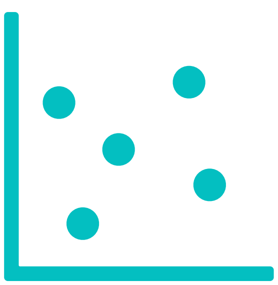
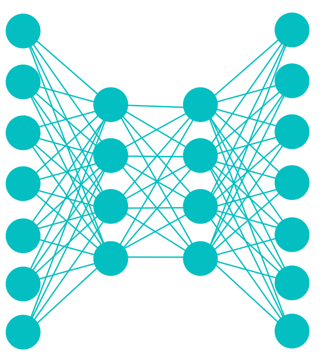
Lecture Notes for

Neural Networks and Machine Learning



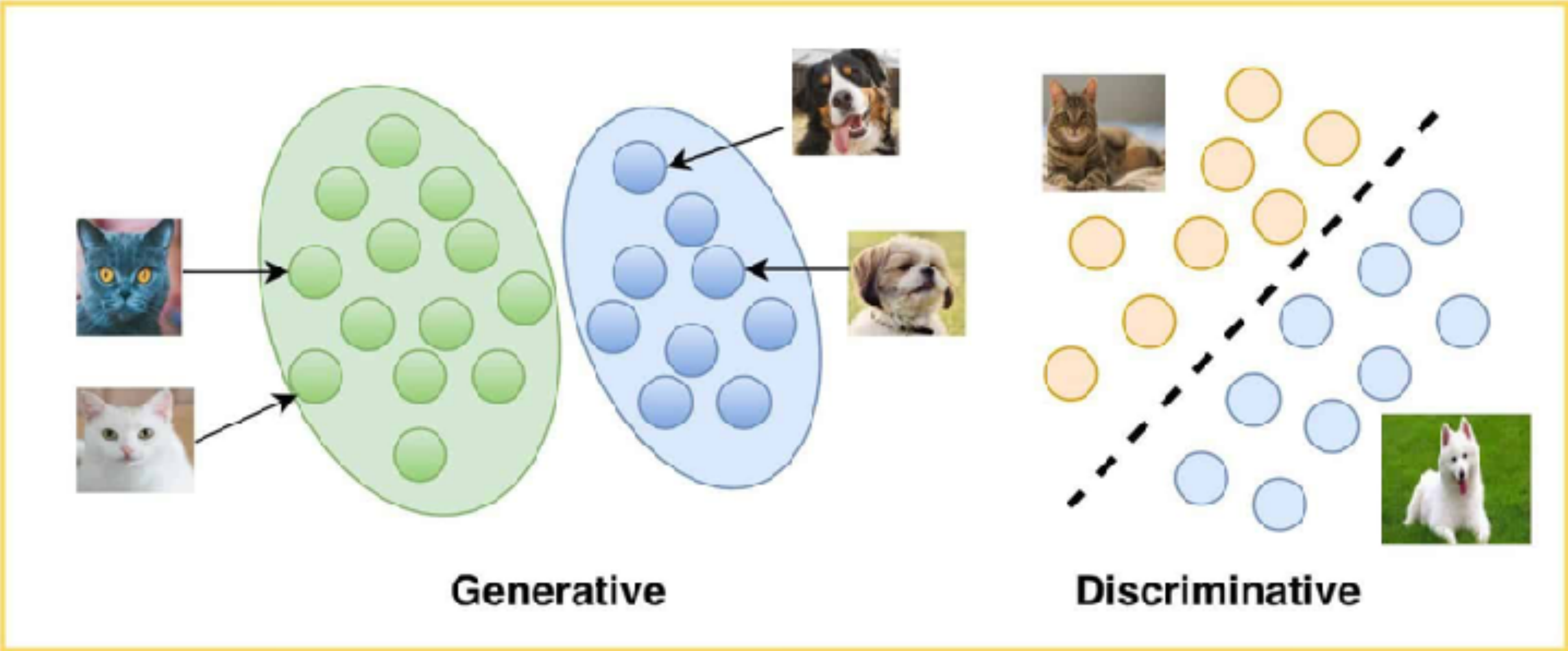
Generative Networks and Auto-Encoding Generators



Logistics and Agenda

- Logistics
  - Office Hours, 12:30-1:30
  - Lab due date
  - Student paper presentation
- Agenda
  - A historical perspective of generative Neural Networks
  - Variational Auto-Encoding
  - VAE in Keras Demo (if time)
  - Adversarial Auto-Encoders (if time)

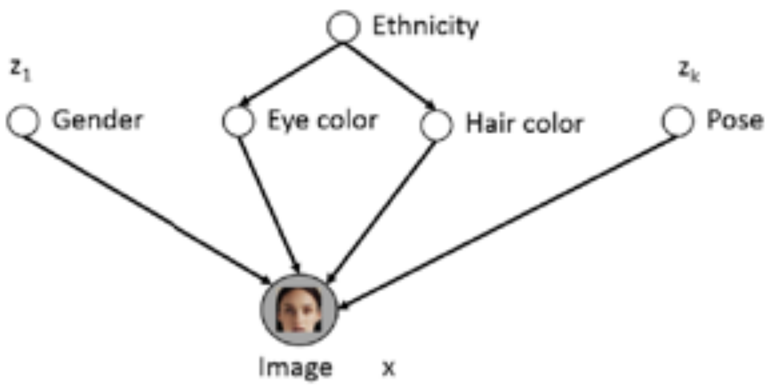
Generative versus Discriminative



https://learnopencv.com/generative-and-discriminative-models/

Motivations: Generative Latent Variables

Latent Space Variables

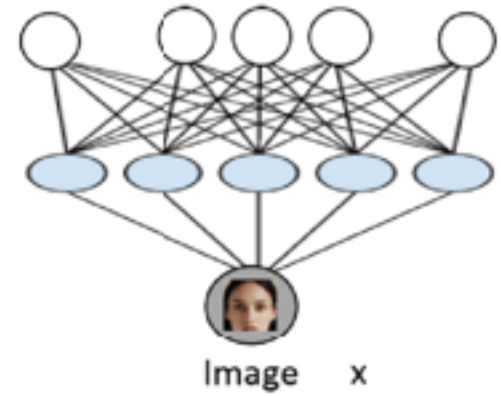


$$p(\mathbf{x} \mid \mathbf{z})$$

Output Observation (e.g., image)

Hard: **z** is expertly chosen

Latent Space Variables



$$p(\mathbf{x} \mid \mathbf{z})$$

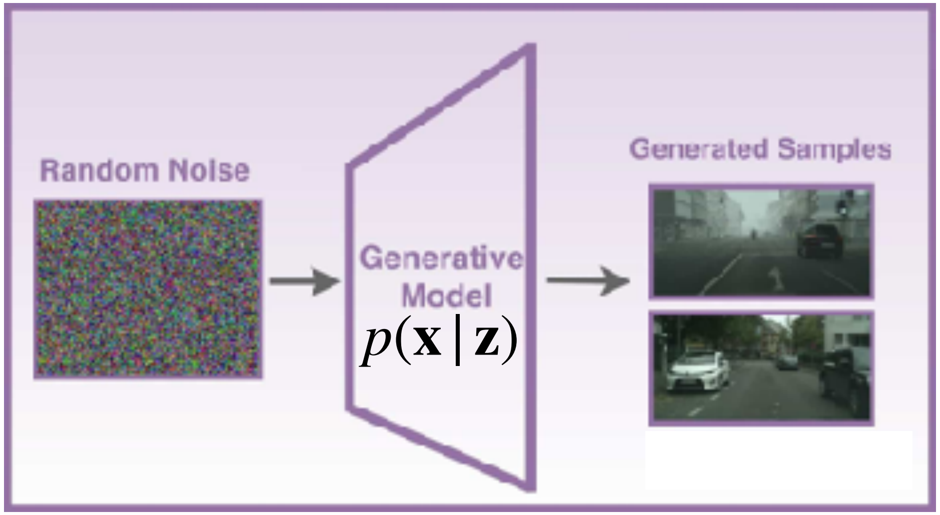
Output Observation (e.g., image)

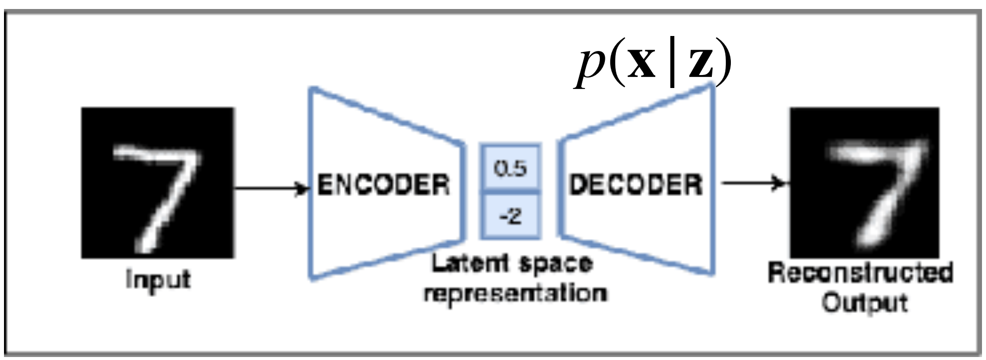
Not as Hard: **z** is trained, latent variables are uncontrolled

Want: $$p(\mathbf{x}) \approx \sum_{\mathbf{z}} p(\mathbf{z}) p_\theta(\mathbf{x} \mid \mathbf{z})$$

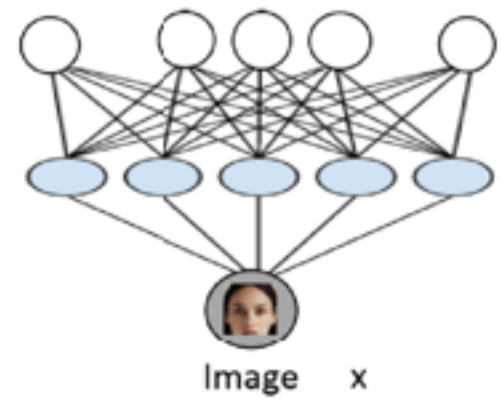
http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture13.pdf

Motivations: Generative Latent Variables





Latent Space Variables



$$p(\mathbf{x}\,|\,\mathbf{z})$$

Output Observation (e.g., image)

Not as Hard: **z** is trained, latent variables are uncontrolled

Want: $$p(\mathbf{x}) \approx \sum_{\mathbf{z}} p(\mathbf{z}) p_\theta(\mathbf{x}\,|\,\mathbf{z})$$

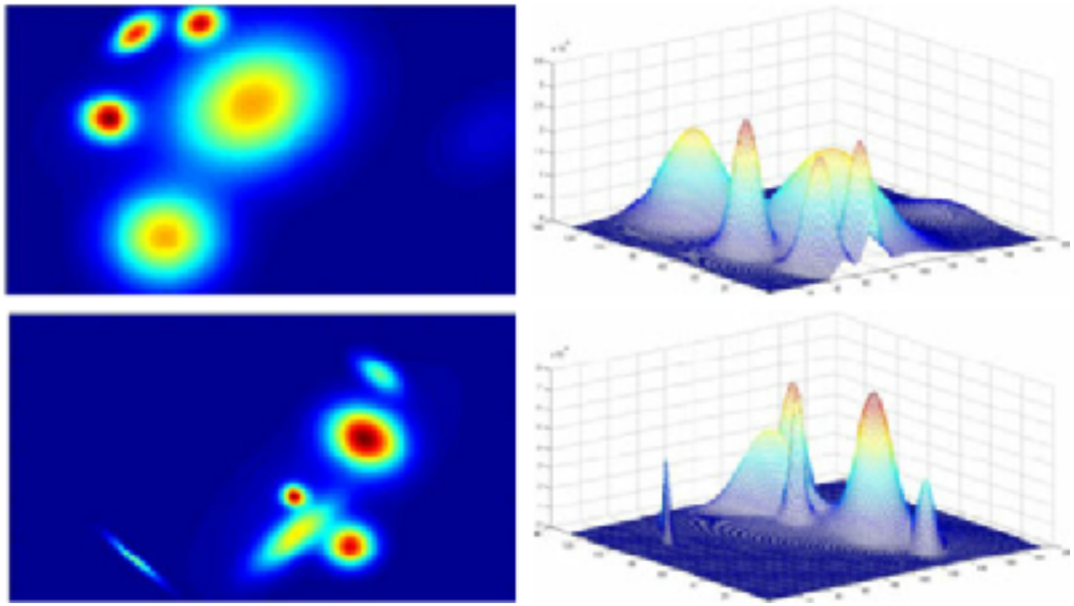
https://learnopencv.com/generative-and-discriminative-models/

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture13.pdf

Motivation: Mixtures for Simplicity

Want: $$p(\mathbf{x}) \approx \sum_{\mathbf{z}} p(\mathbf{z}) p_\theta(\mathbf{x} \,|\, \mathbf{z})$$

- Each latent variable is mostly independent of other latent variables
- The sum of various mixtures can approximate most any distribution
- Good choice for conditional is Normal Distribution
- Can parameterize $p(x|z)$ to be a Neural Network



$$p_\theta(\mathbf{x} \,|\, \mathbf{z} = k) = \mathcal{N}(\mathbf{x}, \mu_k, \Sigma_k)$$

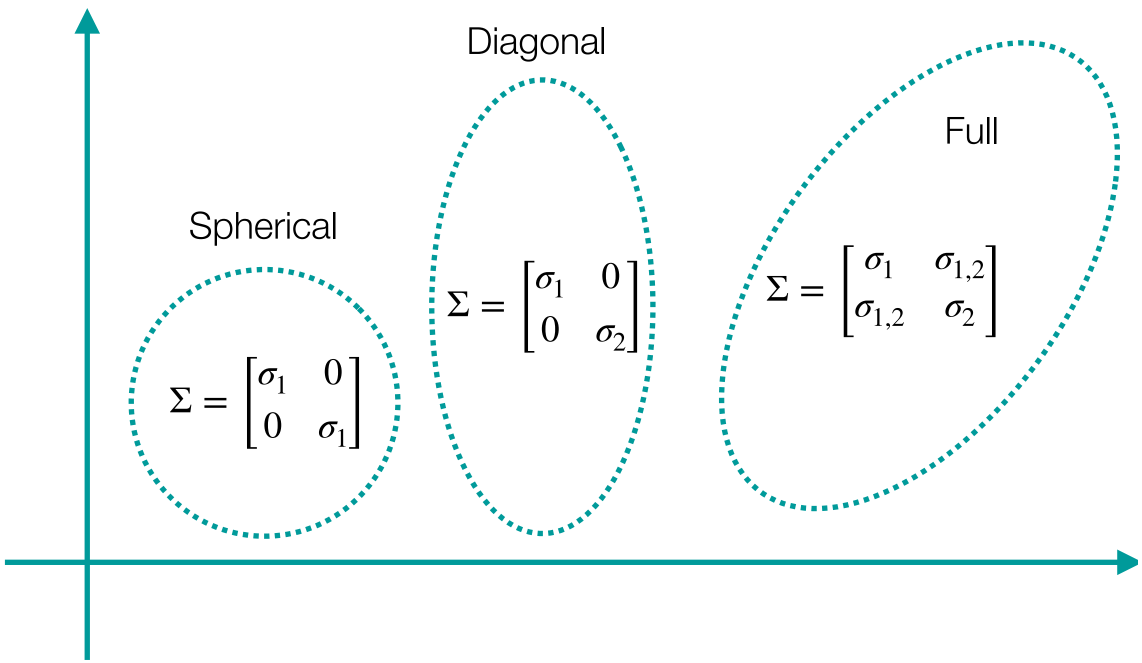
mean and covariance learned

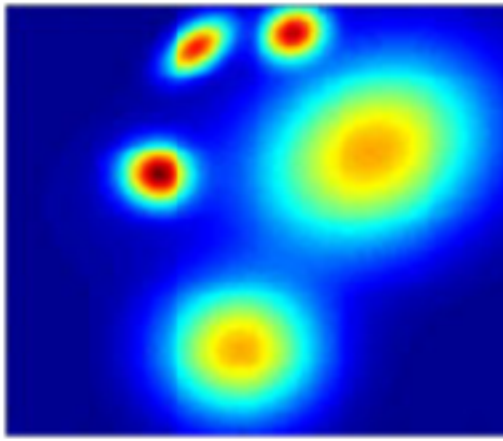
http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture13.pdf

Motivation: Mixtures for Simplicity

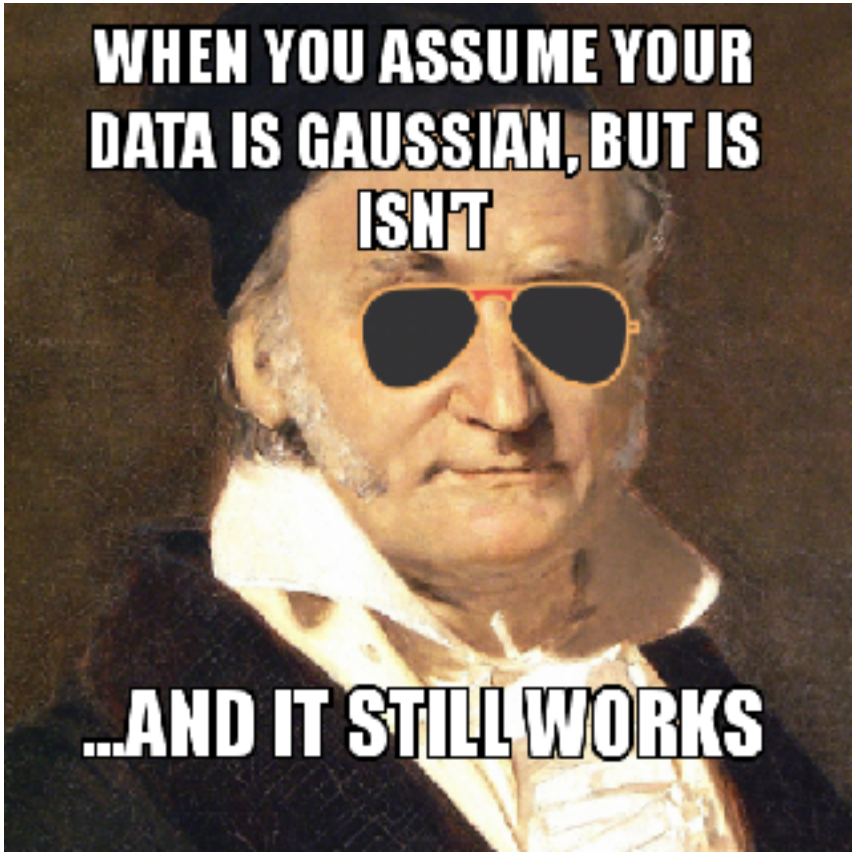
$$= \mathcal{N}(\mathbf{x}, \mu_k, \Sigma_k)$$

mean and covariance learned

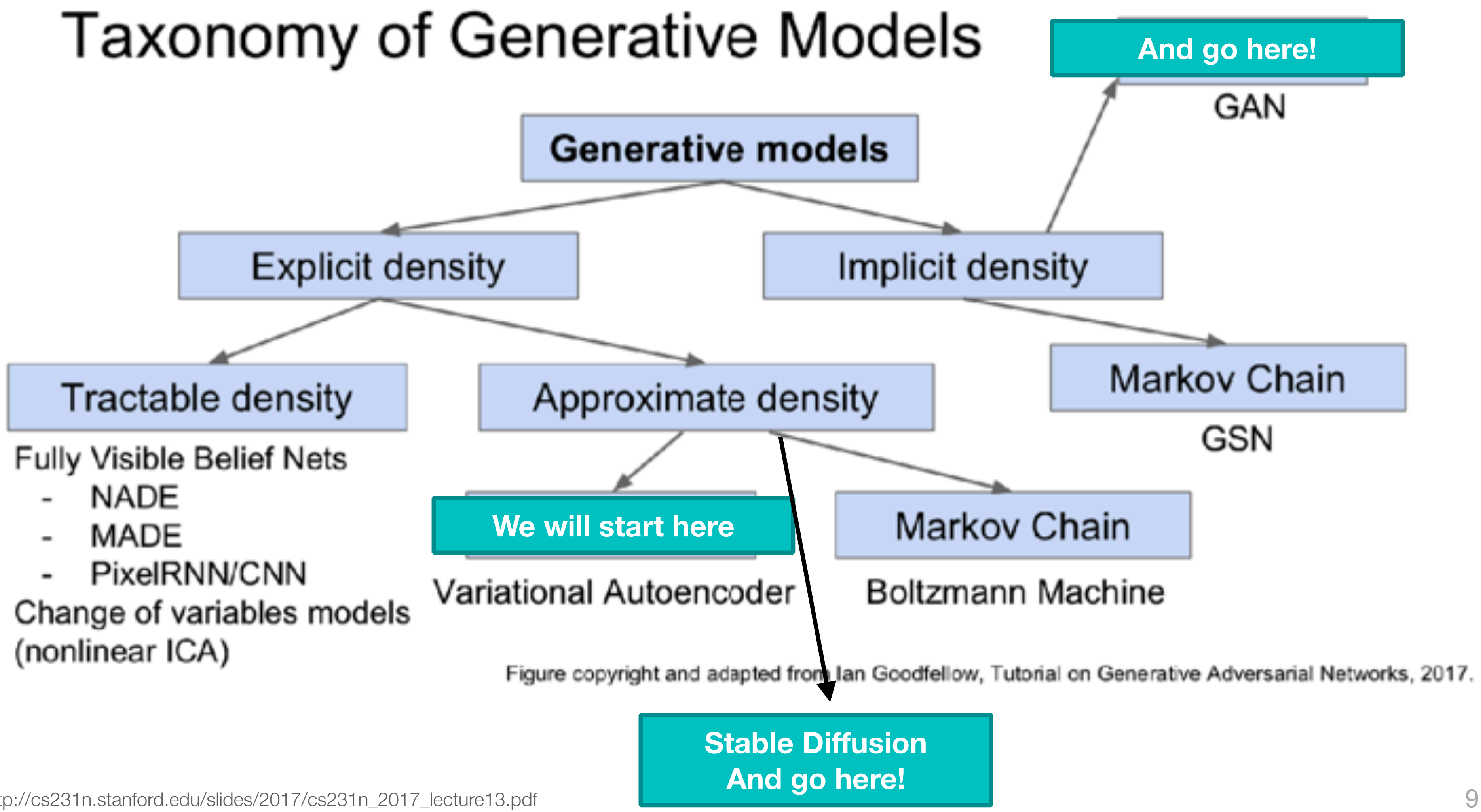




A History of Generative Networks

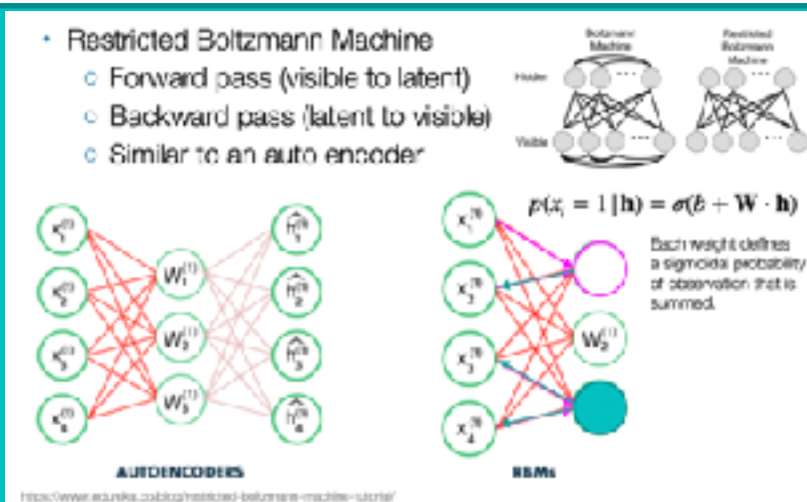


Taxonomy of Generative Models

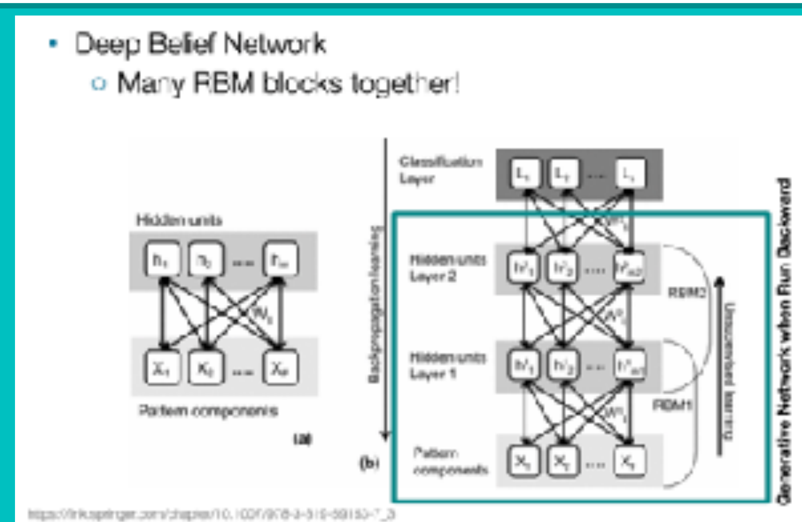


http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture13.pdf

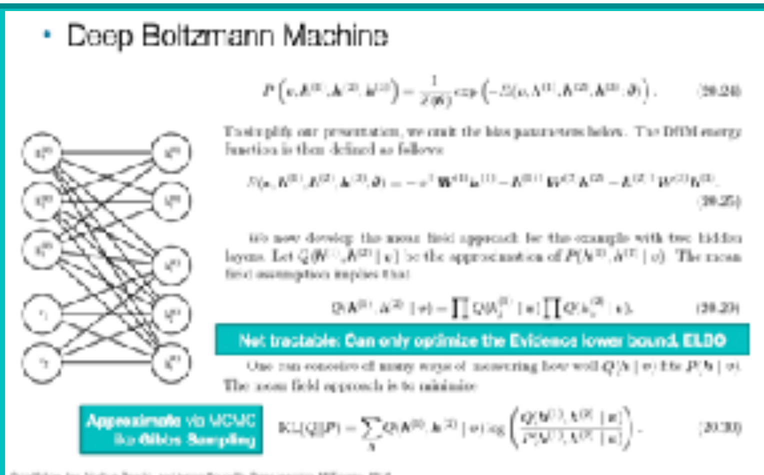
Abridged History of Generative Networks



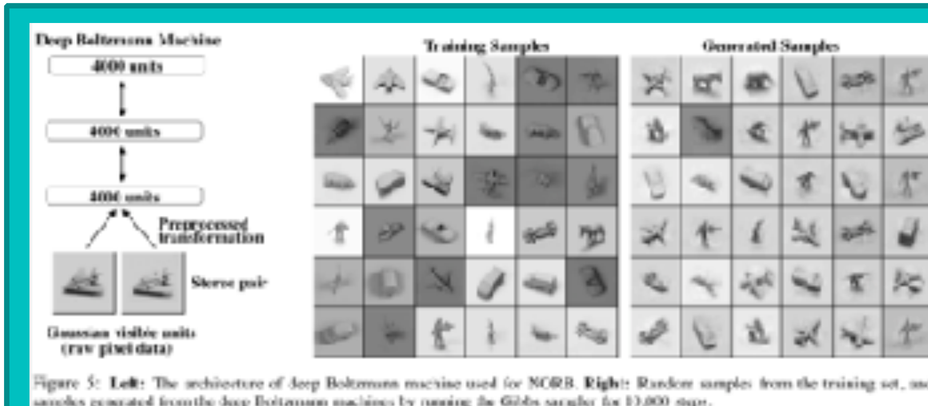
2006 Restricted Boltzmann Machine



2007, Deep Belief Networks
RBMs with many layers

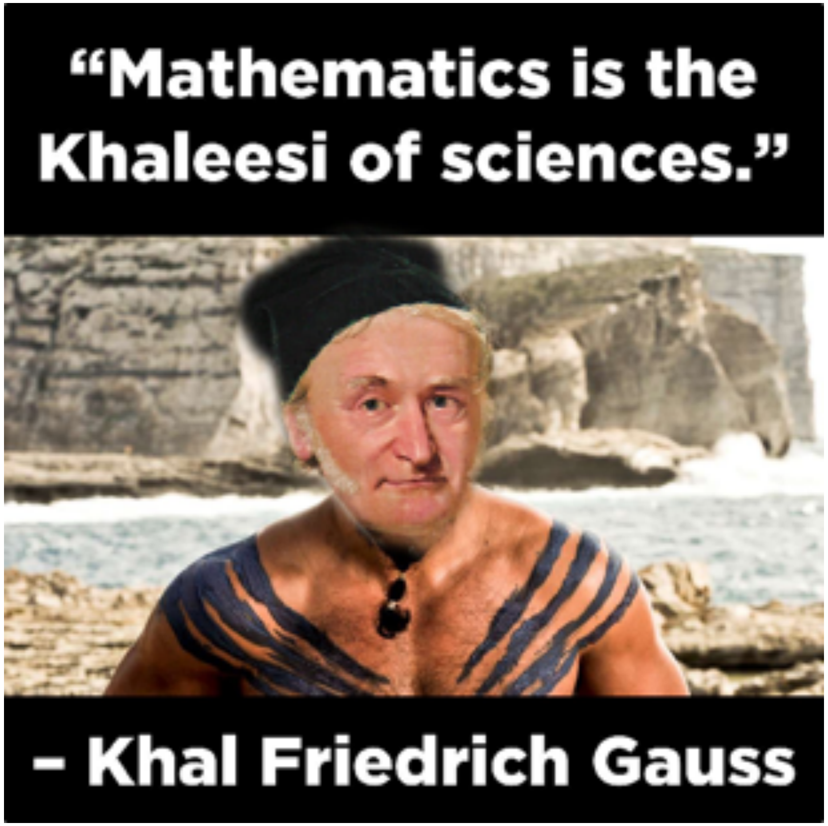


2009 Deep Boltzmann Machine
Goodfellow, Bengio, Courville



2009, Practical Examples
Salakhutdinov and Hinton

Variational Auto Encoding



Aside: Remember These

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

some function

could be neural networks

$$\mathbf{E}_{s \leftarrow q(s|x)}[f(\cdot)] = \int q(s|x) \cdot f(x)\,dx \approx \sum_{\forall i} q(s|x^{(i)}) \cdot f(x^{(i)})$$

Expected value of f under conditional distribution, q

s is latent variable, $x^{(i)}$ is an observation

$$\therefore\ \mathbf{E}_{s \leftarrow q(s|x)}[\log f(\cdot)] = \sum_{\forall i} q(s|x^{(i)}) \cdot \log\left(f(x^{(i)})\right)$$

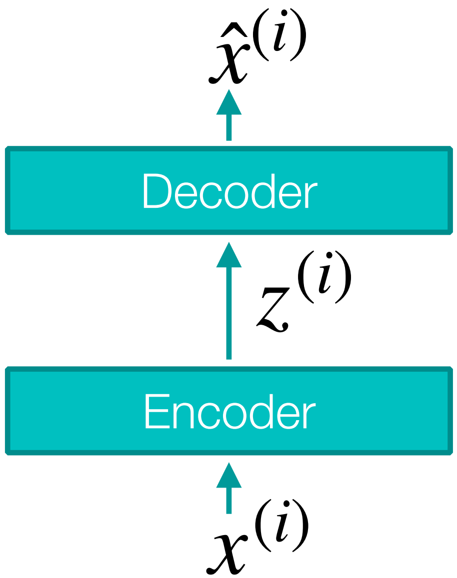
If function is a probability, this is just the negative of cross entropy of distributions:

$$H(q,p) = -\sum_{x} q(x) \cdot \log(p(x))$$

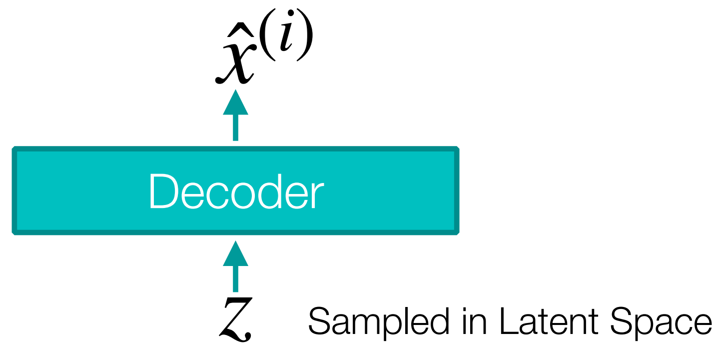
Recall that KL divergence is a measure of difference in two distribution, and is just:

$$D(p||q) = \sum_{x} p(x) \cdot \log\left(\frac{p(x)}{q(x)}\right) = \mathbf{E}_p\left[\log\left(\frac{p(x)}{q(x)}\right)\right]$$

Can Auto Encoding Generate Samples?



Once trained, is it possible to generate data?



- Does this work for simple auto encoding?
  - Yes, but not satisfactory results
- Learned space is not continuous
- Features could be highly correlated, related in complex ways
  - So, how to sample from the latent space?
- Need to define constraints on latent space…

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture13.pdf

Reasonable constraints for p(z)?

$\hat{x}^{(i)}$

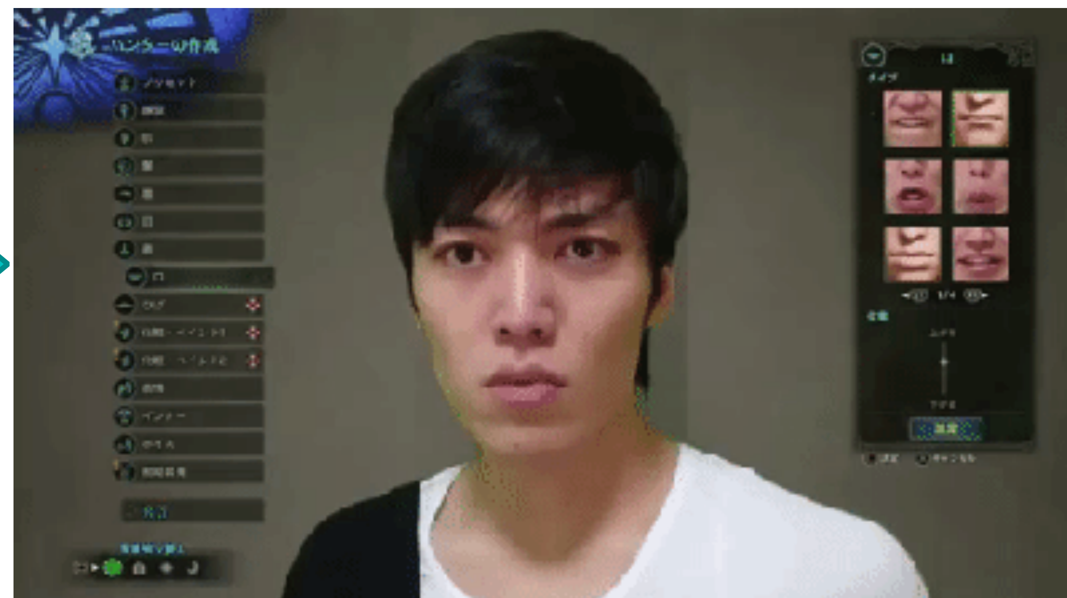
Decoder

z Sampled in Latent Space

- Should be simple, easy to sample from: **Normal**
- Each component should be independent and identically distributed (i.i.d).: **Diag. Covariance**
  - Encourages features that may be semantic, like expert might select

Ideal:
Smiling
Hair Color
Skin Color
Eye Color
...
Nose Position

z → Transform with Decoder → $\hat{x}^{(i)}$



Mathematical Motivation

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

We need this inference in order to compute latent variable

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

Denominator is of this form

- We can't compute! **Intractable computation** for all "z"
- So let's define this with **variational inference**:
  - AKA: Find the best approximation of desired distribution using a parametrized set of distributions (usually normal distributions)
  - Only needs to work for z **with observed** $x^{(i)}$
  - 1. **Encode** observed $x^{(i)}$ via network $q(z \mid x^{(i)})$ (with some constraints)
  - 2. Use $q(z \mid x^{(i)})$ to sample z appropriately, then **decode** with another neural network, $p(x^{(i)} \mid z^{(i)})$
  - 3. Make $q(z \mid x^{(i)})$ largest probability possible via Gaussian Distributions

KL Divergence

$$p(x) = \frac{p(z,x)}{p(z|x)} = \frac{p(x|z)p(z)}{p(z|x)}$$

$$\log p(x) = \log p(x) \cdot \int q(z)dz$$

$$\log p(x) = \int \log p(x) \cdot q(z)dz$$

$$\log p(x) = \int q(z)\log\left[\frac{p(x,z)}{p(z|x)}\right]dz$$

$$\log p(x) = \int q(z)\log\left[\frac{p(x,z)\cdot q(z)}{p(z|x)\cdot q(z)}\right]dz$$

$$\log p(x) = \int q(z)\cdot\left(\log\left[\frac{p(x,z)}{q(z)}\right] + \log\left[\cdot\frac{q(z)}{p(z|x)}\right]\right)dz$$

$$\log p(x) = \int q(z)\cdot\log\left[\frac{p(x,z)}{q(z)}\right]dz + \boxed{\int q(z)\log\left[\frac{q(z)}{p(z|x)}\right]dz}$$

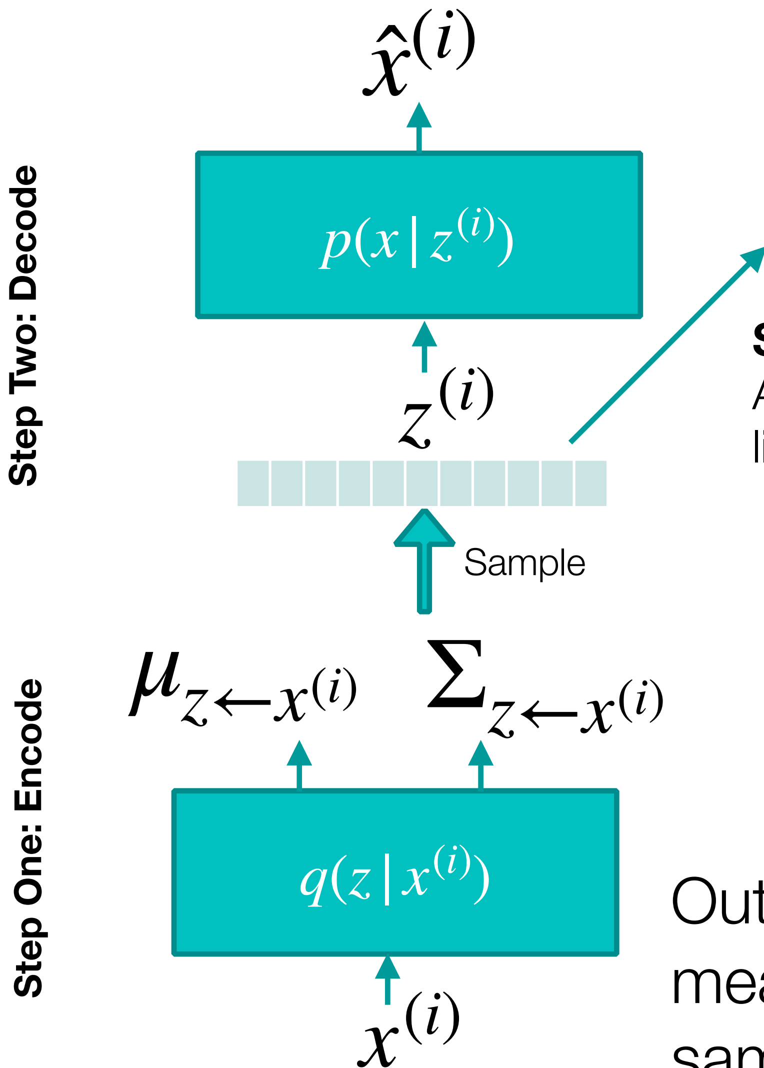
> 0

$$\log p(x) \geq \int q(z)\cdot\log\left[\frac{p(x,z)}{q(z)}\right]dz$$

equal only if
$p(z|x)$ and $q(z)$ are essentially
the same distribution

$$\therefore \min D_{KL}\left[q(z)\|p(z|x)\right]$$

Need a new formulation



Step Three: Make conditional p and q Similar

$$D_{KL}\left[q(z \mid x^{(i)}) \| p(z \mid x^{(i)})\right] = \mathbf{E}_{q(z|x)}\left[\log\left(\frac{q(z \mid x^{(i)})}{p(z \mid x^{(i)})}\right)\right]$$

Step Four: Use Variational Inference
Assume that a family of distributions can maximize likelihood of observing $x^{(i)}$:

$$\log p(x)_{\forall i} \approx \mathbf{E}_{\mathbf{z} \leftarrow q(z|x^{(i)})}\left[\log p(x^{(i)})\right]$$

Max Log Lik:: maximize probability of observed $x^{(i)}$
given family of distributions q
hope this is a good approximation

Output of network, q, are the mean and covariance for sampling a variable z

Need a new formulation

$$\log p(x)_{\forall i} \approx \mathbf{E}_{\mathbf{z} \leftarrow q(z|x)} \left[\log p(x^{(i)})\right]$$ Maximize!

$$= \mathbf{E}_q \left[\log \frac{p(x^{(i)} \mid z) p(z)}{p(z \mid x^{(i)})} \frac{q(z \mid x^{(i)})}{q(z \mid x^{(i)})} \right]$$

Variational + multiply by one

$p(z \mid x^{(i)})$ this is still a problem

$$= \mathbf{E}_q \left[\log p(x^{(i)} \mid z)\right] + \mathbf{E}_q \left[\log \frac{p(z)}{q(z \mid x^{(i)})} \right] + \mathbf{E}_q \left[\log \frac{q(z \mid x^{(i)})}{p(z \mid x^{(i)})} \right]$$

$$= \mathbf{E}_q \left[\log p(x^{(i)} \mid z)\right] - \mathbf{E}_q \left[\log \frac{q(z \mid x^{(i)})}{p(z)} \right] + \mathbf{E}_q \left[\log \frac{q(z \mid x^{(i)})}{p(z \mid x^{(i)})} \right]$$

$$= \mathbf{E}_q \left[\log p(x^{(i)} \mid z)\right] - D_{KL} \left[q(z \mid x^{(i)}) \| p(z) \right] + D_{KL} \left[\cancel{q(z \mid x^{(i)}) \| p(z \mid x^{(i)})} \right]$$

always non-negative

$$\log p(x)_{\forall i} \geq \mathbf{E}_q \left[\log p(x^{(i)} \mid z)\right] - D_{KL} \left[q(z \mid x^{(i)}) \| p(z) \right]$$ Will Maximize Lower Bound

Can we motivate this in a different way?