Stat 212b: Topics in Deep Learning
Lecture 15

Joan Bruna
UC Berkeley

Berkeley
UNIVERSITY OF CALIFORNIA

Today

- Reminder:

Review: Unsupervised Learning

- Given high-dimensional data $X = (x_1, \ldots, x_n)$ we want to estimate a low-dimensional model characterizing the population.

- Why is this an important problem?
- It is an essential building block in most high-dimensional prediction tasks.
  - Inverse Problems (super-resolution, inpainting, denoising, etc.).
  - Structured Output Prediction (translation, Q&A, pose estimation, etc.)
  - "Disentangling" or Posterior Inference.
  - Learning with few labeled examples

Review: Curse of Dimensionality

- *Challenge*: How to model $p(x)$, $x \in \mathbb{R}^N$ (or $x \in \Omega^N$) for large N ?

- An existing hypothesis is that, although the ambient dimensionality is high, the *intrinsic* dimensionality of x is low.

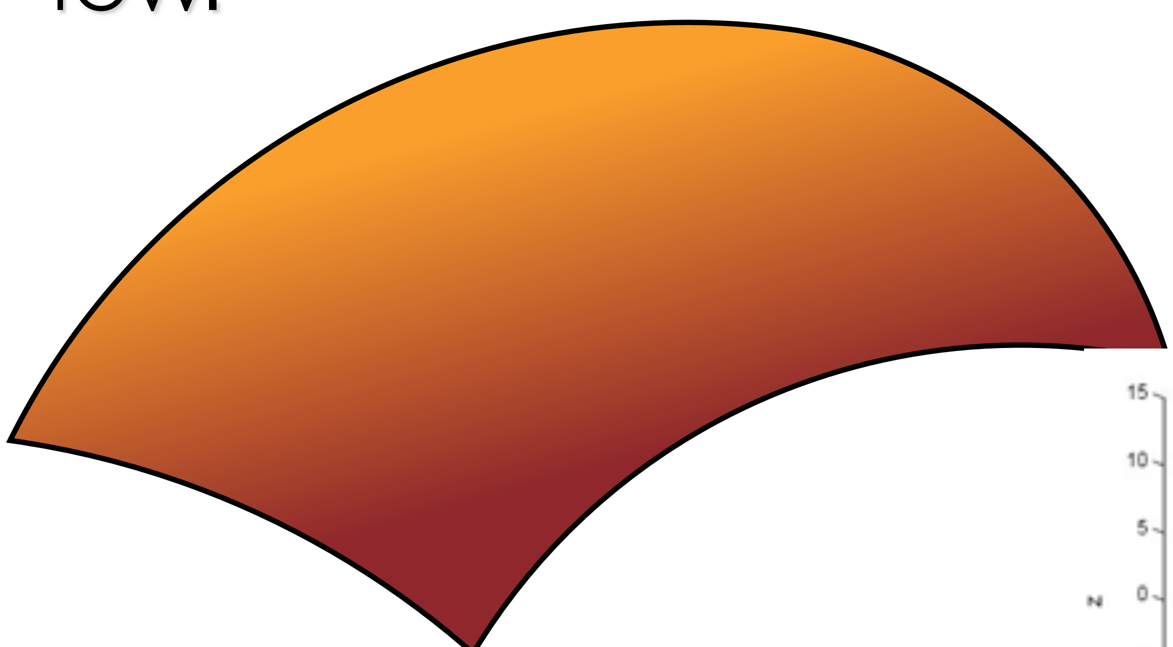
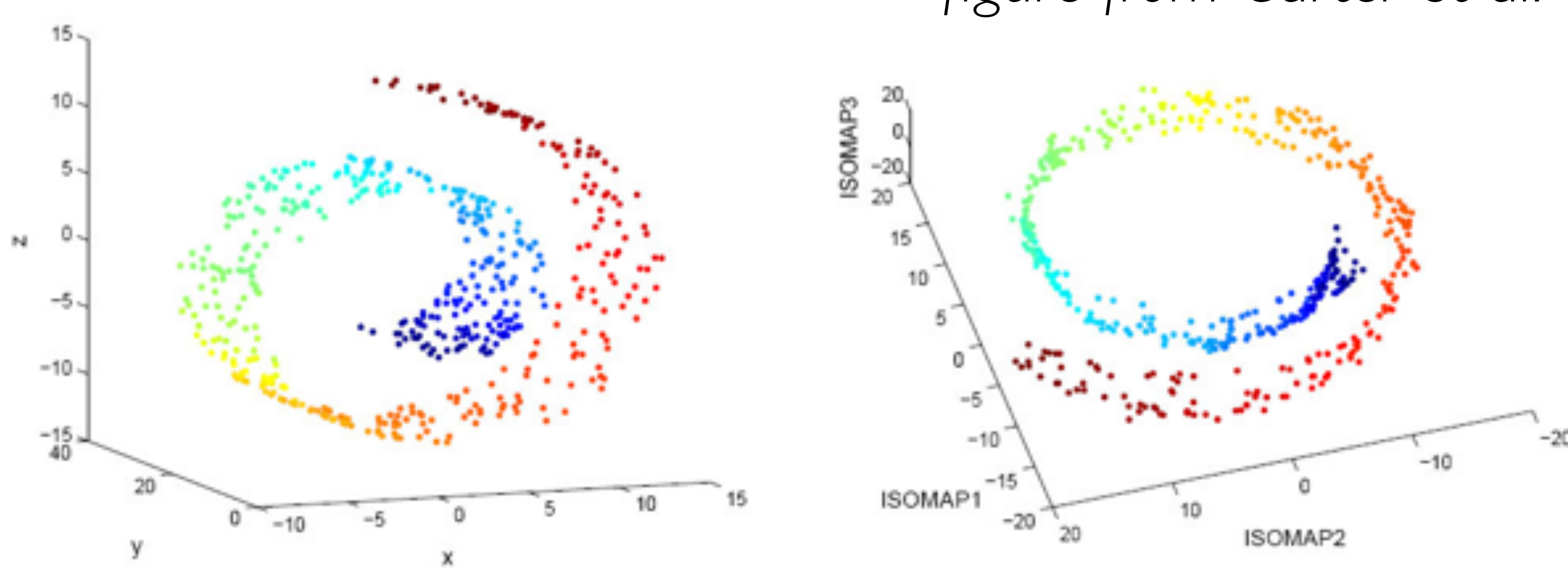


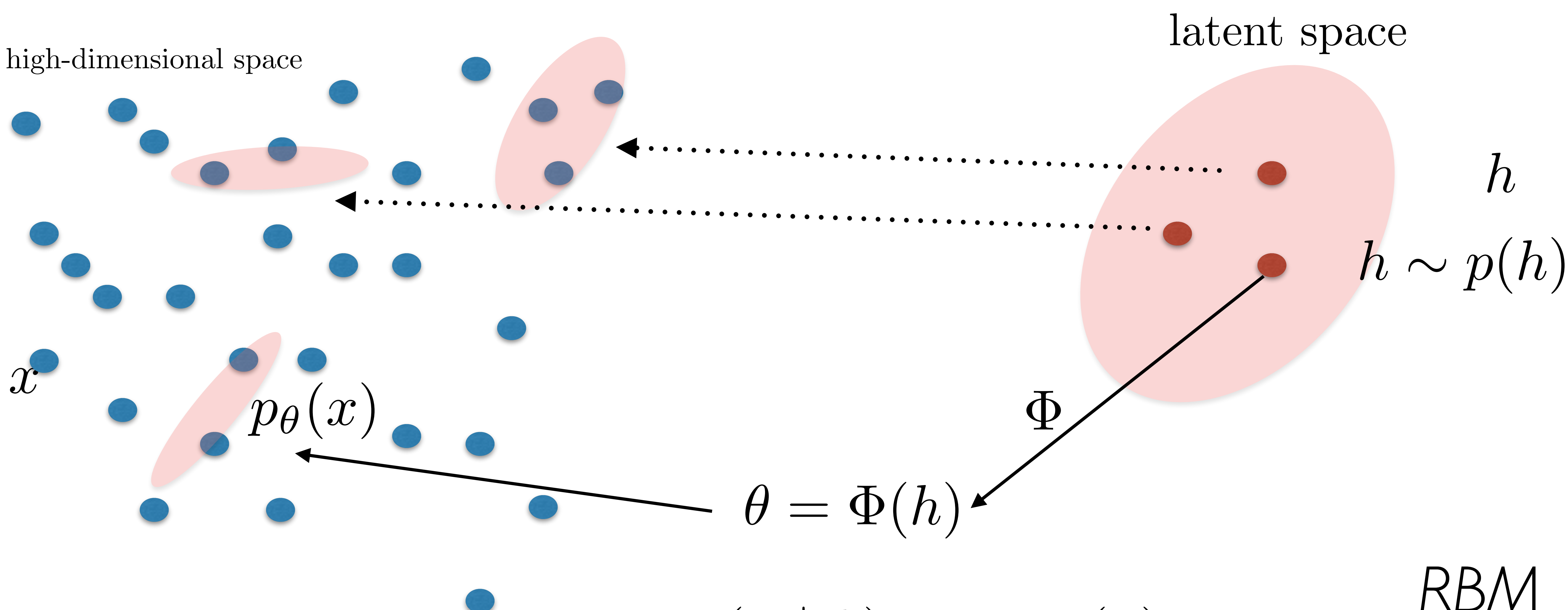
figure from Carter et al.



(a) Swiss Roll (b) Isomap embedding

Review: Latent Graphical Models

- Latent Graphical Models or *Mixtures.*



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

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

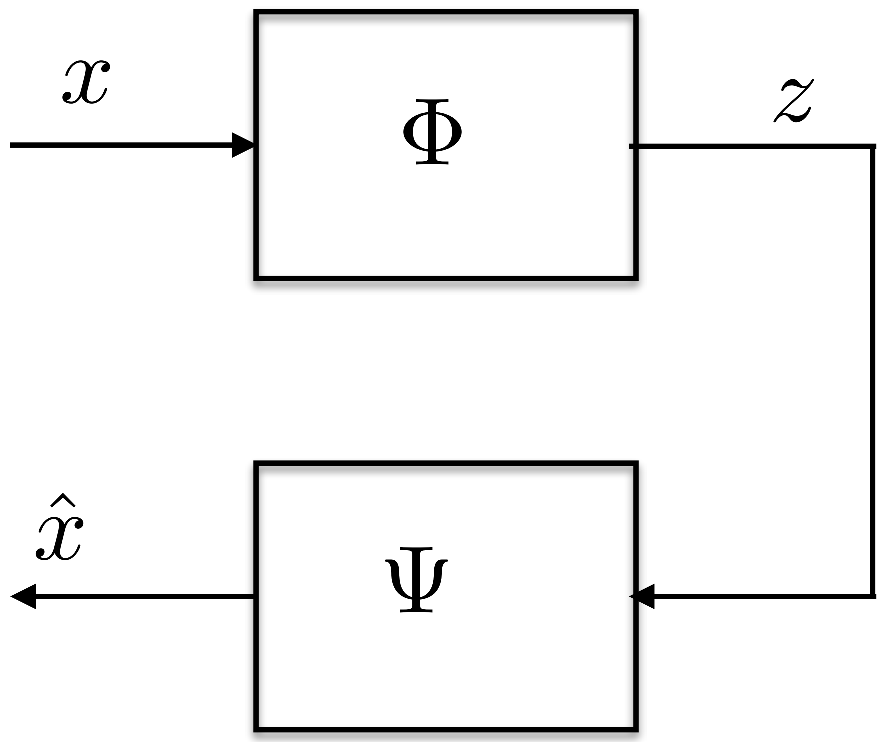
RBM
DBN
DBM
VAE
...

Objectives

- Auto encoders and manifold learning.
- The EM algorithm
- Variational Inference in Exponential Families
- Variational Autoencoders

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):

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

linearization in transformed domain
Slow Feature Analysis

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

- Q: How to limit the reconstruction capacity?

Auto encoders

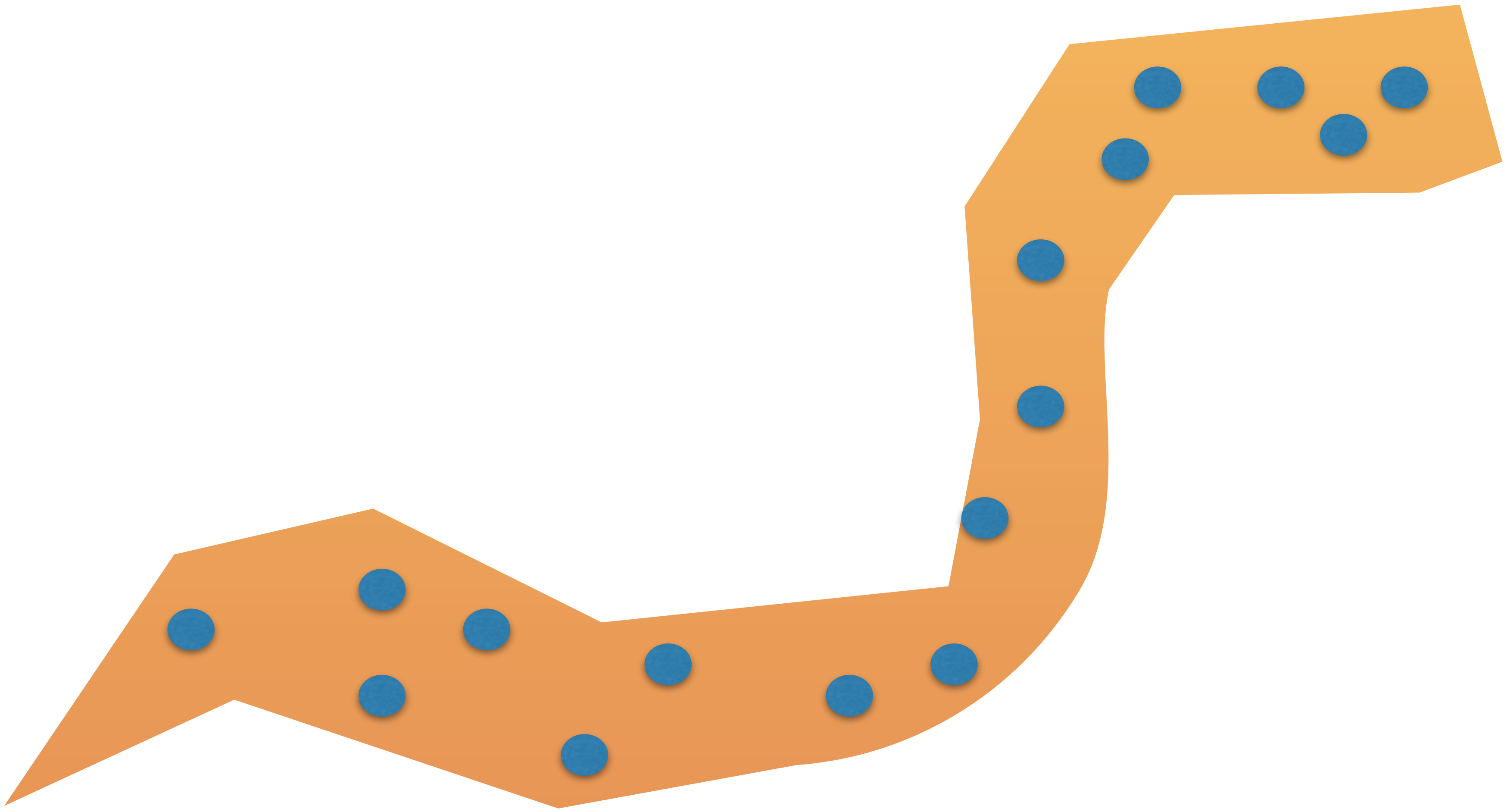
- Optimization set-up:

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

$\ell(x, x')$: 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.

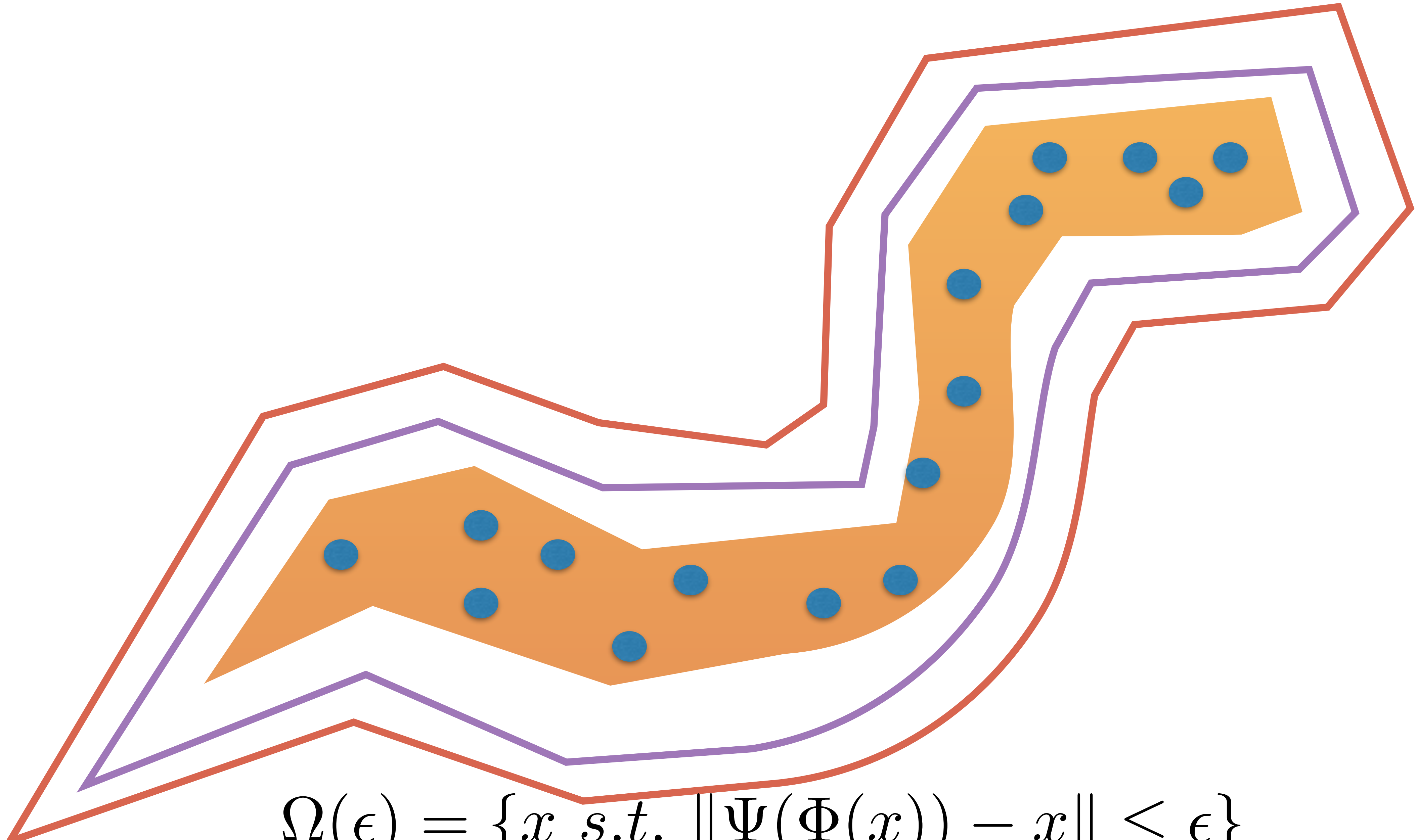
Auto encoders: Geometric Interpretation



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

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

Auto encoders: Geometric Interpretation



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

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

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

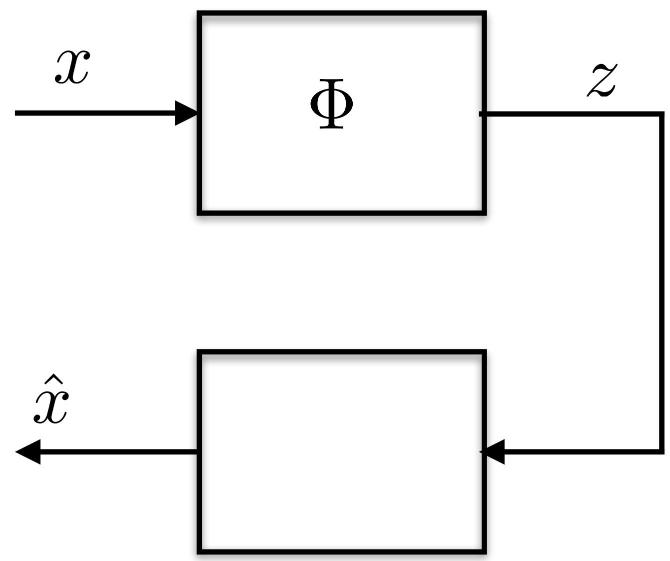
  - Predictive Sparse Decomposition (PSD) [Kavockoglu et al.,'08] considers an Augmented Lagrangian of the Sparse Autoencoder:

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

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

  - 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

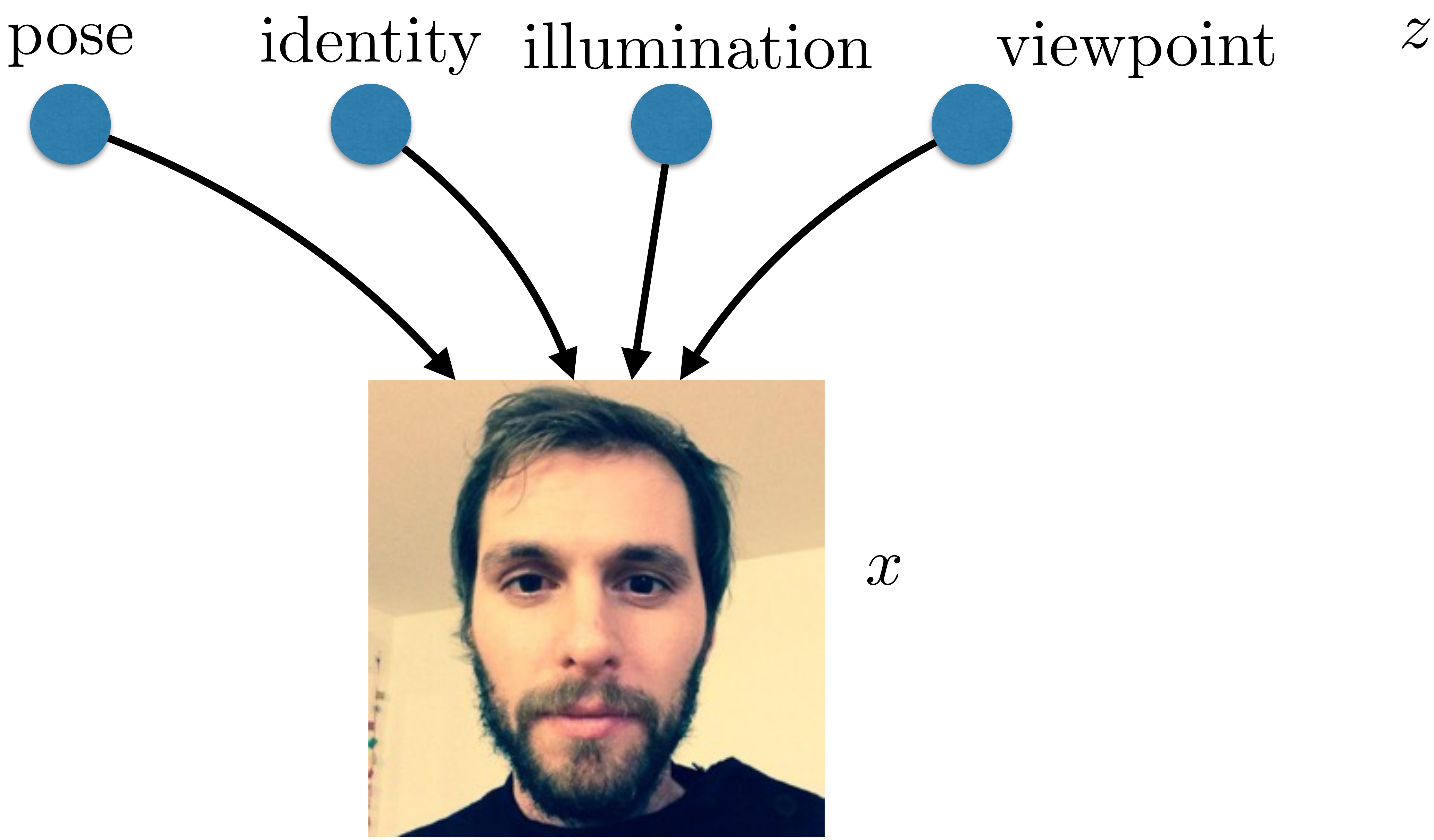
$$p(z \mid x) = \frac{p(z)p(x|z)}{\int p(z')p(x|z')dz'}$$

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

- Q: How to perform "correct" posterior inference?

Approximate Posterior Inference

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



- How to infer z given x ?

The EM algorithm

- It is designed to find MLE solutions of latent variable models.
- In general, we have log-likelihoods of the form

$$\log p(X \mid \theta) = \log \left(\sum_Z p(X, Z \mid \theta) \right) , \quad \begin{array}{l} \theta = \text{model parameters} \; . \\ Z = \text{latent variables} \end{array}$$

The EM algorithm

- It is designed to find MLE solutions of latent variable models.
- In general, we have log-likelihoods of the form

$$\log p(X \mid \theta) = \log \left(\sum_Z p(X, Z \mid \theta) \right) \quad , \quad \begin{aligned} \theta &= \text{model parameters} \; . \\ Z &= \text{latent variables} \end{aligned}$$

- Using current parameters θ_{old} , we compute the expected total likelihood of the model (E-step):

$$\mathcal{Q}(\theta, \theta_{old}) = \mathbb{E}_{Z \sim p(Z \mid X, \theta_{old})} \log p(X, Z \mid \theta)$$

- Then we update the parameters to maximize this likelihood:

$$\theta_{new} = \arg\max_\theta Q(\theta, \theta_{old}) \; .$$

EM and Variational Bound

- Q: Does this algorithm monotonically improve the likelihood?
- Assume for now that latent variables are discrete.
- For any distribution $q(Z)$ over latent variables, we have

$$\log p(X \mid \theta) = \log \left(\sum_Z p(X, Z \mid \theta) \right) = \log \left(\sum_Z q(Z) \frac{p(X, Z \mid \theta)}{q(Z)} \right)$$

$$\geq \sum_Z q(Z) \log \left(\frac{p(X, Z \mid \theta)}{q(Z)} \right) = \mathcal{L}(q, \theta) \ .$$

$$(\text{Jensen's Inequality: } \mathbb{E}(f(X)) \geq f(\mathbb{E}(X)) \text{ if } f \text{ is convex })$$

Variational Bound

- We can express the variational lower bound as

$$\begin{aligned}\mathcal{L}(q,\theta) &= \mathbb{E}_{q(Z)}\left[\log p(X,Z \mid \theta)\right] - \mathbb{E}_{q(Z)} \log q(Z) \\ &= \mathbb{E}_{q(Z)}\left[\log p(X,Z \mid \theta)\right] + H(q)\ .\end{aligned}$$

$H(q)$: Entropy of $q(Z)$.

- Also, we have

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

$$KL(q||p) = -\sum_z q(z)\log\left(\frac{p(z)}{q(z)}\right)$$

is the Kullback-Leibler divergence.

Variational Bound

- Thus, the divergence $KL(q||p)$ measures how far our variational approximation $q(z)$ is from the true posterior, and directly controls the bound on the log-likelihood.

- Using

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

- E-step: maximize lower bound $\mathcal{L}(q, \theta)$ with respect to q, holding parameters fixed.
- M-step: maximize lower bound $\mathcal{L}(q, \theta)$ with respect to parameters, holding q fixed.

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:

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

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

- 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

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

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

- It is well defined for the family of parameters

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

Exponential Families

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

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.

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 X and latent variables Z, 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 θ 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:

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

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

- 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:

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

- ***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

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

- 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.

- Faster alternative?

Variational Autoencoders

- 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

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

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

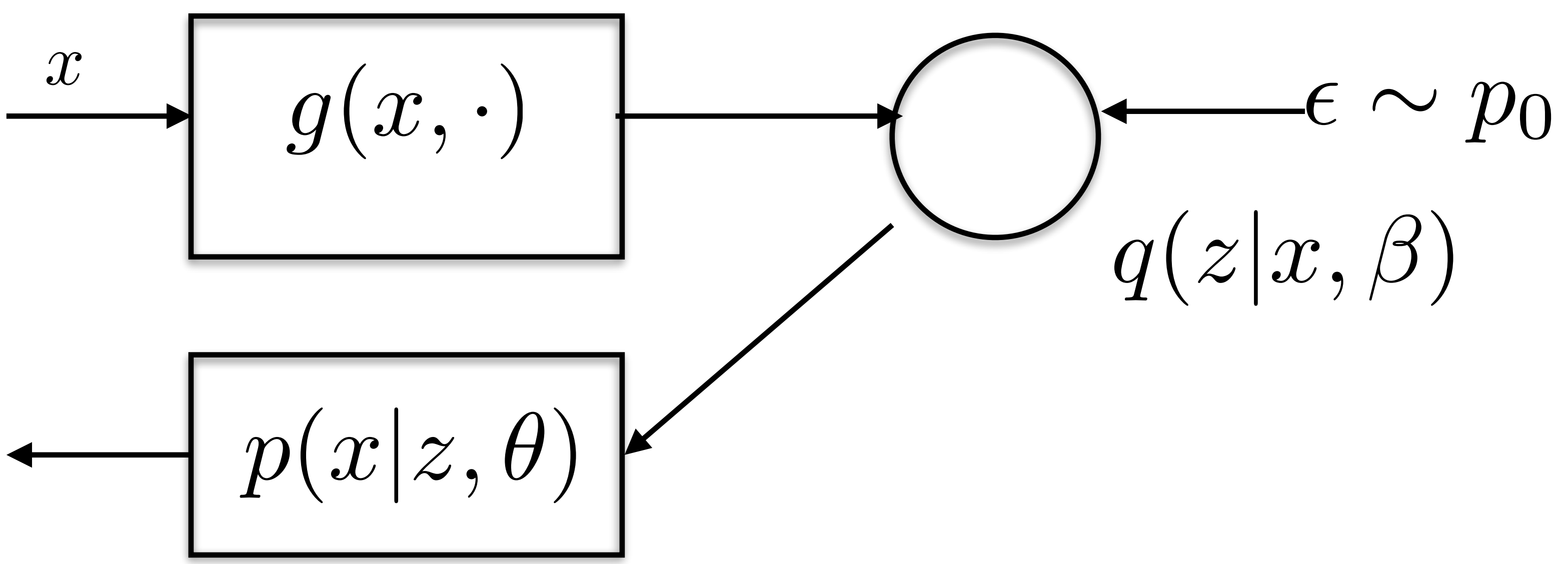
$$\widehat{\mathcal{L}(\theta, \beta, X)} = -D_{KL}(q_\beta(z|X)||p_\theta(z)) + \frac{1}{S}\sum_{s \leq S} \log p(X|z^{(s)}, \theta)$$

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

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

Variational Autoencoders

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



Variational Autoencoder

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

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

- 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:

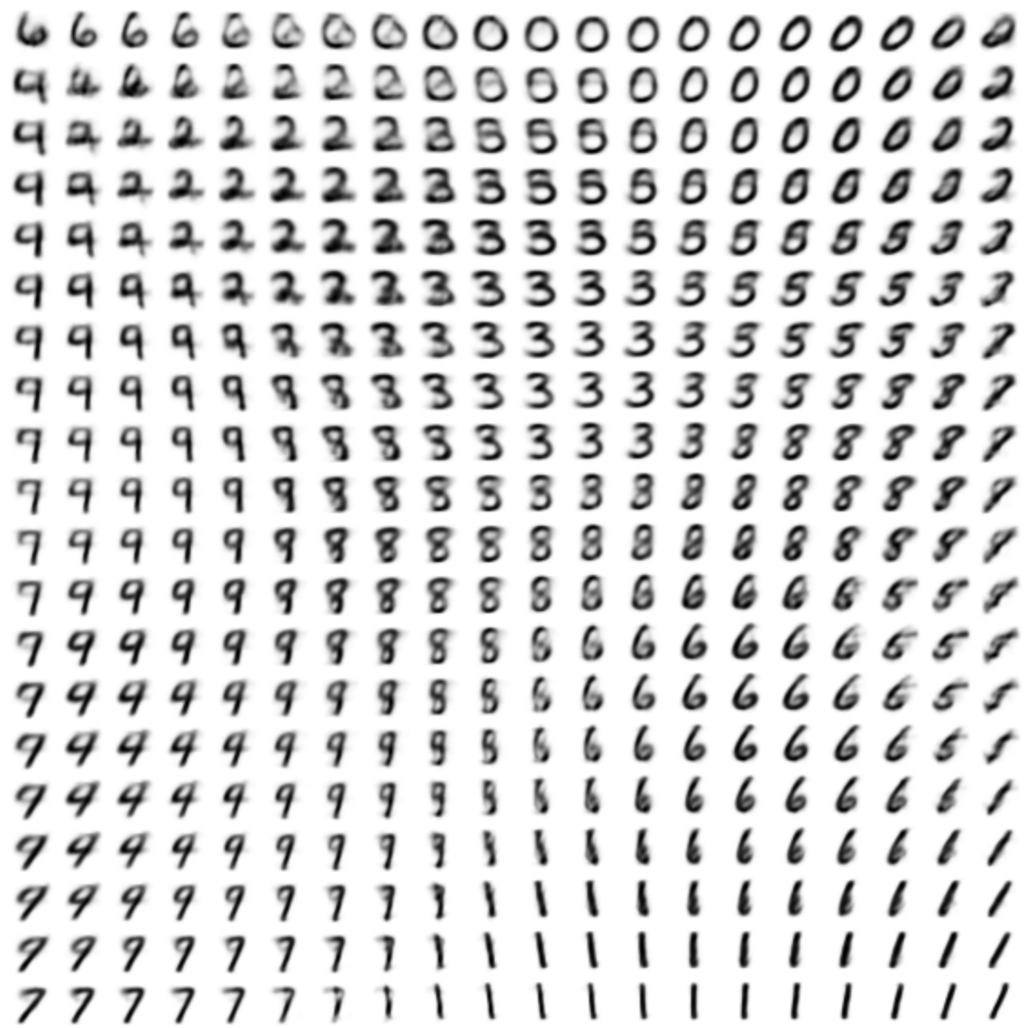
$$q_\beta(z|x) = \mathcal{N}(z; \overline{\mu}(x), \overline{\Sigma}(x))$$

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

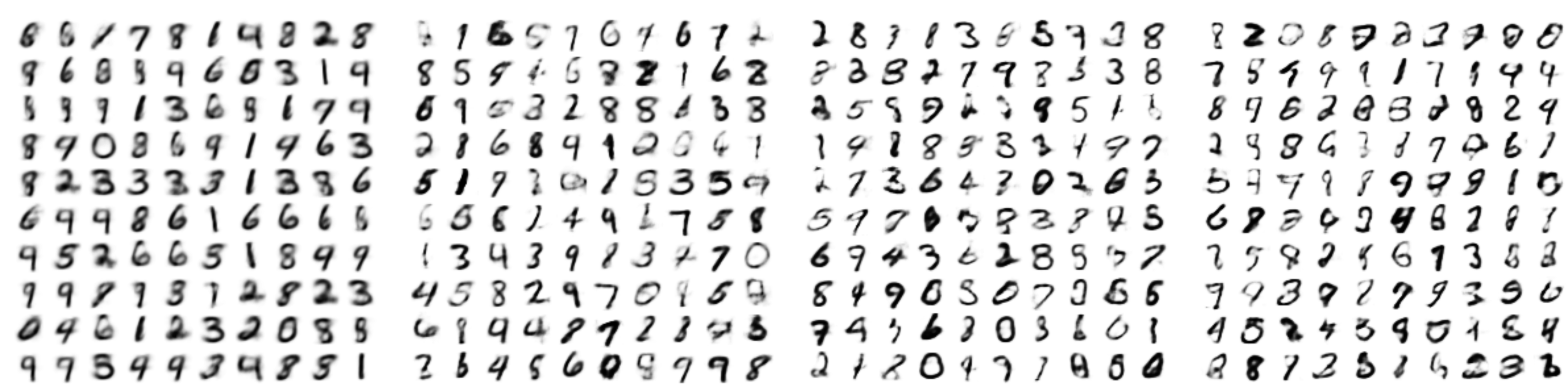
$$Z \sim q_\beta(z|x) \Leftrightarrow Z = \overline{\mu}(x) + \overline{\Sigma}(x)\epsilon \;,\; \epsilon \sim \mathcal{N}(0, \mathbf{1})$$

Examples

(a) Learned Frey Face manifold



(b) Learned MNIST manifold



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

Extensions

- Importance Sampling Variational Autoencoders