$$

- Several well-known models belong to the exponential family:
  - Energy based models
  - Gaussian Mixtures
  - Latent Dirichlet Allocation

Exponential Families

- **Proposition:** The log-partition function $A(\theta)$ satisfies

  - $$\frac{\partial A}{\partial \theta_k}(\theta) = \mathbb{E}_\theta\{\phi_k(X)\} = \int \phi_k(x) p_\theta(x) dx \ .$$

  - $A(\theta)$ is convex in its domain Ω.

- Higher order derivatives always exist.

Legendre Transform

Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function.

The Legendre Transform f^* of f is defined as

$$f^*(u) = \sup_x (xu - f(x))$$

- f^* is the convex conjugate of f.

Legendre Transform

Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function.

The Legendre Transform f^* of f is defined as

$$f^*(u) = \sup_x (xu - f(x))$$

- f^* is the convex conjugate of f.
- Equivalent definition in the differentiable case:

f and g are Legendre transforms of each other if their first derivatives are inverses of each other:

$$\forall\ x,\ \ g'(f'(x)) = x\ ,\ \forall\ u,\ \ f'(g'(u)) = u\ .$$

- It follows that $f^{**} = f$

Conjugate Duality

- Conjugate duality representation of convex functions:

$$A^*(\mu) = \sup_{\theta \in \Omega}\{\langle \mu, \theta \rangle - A(\theta)\}$$

$$\text{canonical parameters} \longleftrightarrow \text{moment parameters}$$

$$\theta_k \qquad\qquad \mu_k$$

- Q: How to interpret the dual conjugate?

Conjugate Duality

- Conjugate duality representation of convex functions:

$$A^*(\mu) = \sup_{\theta \in \Omega} \{\langle \mu, \theta \rangle - A(\theta)\}$$

$$\text{canonical parameters} \longleftrightarrow \text{moment parameters}$$

$$\theta_k \qquad\qquad \mu_k$$

- Q: How to interpret the dual conjugate?

$A^*(\mu)$: Negative entropy of $p_{\theta(\mu)}$, where
$p_{\theta(\mu)}$ is the exponential family distribution such that
$\mathbb{E}_{\theta(\mu)}\phi(X) = \mu$.

Conjugate Duality

- Conjugate duality representation of convex functions:

$$A^*(\mu) = \sup_{\theta \in \Omega} \{\langle \mu, \theta \rangle - A(\theta)\}$$

$$\text{canonical parameters} \longleftrightarrow \text{moment parameters}$$

$$\theta_k \qquad\qquad \mu_k$$

- Q: How to interpret the dual conjugate?

$A^*(\mu)$: Negative entropy of $p_{\theta(\mu)}$, where $p_{\theta(\mu)}$ is the exponential family distribution such that $\mathbb{E}_{\theta(\mu)}\phi(X) = \mu$.

- Variational representation:

$$A(\theta) = \sup_{\mu} \{\langle \theta, \mu \rangle - A^*(\mu)\}$$

Variational Inference and Duality

- We derive the exact EM algorithm for exponential families with latent variables. Given observed variables Z and latent variables X, we consider

$$p_\theta(x,z) = \exp\{\langle \theta, \phi(x,z)\rangle - A(\theta)\} \ \ , \ \text{with}$$

$$A(\theta) = \log \int_{x,z} \exp\{\langle \theta, \phi(x,z)\rangle\} dx dz$$

Variational Inference and Duality

- We derive the exact EM algorithm for exponential families with latent variables. Given observed variables Z and latent variables X, we consider

$$p_\theta(x,z) = \exp\left\{\langle \theta, \phi(x,z)\rangle - A(\theta)\right\} \ \ , \ \ \text{with}$$

$$A(\theta) = \log \int_{x,z} \exp\{\langle \theta, \phi(x,z)\rangle\} dx dz$$

- Given observation $X = x$, the posterior distribution is

$$p(z \mid x) = \frac{\exp\{\langle \theta, \phi(x,z)\rangle\}}{\int \exp\{\langle \theta, \phi(x,z')\rangle\} dz'} = \exp\{\langle \theta \phi(x,z)\rangle - A_x(\theta)\}$$

$$A_x(\theta) = \log \int_z \exp\{\langle \theta, \phi(x,z)\rangle\} dz$$

Variational Inference and Conjugate Duality

- The MLE for our parameters $\boldsymbol{\theta}$ is obtained by maximizing the incomplete log-likelihood of the data:

$$\mathcal{L}(\theta, x) = \log \int_z \exp\{\langle \theta, \phi(x, z) \rangle - A(\theta)\} dz = A_x(\theta) - A(\theta) \ .$$

Variational Inference and Conjugate Duality

- The MLE for our parameters $\boldsymbol{\theta}$ is obtained by maximizing the incomplete log-likelihood of the data:

$$\mathcal{L}(\theta, x) = \log \int_z \exp\{\langle \theta, \phi(x,z) \rangle - A(\theta)\} dz = A_x(\theta) - A(\theta) \ .$$

- The variational representation gives

$$A_x(\theta) = \sup_{\mu_x} \{\langle \theta, \mu_x \rangle - A_x^*(\mu_x)\}$$

$$A_x^*(\mu_x) = \sup_{\theta} \{\langle \theta, \mu_x \rangle - A_x(\theta)\}$$

Variational Inference and Conjugate Duality

- The MLE for our parameters $\boldsymbol{\theta}$ is obtained by maximizing the incomplete log-likelihood of the data:

$$\mathcal{L}(\theta, x) = \log \int_z \exp\{\langle \theta, \phi(x,z) \rangle - A(\theta)\} dz = A_x(\theta) - A(\theta) \ .$$

- The variational representation gives

$$A_x(\theta) = \sup_{\mu_x} \{\langle \theta, \mu_x \rangle - A_x^*(\mu_x)\}$$

$$A_x^*(\mu_x) = \sup_{\theta} \{\langle \theta, \mu_x \rangle - A_x(\theta)\}$$

- It results in the lower-bound for the incomplete log-likelihood:

$$\mathcal{L}(\theta, x) \geq \langle \mu_x, \theta \rangle - A_x^*(\mu_x) - A(\theta) = \widetilde{\mathcal{L}}(\mu_x, \theta)$$

- EM is thus a coordinate ascent on the lower bound:

$$\mu_x^{(t+1)} = \arg\max_{\mu_x} \widetilde{\mathcal{L}}(\mu_x, \theta^{(t)}) \qquad \text{(E step)}$$

$$\theta^{(t+1)} = \arg\max_{\theta} \widetilde{\mathcal{L}}(\mu_x^{(t+1)}, \theta) \qquad \text{(M step)}$$

Variational Inference and Conjugate Duality

- EM is thus a coordinate ascent on the lower bound:

$$\mu_x^{(t+1)} = \arg\max_{\mu_x} \widetilde{\mathcal{L}}(\mu_x, \theta^{(t)}) \qquad \text{(E step)}$$

$$\theta^{(t+1)} = \arg\max_{\theta} \widetilde{\mathcal{L}}(\mu_x^{(t+1)}, \theta) \qquad \text{(M step)}$$

- E step is called expectation because the maximizer of $\widetilde{\mathcal{L}}(\mu_x, \theta)$ is, by duality, the expectation $\mu_x^{(t+1)} = \mathbb{E}_{\theta^{(t)}} \phi(x, Z)$

Variational Inference and Conjugate Duality

- EM is thus a coordinate ascent on the lower bound:

$$\mu_x^{(t+1)} = \arg\max_{\mu_x} \widetilde{\mathcal{L}}(\mu_x, \theta^{(t)}) \qquad \text{(E step)}$$

$$\theta^{(t+1)} = \arg\max_{\theta} \widetilde{\mathcal{L}}(\mu_x^{(t+1)}, \theta) \qquad \text{(M step)}$$

- E step is called expectation because the maximizer of $\widetilde{\mathcal{L}}(\mu_x, \theta)$ is, by duality, the expectation $\mu_x^{(t+1)} = \mathbb{E}_{\theta^{(t)}} \phi(x, Z)$

- Also, because $\max_{\mu} \{\langle \mu_x, \theta^{(t)} \rangle - A_x^*(\mu_x)\} = A_x(\theta^{(t)})$, after each E step the inequality becomes an equality, thus M step increases log-likelihood.

Approximate Posterior Inference

- For most models, the posterior is analytically intractable:

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

- MCMC Route:
  - Express $p(z|x)$ as the stationary distribution of an appropriate Markov Chain.
  - Then query properties of $p(z|x)$ by Monte-Carlo Sampling.

Approximate Posterior Inference

- For most models, the posterior is analytically intractable:

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

- <u>MCMC Route:</u>
  - Express $p(z|x)$ as the stationary distribution of an appropriate Markov Chain.
  - Then query properties of $p(z|x)$ by Monte-Carlo Sampling.

- ***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:

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

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

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

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

Mean Field Variational Bayes

- Joint likelihood of observed and latent variables:

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

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

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

  - 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.

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:

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

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

Mean Field Variational Bayes

[Paiskey, Blei, Jordan,'12]

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

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

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

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

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

Mean Field Variational Bayes

[Paiskey, Blei, Jordan,'12]

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

$$
\begin{aligned}
\nabla_\beta \mathbb{E}_{q(z|\beta)} f(Z) &= \nabla_\beta \int f(z) q(z|\beta) dz \\
&= \int f(z) \nabla_\beta q(z|\beta) dz \\
&= \int f(z) q(z|\beta) \nabla_\beta \log q(z|\beta) dz \\
&= \mathbb{E}_q \{ f(Z) \nabla_\beta \log q(z|\beta) \}
\end{aligned}
$$

- 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 (Williams'95).

Rao Blackwellization

- Variance reduction technique introduced in [Casella Robert '96]
- Applied to Variational Inference in [Ranganath et al'14].
- Consider two random variables X, Y and a function $J(X, Y)$.
- We are interested in computing and estimator for $\mathbb{E}_{X,Y} J(X, Y)$
- Define $\bar{J}(X) = \mathbb{E}_Y[J(X, Y)|X]$. What is $\mathbb{E}_X \bar{J}(X)$?
- It is $\mathbb{E}_X \bar{J}(X) = \mathbb{E}_{X,Y} J(X, Y)$, so we can use it instead of $J(X, Y)$ in a Monte-Carlo setup.
- Observe that

$$\mathbb{V}[\bar{J}(X)] = \mathbb{V}[J(X, Y)] - \mathbb{E}[(J(X, Y) - \bar{J}(X))^2] \leq \mathbb{V}[J(X, Y)] \,.$$

- We can rewrite the gradient of $\mathcal{L}$ as a conditional expectation wrt the Markov Blanket variables:

$$\nabla_{\beta_i} \mathbb{E}_{q(z|\beta)} \log p(X, Z|\theta) = \mathbb{E}_{z \sim q_i(z|\beta_i)} \left[\nabla_{\beta_i} \log q(z_i|\beta_i) \log p_i(x, z_{(i)}) \right]$$

MCMC vs Variational Inference

- Q: What are the pros/cons of MCMC versus VI?

MCMC vs Variational Inference

- Q: What are the pros/cons of MCMC versus VI?

MCMC	VI
Asymptotically Exact (why?)	Not exact (why?)
Computationally Expensive	Computationally Tractable
Robust Assumptions	Domain Knowledge

MCMC vs Variational Inference

- In particular, Variational Inference posterior approximations

$$q^*(z) = \arg\min_{q(z)\in\mathcal{F}} KL[q(z) \mid\mid p(z \mid x)]$$

tend to underestimate the variance of the posterior distribution:

$$\sigma_x^2 = \mathrm{Var}(Z)\ ,\ \ Z \sim p(z|x)\ .$$

MCMC vs Variational Inference

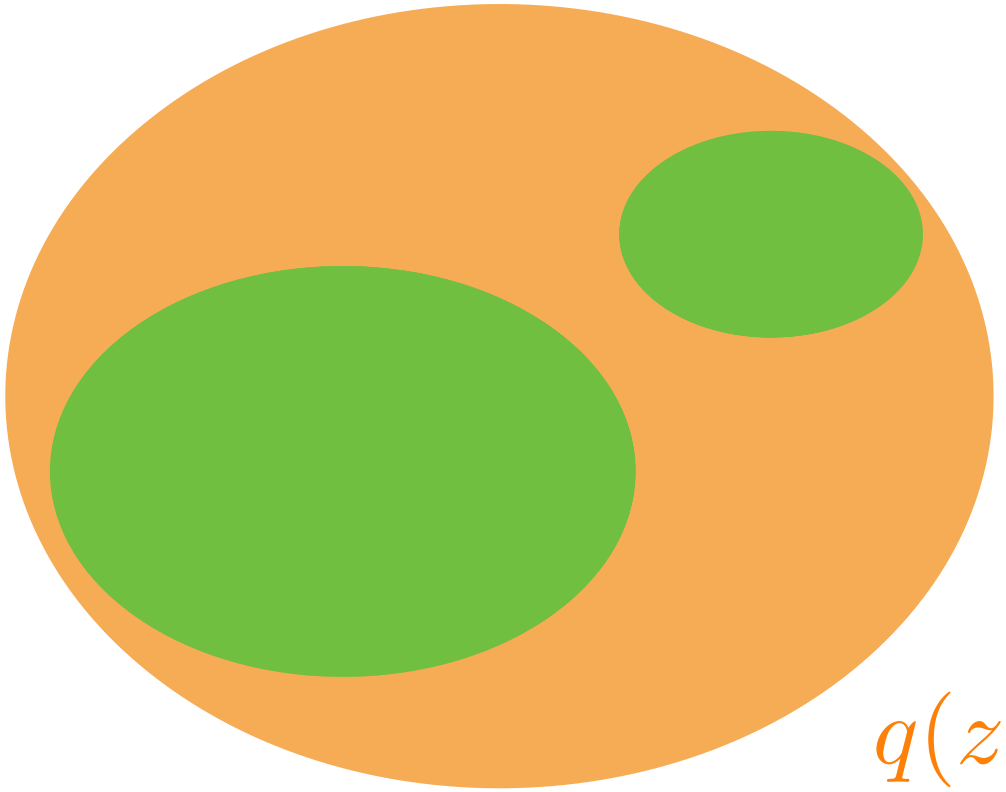
- In particular, Variational Inference posterior approximations

$$q^*(z) = \arg\min_{q(z)\in\mathcal{F}} KL[q(z) \mid\mid p(z \mid x)]$$

tend to underestimate the variance of the posterior distribution:

$$\sigma_x^2 = \mathrm{Var}(Z)\ ,\ \ Z \sim p(z|x)\ .$$

true posterior $p(z|x)$
variational approximation $q(z)$



$$KL[q(z) \mid\mid p(z)] = \sum_z q(z) \log\left(\frac{q(z)}{p(z)}\right)$$

$$q(z) \geq \delta > 0, p(z|x) \approx 0 \Rightarrow\ KL(q||p) \text{ large !}$$

MCMC vs Variational Inference

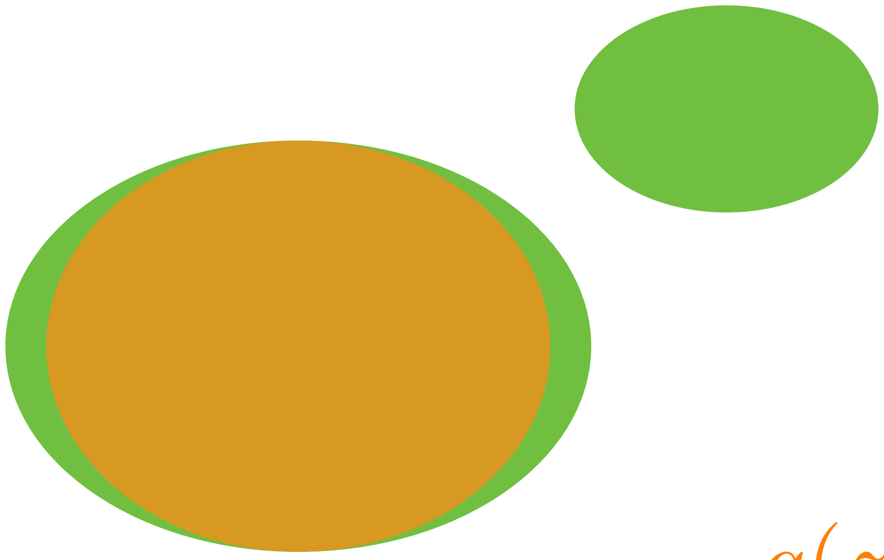
- In particular, Variational Inference posterior approximations

$$q^*(z) = \arg \min_{q(z) \in \mathcal{F}} KL[q(z) \mid\mid p(z \mid x)]$$

tend to underestimate the variance of the posterior distribution:

$$\sigma_x^2 = \mathrm{Var}(Z) \ , \ Z \sim p(z|x) \ .$$

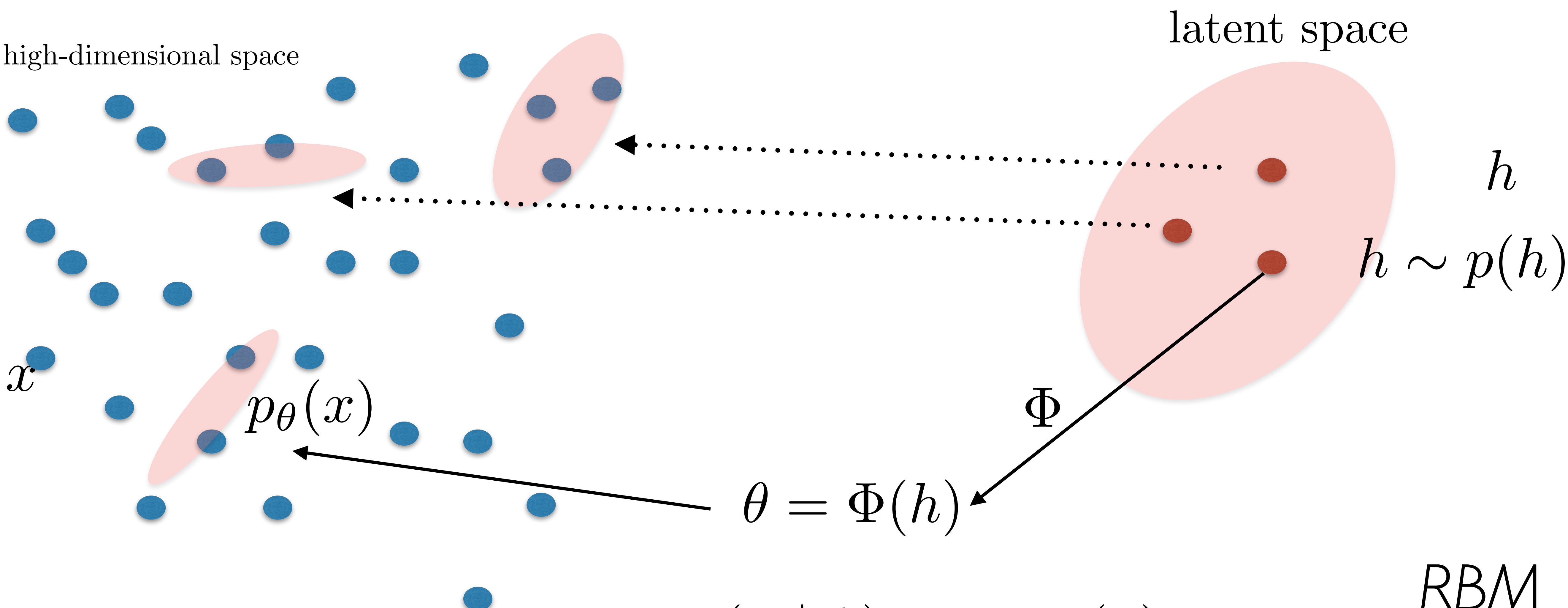
true posterior $p(z|x)$
variational approximation $q(z)$



$$KL[q(z) \mid\mid p(z)] = \sum_z q(z) \log \left(\frac{q(z)}{p(z)} \right)$$

$$q(z) \approx 0 \ , \ p(z|x) \geq \delta > 0 \Rightarrow \ KL(q||p) \text{ small!}$$

- Latent Graphical Models or *Mixtures*.



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

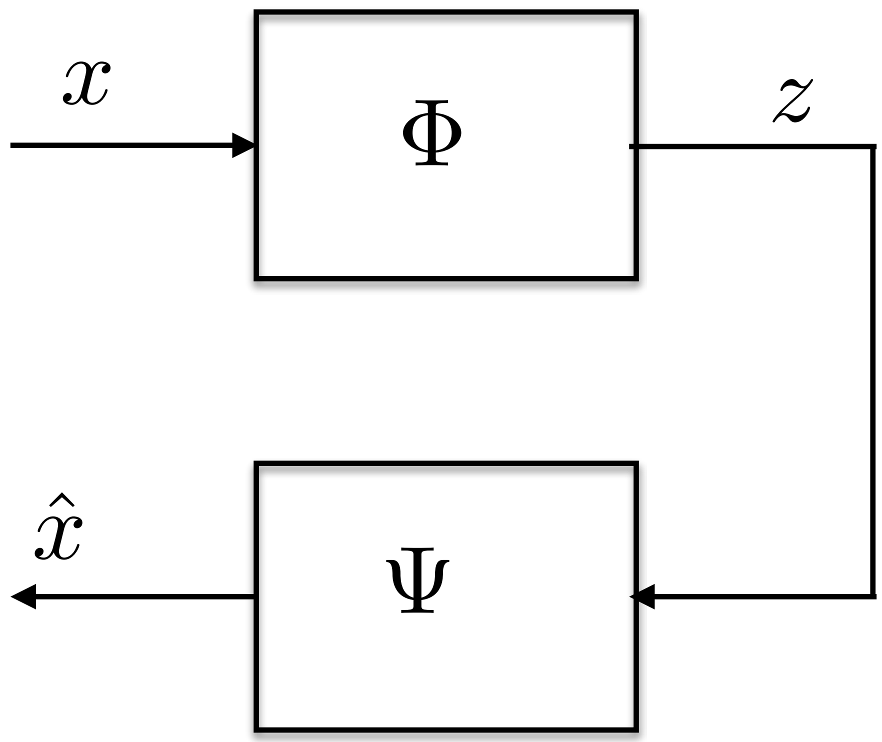
$$p(x) = \int p(x, h) dh = \int p(x \mid h) p(h) dh$$

RBM
DBN
DBM
VAE
…

Model: additive combination of simple parametric models

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:

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

  - 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

$\mathcal{R}$: Regularization term

Auto encoders

- Optimization set-up:

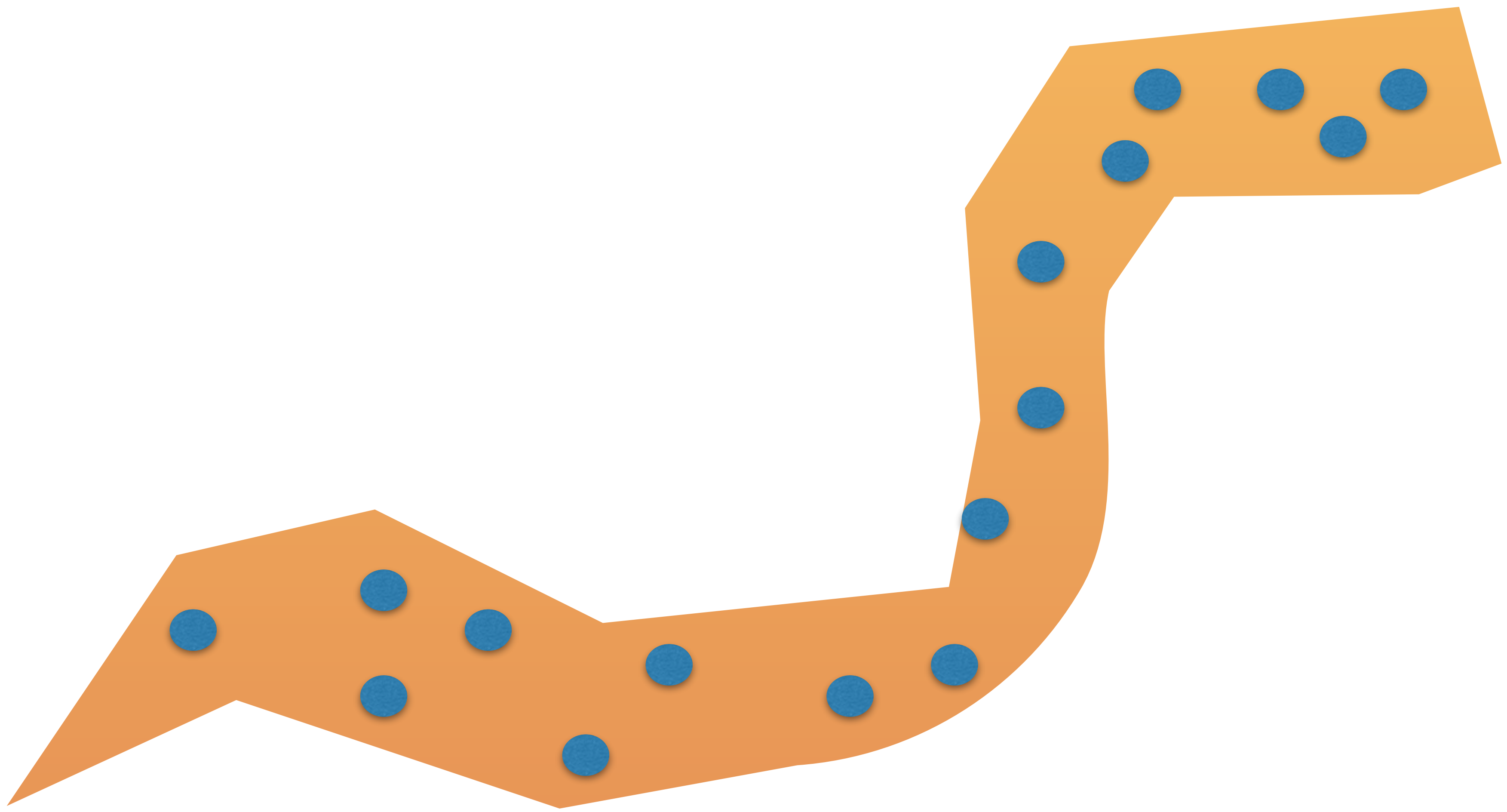
$$\min_{\Phi,\Psi} \frac{1}{n} \sum_{i \leq n} \ell\left(x_i, \Psi(\Phi(x_i))\right) + \mathcal{R}(\Phi(X))$$

$\ell(x, x')$: 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.
  - Denoising autoencoders: limit the capacity of the channel by making it noisy.

Auto encoders: Geometric Interpretation

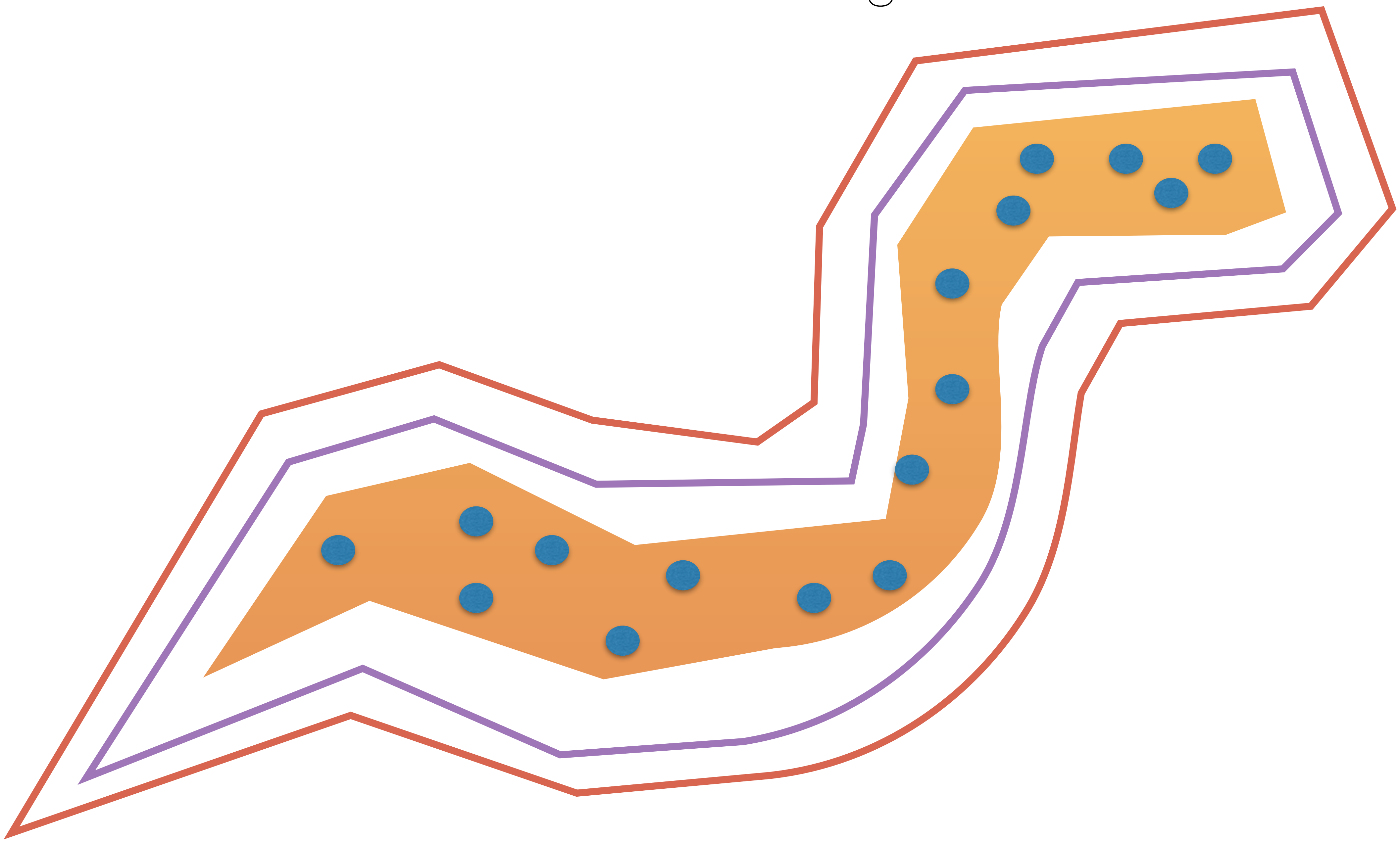
- The reconstruction error approximates a distance to a covering manifold of X



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

Auto encoders: Geometric Interpretation

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



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

More Examples

- Sparse Coding approximations

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

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

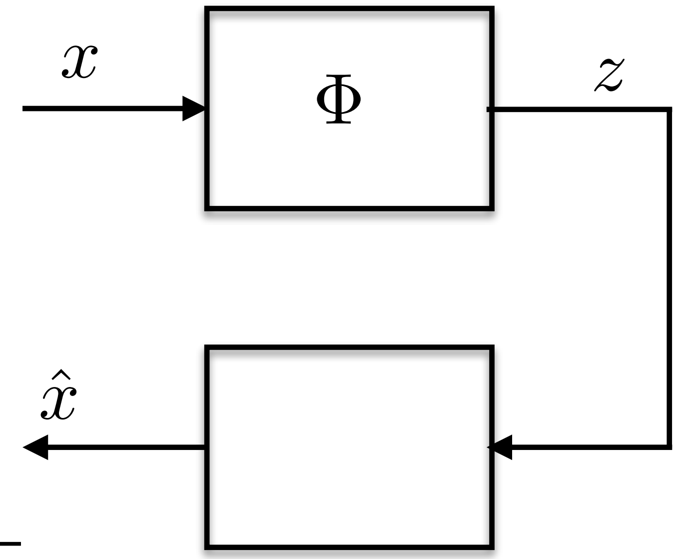
$$\Phi(X) = \mathrm{diag}(\beta) \tanh(WX + b)$$

  - 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 $p(z \mid x) = \delta(z - \Phi(x))$

  - It can model the mode (MAP) or the mean of the posterior.

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

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

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

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

Variational Autoencoders

- It follows 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

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

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

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

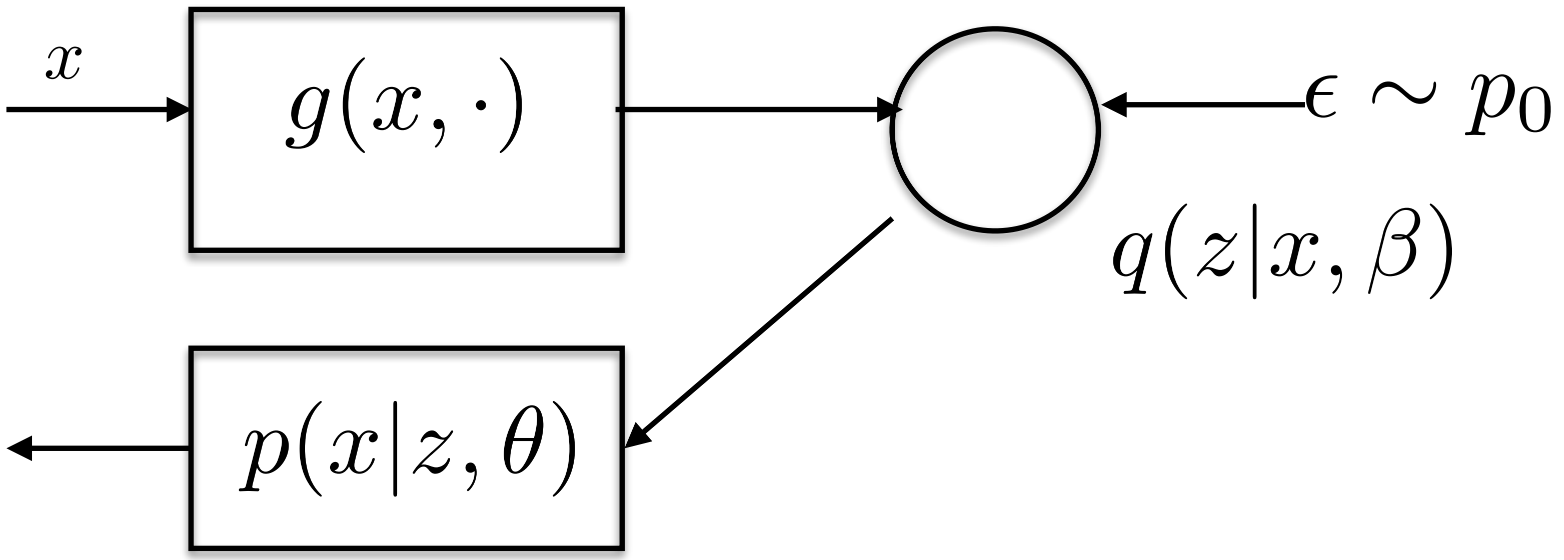
Variational Autoencoders

- 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:

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

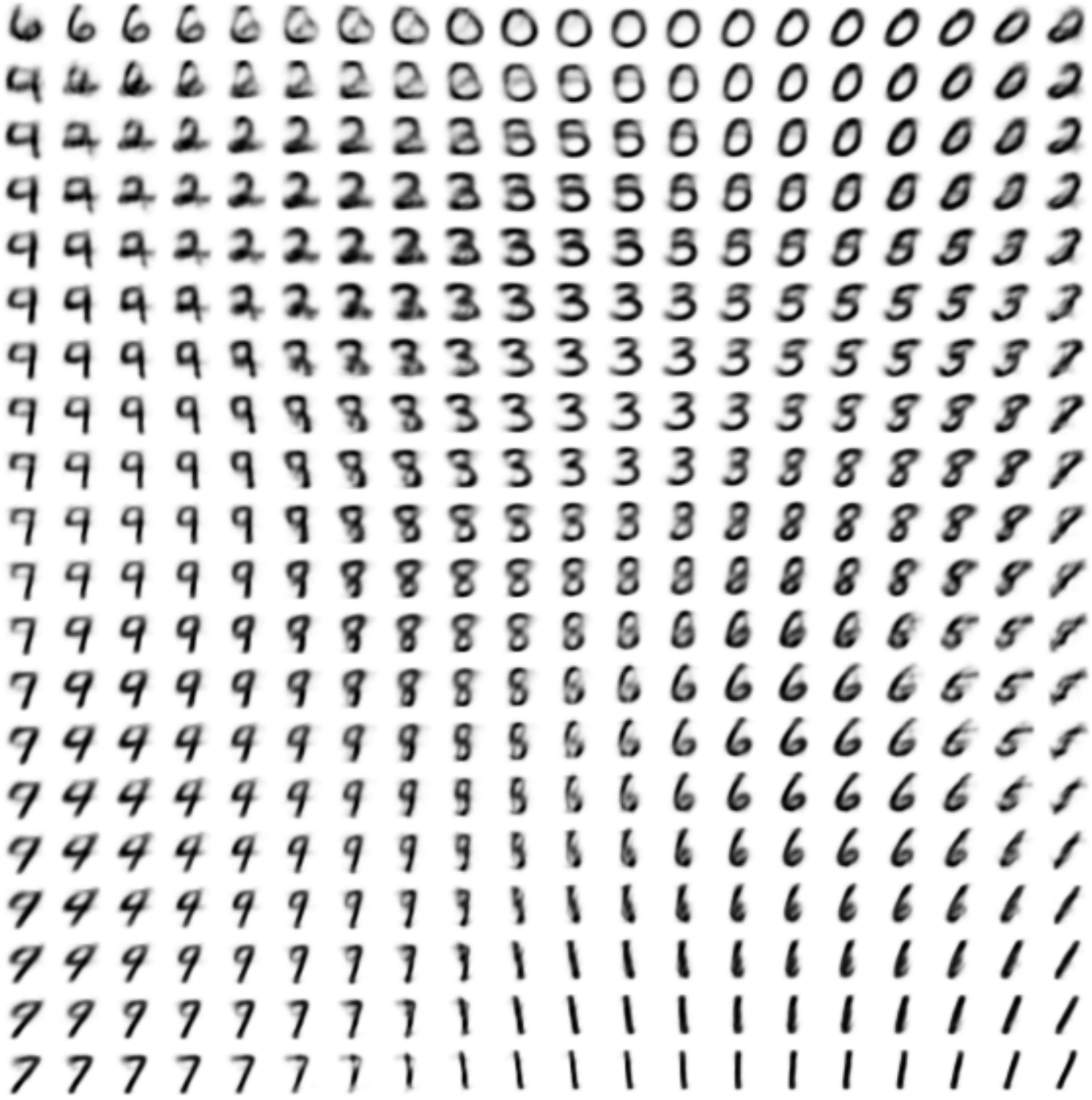
$$\overline{\mu}(z), \overline{\Sigma}(z) \ : \ \text{Neural networks, } (\overline{\Sigma} \text{ diagonal})$$

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

Variational Autoencoder

- Examples using a two-dimensional latent space:

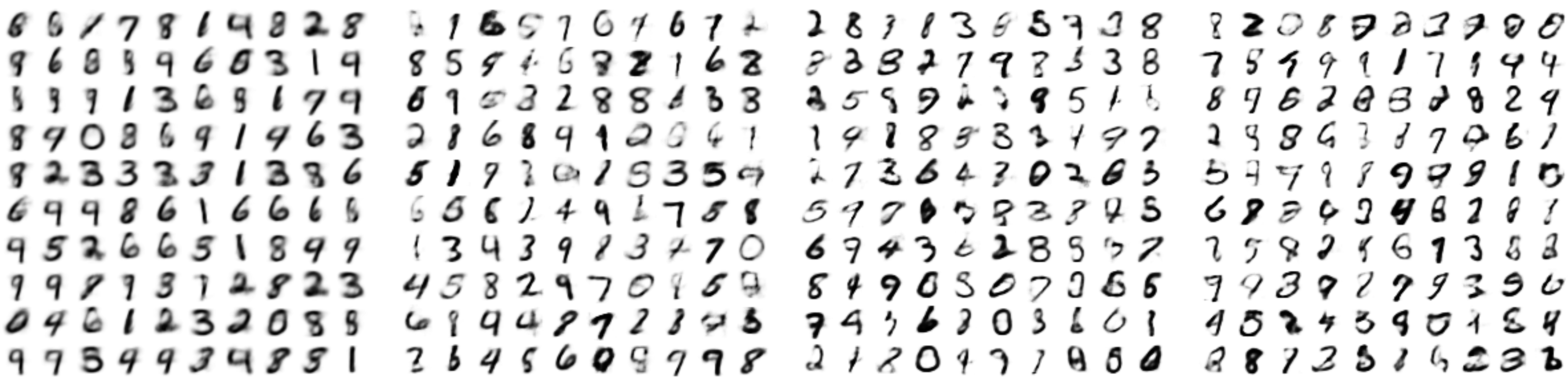
(a) Learned Frey Face manifold



(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:

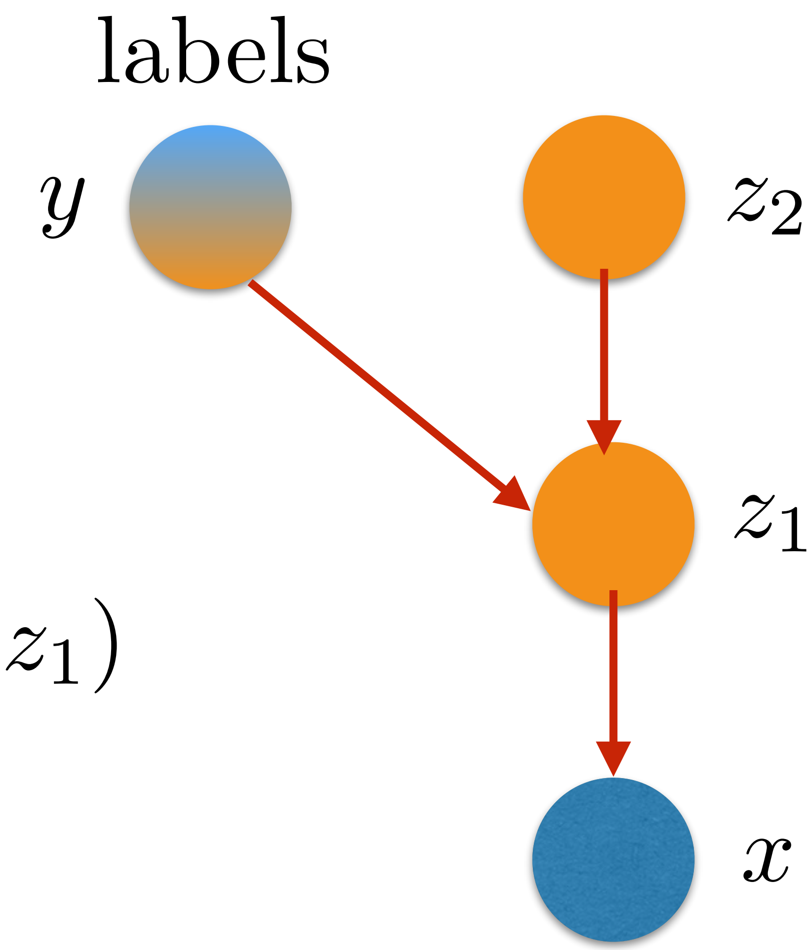
We observe $\{x_i\}_{i \leq L_1}$ and $\{x_j, y_j\}_{j \leq L_2}$, with $x_i \sim p(x)$, $x_j \sim p(x)$.

$$L_1 \gg L_2$$

Extensions 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.

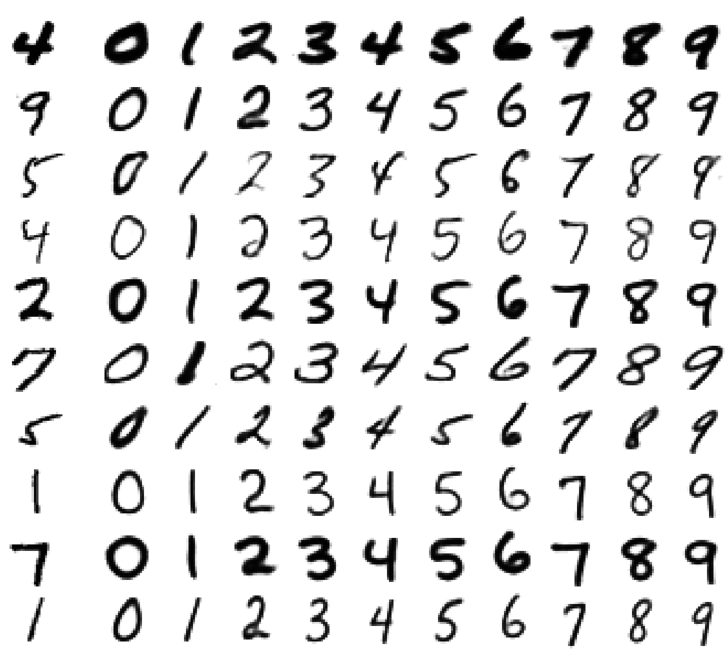
N	NN	CNN	TSVM	CAE	MTC	AtlasRBF	M1+TSVM	M2	M1+M2
100	25.81	22.98	16.81	13.47	12.03	8.10 ($\pm$ 0.95)	11.82 ($\pm$ 0.25)	11.97 ($\pm$ 1.71)	**3.33** ($\pm$ 0.14)
600	11.44	7.68	6.16	6.3	5.13	–	5.72 ($\pm$ 0.049)	4.94 ($\pm$ 0.13)	**2.59** ($\pm$ 0.05)
1000	10.7	6.45	5.38	4.77	3.64	3.68 ($\pm$ 0.12)	4.24 ($\pm$ 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)

- Now there are stronger models on that task.
  - Ladder-Networks
  - GANs.
  - Graph Neural Networks

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.

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

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

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

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

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

- 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

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

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

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

- It results that

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

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