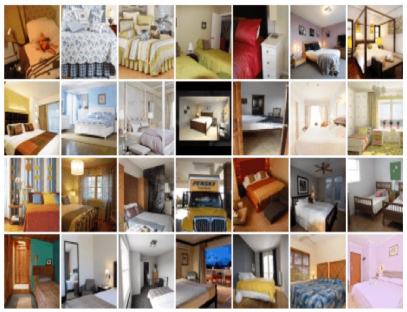
10707
Deep Learning

Russ Salakhutdinov

Machine Learning Department
rsalakhu@cs.cmu.edu

Variational Inference

Statistical Generative Models



+

Model family, loss function, optimization algorithm, etc.

Data

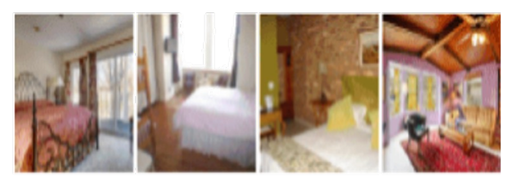
Learning

Prior Knowledge

Image x

→ A probability distribution p(x) → probability p(x)

Sampling from p(x) **generates** new images:



Grover and Ermon, DGM Tutorial

Statistical Generative Models

Training Data(CelebA)

Model Samples (Karras et.al., 2018)

4 years of progression on Faces

Brundage et al., 2017

Conditional Generation

- Conditional generative model P(zebra images| horse images)

- Style Transfer

Input Image

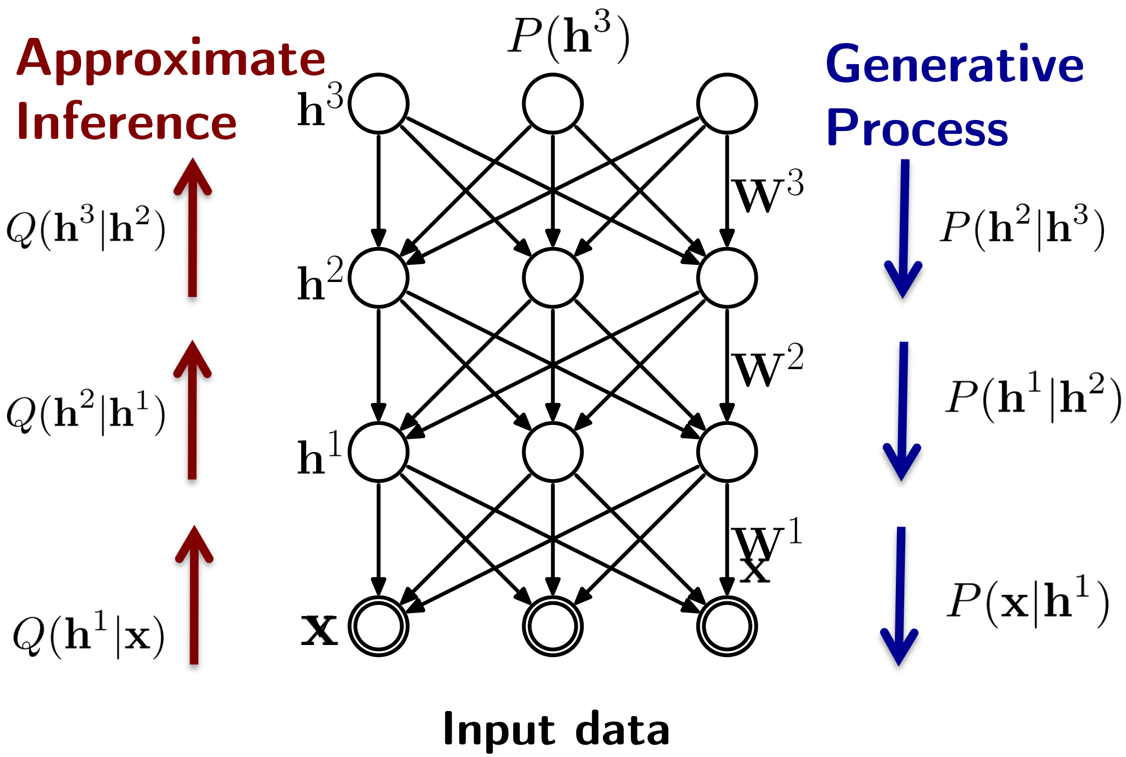
Monet

Van Gogh

Zhou el al., Cycle GAN 2017

Helmholtz Machines

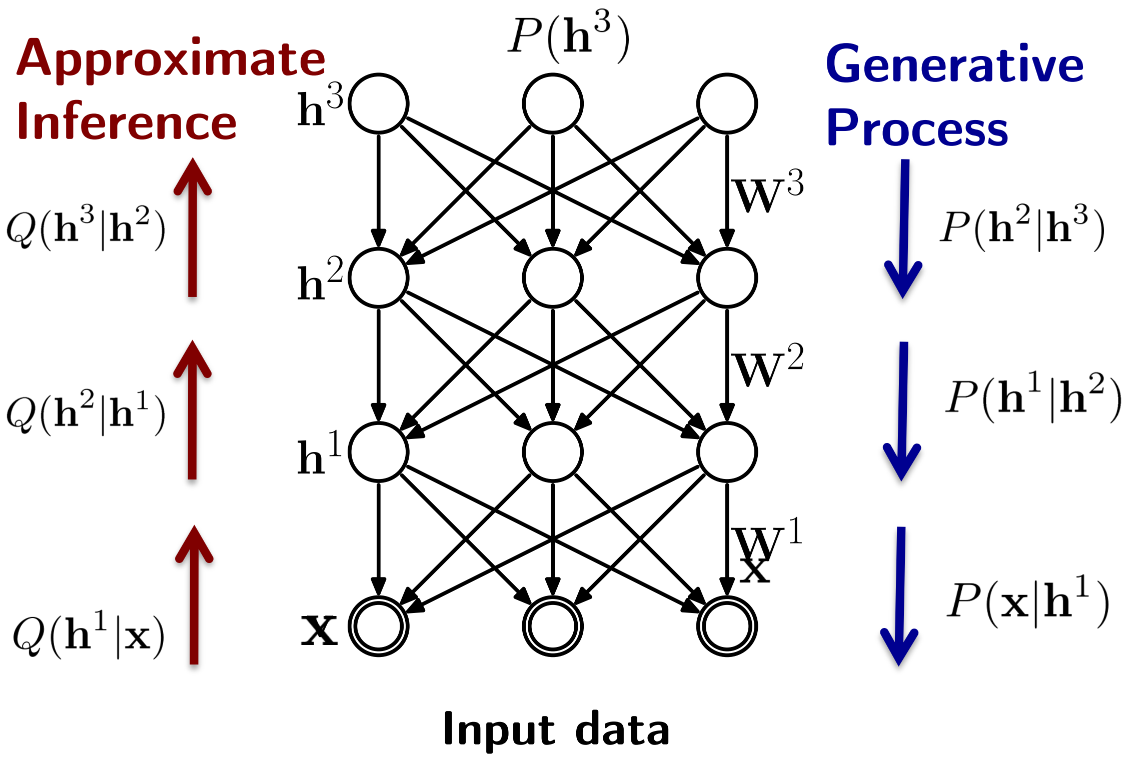
- Hinton, G. E., Dayan, P., Frey, B. J. and Neal, R., Science 1995



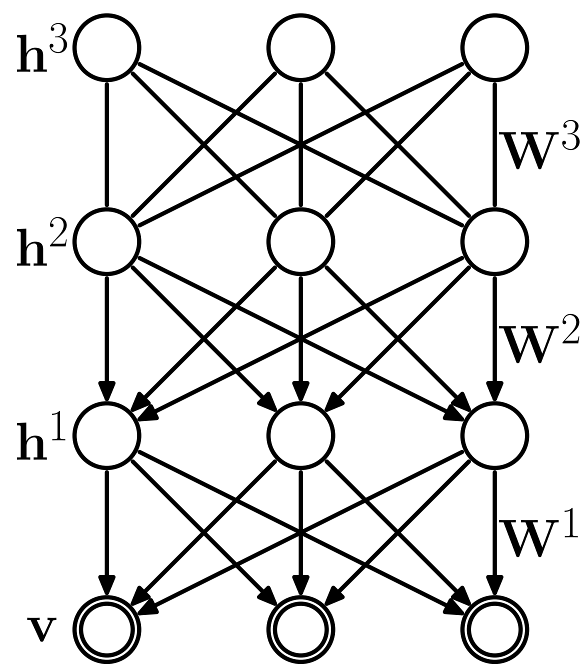
- Kingma & Welling, 2014
- Rezende, Mohamed, Daan, 2014
- Mnih & Gregor, 2014
- Bornschein & Bengio, 2015
- Tang & Salakhutdinov, 2013

Helmholtz Machines

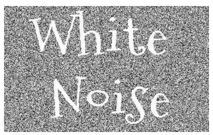
Helmholtz Machine

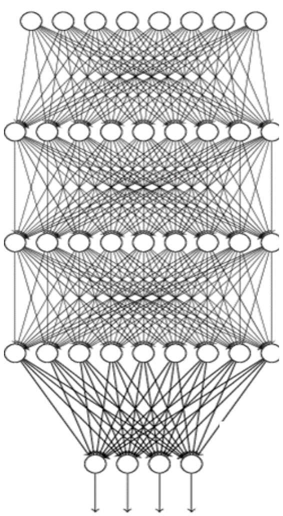


Deep Belief Network

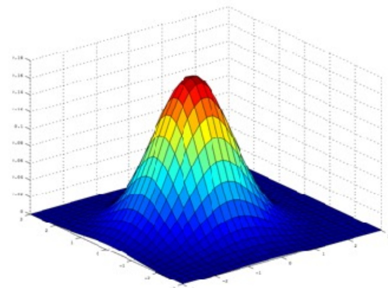


Deep Directed Generative Models



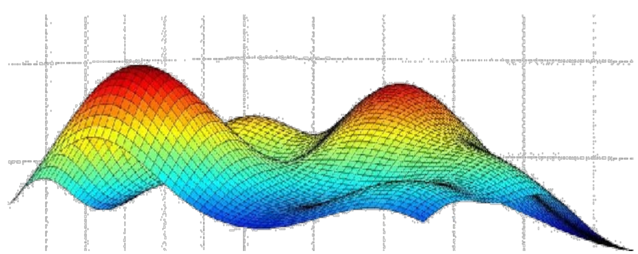


Code Z



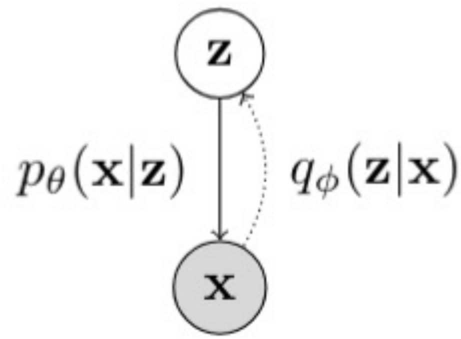
- Recognition
- Bottom-up
- Q(z|x)

- Generative
- Top-Down
- P(x|z)



D_{real}

- Latent Variable Models

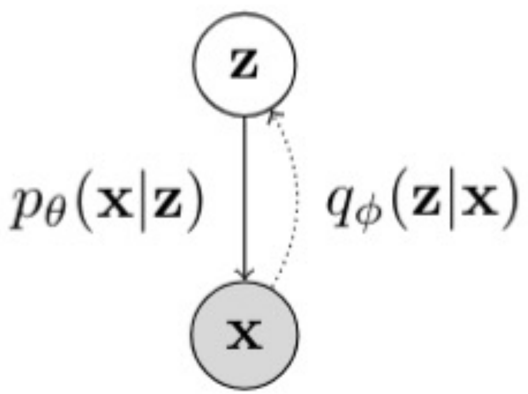


$$\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

- Conditional distributions are parameterized by deep neural networks

Directed Deep Generative Models

- Directed Latent Variable Models with Inference Network



- Maximum log-likelihood objective

$$\max_\theta \sum_{\mathbf{x} \in \mathcal{D}} \log p_\theta(\mathbf{x})$$

- Marginal log-likelihood is intractable:

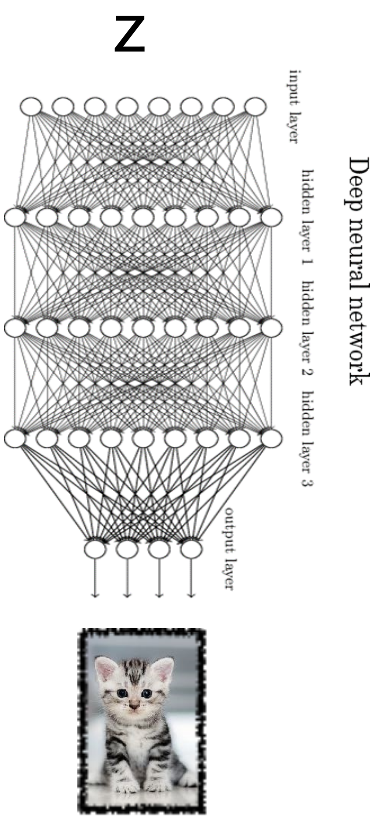
$$\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

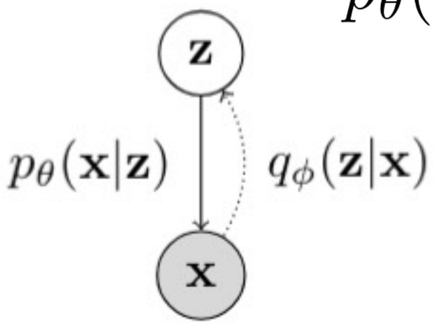
- Key idea: Approximate true posterior p(z|x) with a simple, tractable distribution q(z|x) (inference/recognition network).

Grover and Ermon, DGM Tutorial

Variational Autoencoders (VAEs)

- Single stochastic (Gaussian) layer, followed by many deterministic layers





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

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(\mu(\mathbf{z}, \theta), \Sigma(\mathbf{z}, \theta)\right)$$

Deep neural network parameterized by θ.
(Can use different noise models)

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mu(\mathbf{x}, \phi), \Sigma(\mathbf{x}, \phi)\right)$$

Deep neural network parameterized by ϕ.

Approximate Inference

- When using probabilistic graphical models, we will be interested in evaluating the posterior distribution $p(\mathbf{Z}|\mathbf{X})$ of the latent variables $\mathbf{Z}$ given the observed data $\mathbf{X}$.

- For example, in the EM algorithm, we need to evaluate the expectation of the complete-data log-likelihood with respect to the posterior distribution over the latent variables.

- For more complex models, it may be infeasible to evaluate the posterior distribution, or compute expectations with respect to this distribution.

- This typically occurs when working with high-dimensional latent spaces, or when the posterior distribution has a complex form, for which expectations are not analytically tractable (e.g. Boltzmann machines).

Probabilistic Model

• The model may have latent variables and parameters, and we will denote the set of all latent variables and parameters by **Z**.

• We will also denote the set of all observed variables by **X**.

• For example, we may be given a set of N i.i.d data points, so that $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ (as we saw in our previous class).

• Our probabilistic model specifies the joint distribution $P(\mathbf{X}, \mathbf{Z})$.

• Our goal is to find approximate posterior distribution $P(\mathbf{Z}|\mathbf{X})$ and the model evidence $p(\mathbf{X})$.

Variational Bound

- Given a joint distribution p(**Z**,**X**|θ) over observed and latent variables governed by parameters θ, the goal is to maximize the likelihood function p(**X**|θ) with respect to θ:

$$p(\mathbf{X}|\theta) = \sum_{Z} p(\mathbf{X}, \mathbf{Z}|\theta).$$

- We will assume that **Z** is discrete, although derivations are identical if **Z** contains continuous, or a combination of discrete and continuous variables.
- For any distribution q(**Z**) over latent variables we can derive the following variational lower bound:

$$\ln p(\mathbf{X}|\theta) = \ln \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) = \ln \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})}$$

Jensen's inequality

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} = \mathcal{L}(q, \theta).$$

Variational Bound

- Variational lower-bound:

$$\ln p(\mathbf{X}|\theta) = \ln \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta) = \ln \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})}$$

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})}$$

$$= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X},\mathbf{Z}|\theta) + \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{1}{q(\mathbf{Z})}$$

$$= \mathbb{E}_{q(\mathbf{Z})}\big[\ln p(\mathbf{X},\mathbf{Z}|\theta)\big] + \mathcal{H}(q(\mathbf{Z})) = \mathcal{L}(q,\theta).$$

Expected complete log-likelihood

Entropy functional.

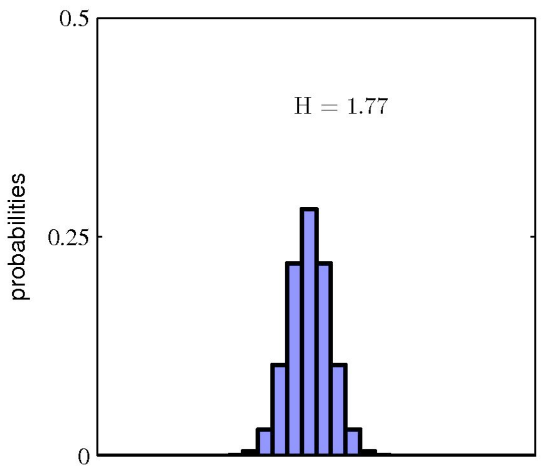
Variational lower-bound

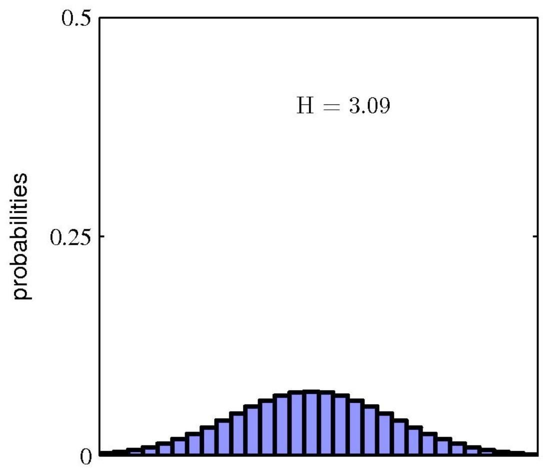
Entropy

- For a discrete random variable X, where $P(X=x_i) = p(x_i)$, the entropy of a random variable is:

$$\mathcal{H}(p) = -\sum_i p(x_i) \log p(x_i).$$

- Distributions that are sharply picked around a few values will have a relatively low entropy, whereas those that are spread more evenly across many values will have higher entropy





- Histograms of two probability distributions over 30 bins.

- The largest entropy will arise from a uniform distribution $H = -\ln(1/30) = 3.40$.

- For a density defined over continuous random variable, the differential entropy is given by:

$$\mathcal{H}(p) = -\int p(x) \log p(x) \mathrm{d}x.$$

Variational Bound

- We saw:

$$\ln p(\mathbf{X}|\theta) \geq \mathbb{E}_{q(\mathbf{Z})}\big[\ln p(\mathbf{X},\mathbf{Z}|\theta)\big] + \mathcal{H}(q(\mathbf{Z})) = \mathcal{L}(q,\theta).$$

- We also note that the following decomposition holds:

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q,\theta) + \mathrm{KL}(q||p),$$

where

$$\mathcal{L}(q,\theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})},$$

Variational lower-bound

$$\mathrm{KL}(q||p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})}.$$

Kullback-Leibler (KL) divergence.

Also known as Relative Entropy.

- KL divergence is not symmetric.
- KL(q||p) $\geq$ 0 with equality iff p(x) = q(x).
- Intuitively, it measures the “distance” between the two distributions.

Variational Bound

- Let us derive that:

$$\log p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \mathrm{KL}(q||p),$$

- We can write:

$$\ln p(\mathbf{X}, \mathbf{Z}|\theta) = \ln p(\mathbf{Z}|\mathbf{X}, \theta) + \ln p(\mathbf{X}|\theta),$$

and plugging into the definition of $\mathcal{L}(q, \theta)$, gives the desired result.

- Note that variational bound becomes tight iff q(**Z**) = p(**Z** | **X**,θ).

- In other words the distribution q(**Z**) is equal to the true posterior distribution over the latent variables, so that KL(q||p) = 0.

- As KL(q||p) $\geq$ 0, it immediately follows that:

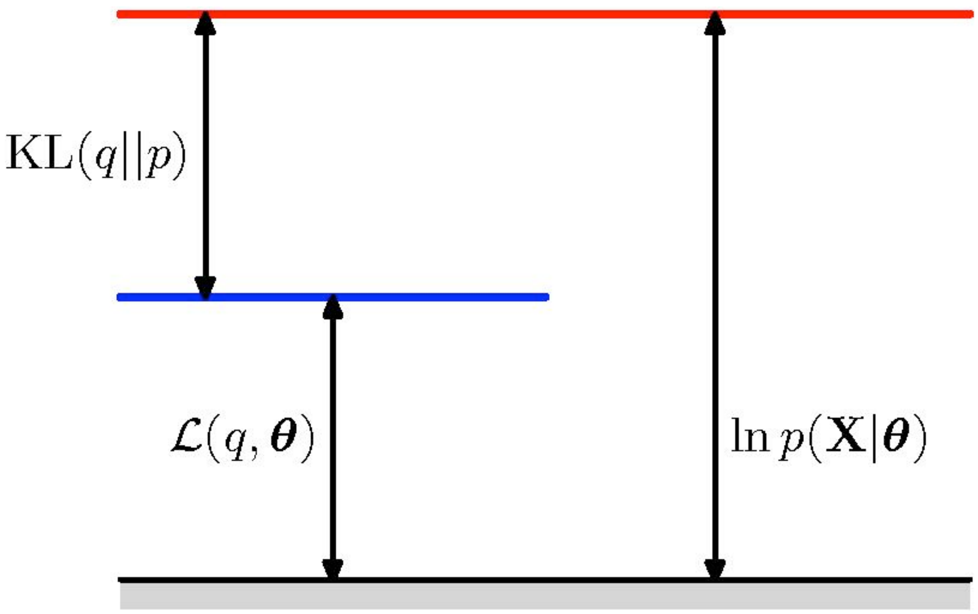
$$\ln p(\mathbf{X}|\theta) \geq \mathcal{L}(q, \theta),$$

which also showed using Jensen's inequality.

Decomposition

- Illustration of the decomposition which holds for any distribution q(**Z**).

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \mathrm{KL}(q||p),$$



Variational Bound

- We can decompose the marginal log-probability as:

$$\log p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p),$$

where

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q||p) = -\int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

- We can maximize the variational lower bound $\mathcal{L}(q)$ with respect to the distribution q(**Z**), which is equivalent to minimizing the KL divergence.

- If we allow any possible choice of q(**Z**), then the maximum of the lower bound occurs when:

$$q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}).$$

In this case KL divergence becomes zero.

Variational Bound

- As in our previous lecture, we can decompose the marginal log-probability as:

$$\log p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p),$$

- We will assume that the true posterior distribution is intractable.

- We can consider a restricted family of distributions q(**Z**) and then find the member of this family for which KL is minimized.

- Our goal is to restrict the family of distributions so that it contains only tractable distributions.

- At the same time, we want to allow the family to be sufficiently rich and flexible, so that it can provide a good approximation to the posterior.

- One option is to use parametric distributions q(**Z**|ω), governed by parameters ω.

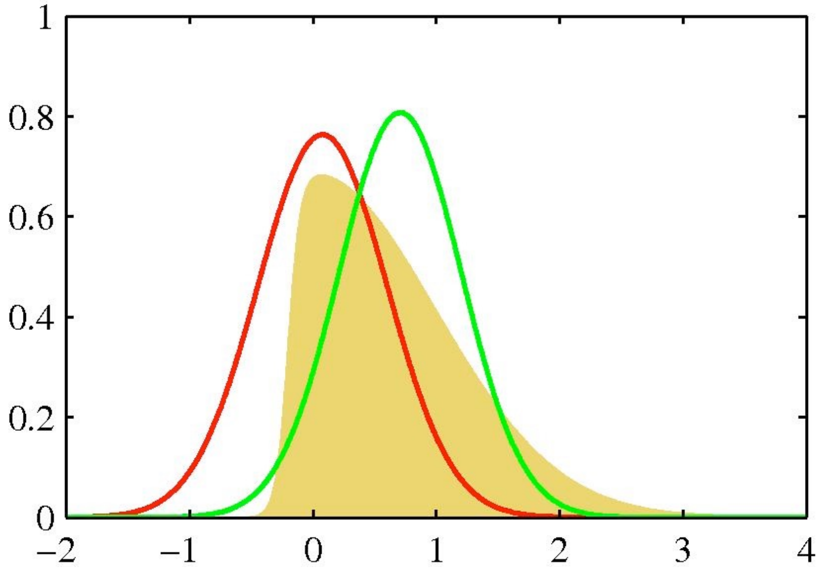
- The lower bound then becomes a function of ω, and we can optimize the lower-bound to determine the optimal values for the parameters.

Example

- One option is to use parametric distributions q(**Z**|ω), governed by parameters ω.

$$\log p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p),$$

- Here is an example, in which the variational distribution is Gaussian. We can optimize with respect to its mean and variance.



The original distribution (yellow), along with Laplace (red), and variational (green) approximations.

Mean-Field

- We now consider restricting the family of distributions.
- Partition the elements of **Z** into M disjoint groups, denoted by $\mathbf{Z_i}$, i=1,…,M.
- We assume that the q distribution factorizes with respect to these groups:

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i).$$

- Note that we place no restrictions on the functional form of the individual factors q_i (we will often denote $q_i(\mathbf{Z_i})$ as simply q_i).

- This approximation framework, developed in physics, is called **mean-field theory**.

Factorized Distributions

- Among all factorized distributions, we look for a distribution for which the variational lower bound is maximized.
- Denoting $q_i(\mathbf{Z}_i)$ as simply q_i, we have:

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i).$$

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} \\
&= \int \prod_i q_i \left[\ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right] \mathrm{d}\mathbf{Z} \\
&= \int q_j \left[\int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right] \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \mathrm{const} \\
&= \int q_j \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \mathrm{const}
\end{aligned}$$

where we denote a new distribution:

$$\tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \mathrm{const}.$$

Factorized Distributions

- Among all factorized distributions, we look for a distribution for which the variational lower bound is maximized.

- Denoting $q_i(\mathbf{Z_i})$ as simply q_i, we have:

$$\begin{aligned}\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} \\ &= \int q_j \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j)\, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \mathrm{const}\end{aligned}$$

where

$$\ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \mathrm{const}.$$

- Here we take an expectation with respect to the q distribution over all variables $\mathbf{Z_i}$ for $i \neq j$, so that:

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i.$$

Maximizing Lower Bound

- Now suppose that we keep $\{q_{i \neq j}\}$ fixed, and optimize the lower bound with respect to all possible forms of the distribution $q_j(\mathbf{Z}_j)$.
- This optimization is easily done by recognizing that:

$$\mathcal{L}(q) = \int q_j \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j)\, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const}$$

$$= -\mathrm{KL}(q_j(\mathbf{Z}_j) || \tilde{p}(\mathbf{X}, \mathbf{Z}_j) + \text{const},$$

constant: does not depend on q.

$$\mathcal{L}(q) = \log p(\mathbf{X}) - \mathrm{KL}(q||p)$$

so the minimum occurs when

$$q_j^*(\mathbf{Z}_j) = \tilde{p}(\mathbf{X}, \mathbf{Z}), \quad \text{or} \quad \ln q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}.$$

- Observe: the log of the optimum solution for factor q_j is given by:
  - Considering the log of the joint distribution over all hidden and visible variables
  - Taking the expectation with respect to all other factors $\{q_i\}$ for $i \neq j$.

Maximizing Lower Bound

• Exponentiating and normalizing, we obtain:

$$q_j^*(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})]\right)\mathrm{d}\mathbf{Z}_j}.$$

• The set of these equations for j=1,…,M represent the set of consistency conditions for the maximum of the lower bound subject to factorization constraint.

• To obtain a solution, we initialize all of the factors and then cycle through factors, replacing each in tern with a revised estimate.

• Convergence is guaranteed because the bound is convex with respect to each of the individual factors.

Factroized Gaussian

• Consider a problem of approximating a general distribution by a factorized distribution.

• To get some insight, let us look at the problem of approximating a Gaussian distribution using a factorized Gaussian distribution.

• Consider a Gaussian distribution over two correlated variables **z** = (z_1,z_2).

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}),$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{12} & \beta_{22}. \end{pmatrix}$$

• Let us approximate this distribution using a factorized Gaussian of the form:

$$q(\mathbf{z}) = q_1(z_1)q_2(z_2).$$

Factroized Gaussian

- Remember:

$$\ln q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const.}$$

- Consider an expression for the optimal factor q_1:

$$\begin{aligned}
\ln q_1^*(z_1) &= \mathbb{E}_{q_2(z_2)}[\ln p(\mathbf{z})] + \text{const} \\
&= \mathbb{E}_{q_2(z_2)}\left[-\frac{\beta_{11}}{2}(z_1 - \mu_1)^2 - \beta_{12}(z_1 - \mu_1)(z_2 - \mu_2) \right] + \text{const} \\
&= -\frac{\beta_{11}}{2} z_1^2 + \beta_{11} z_1 \mu_1 - \beta_{12} z_1 (\mathbb{E}[z_2] - \mu_2) + \text{const.}
\end{aligned}$$

- Note that we have a quadratic function of z_1, and so we can identify $q_1(z_1)$ as a Gaussian distribution:

$$q_1^*(z_1) = \mathcal{N}(z_1 | m_1, \beta_{11}^{-1}), \quad m_1 = \mu_1 - \frac{\beta_{12}}{\beta_{11}}(\mathbb{E}[z_2] - \mu_2).$$

Factroized Gaussian

- By symmetry, we also obtain:

$$q_1^*(z_1) = \mathcal{N}(z_1|m_1, \beta_{11}^{-1}), \quad m_1 = \mu_1 - \frac{\beta_{12}}{\beta_{11}}(\mathbb{E}[z_2] - \mu_2).$$

$$q_2^*(z_2) = \mathcal{N}(z_2|m_2, \beta_{22}^{-1}), \quad m_2 = \mu_2 - \frac{\beta_{12}}{\beta_{22}}(\mathbb{E}[z_1] - \mu_1).$$

- There are two observations to make:
  - We did not assume that $q_i^*(z_i)$ is Gaussian, but rather we derived this result by optimizing variational bound over all possible distributions.
  - The solutions are coupled. The optimal $q_1^*(z_1)$ depends on expectation computed with respect to $q_2^*(z_2)$.

- One option is to cycle through the variables in turn and update them until convergence.

Factroized Gaussian

- By symmetry, we also obtain:

$$q_1^*(z_1) = \mathcal{N}(z_1|m_1, \beta_{11}^{-1}), \quad m_1 = \mu_1 - \frac{\beta_{12}}{\beta_{11}}(\mathbb{E}[z_2] - \mu_2).$$

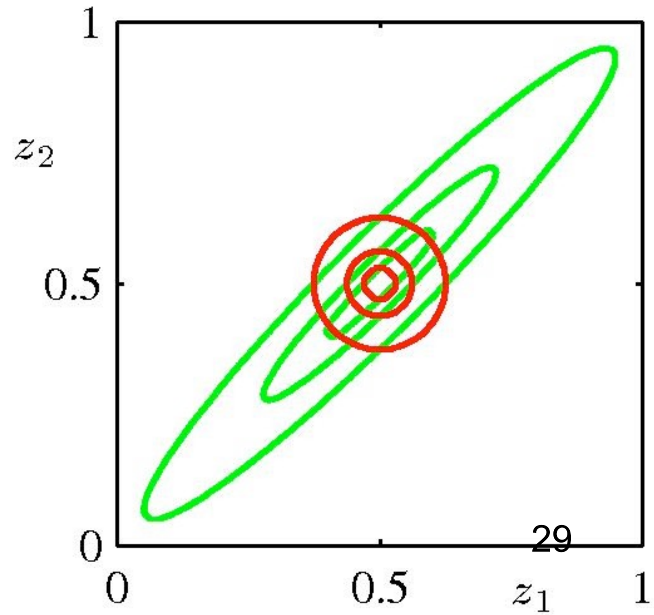
$$q_2^*(z_2) = \mathcal{N}(z_2|m_2, \beta_{22}^{-1}), \quad m_2 = \mu_2 - \frac{\beta_{12}}{\beta_{22}}(\mathbb{E}[z_1] - \mu_1).$$

- However, in our case, $\mathbb{E}[z_1] = \mu_1, \ \mathbb{E}[z_2] = \mu_2.$

- The green contours correspond to 1,2, and 3 standard deviations of the correlated Gaussian.

- The red contours correspond to the factorial approximation q(**z**) over the same two variables.

- Observe that a factorized variational approximation tends to give approximations that are too compact.



Alternative Form of KL Divergence

- We have looked at the variational approximation that minimizes KL(q||p).
- For comparison, suppose that we were minimizing KL(p||q).

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{Z}) \ln \frac{q(\mathbf{Z})}{p(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{Z}) \left[\sum_{i=1}^{M} \ln q_i(\mathbf{Z}_i) \right] \mathrm{d}\mathbf{Z} + \int p(\mathbf{Z}) \ln \frac{1}{p(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

constant: does not depend on q.

- It is easy to show that:

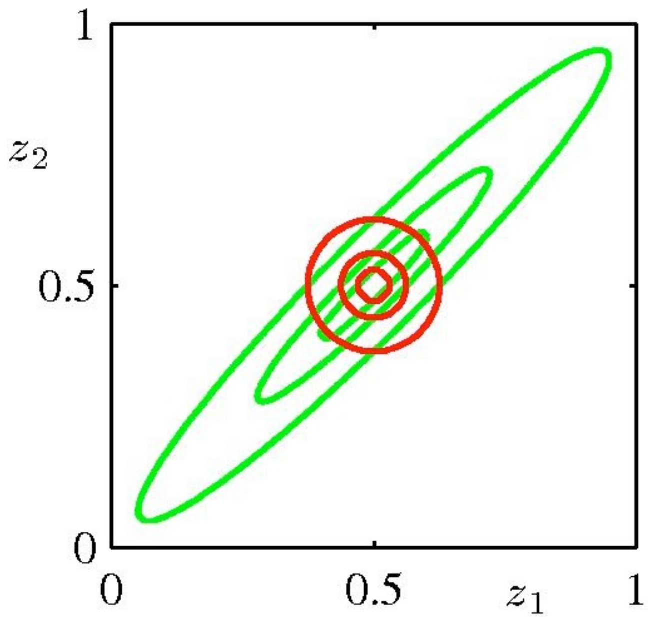
$$q_j^*(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} d\mathbf{Z}_i = p(\mathbf{Z}_j).$$

- The optimal factor is given by the marginal distribution of p(**Z**).

Comparison of two KLs

- Comparison of two the alternative forms for the KL divergence.

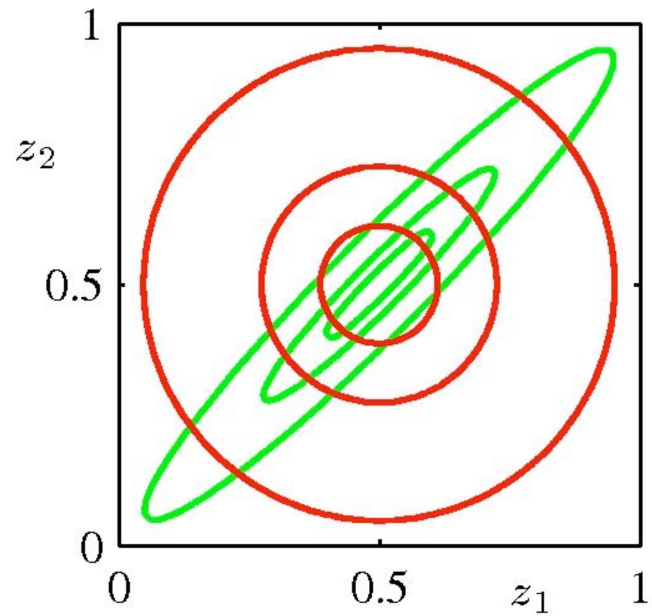
KL(q||p)



(a)

Approximation is too compact.

KL(p||q)



(b)

Approximation is too spread.

Comparison of two KLs

• The difference between these two approximations can be understood as follows:

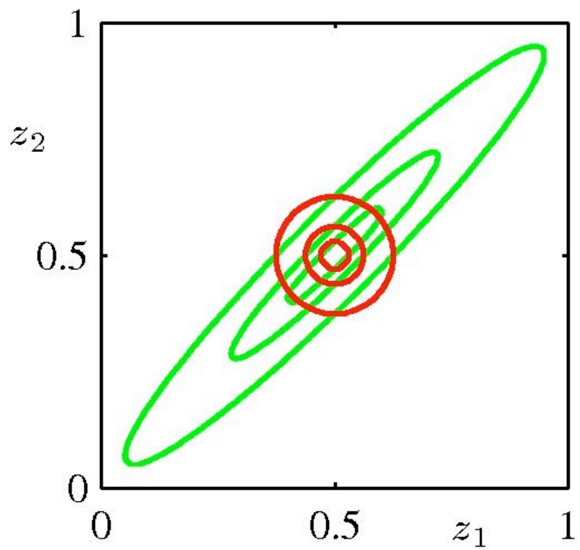
$$\mathrm{KL}(q||p) = -\int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

• There is a large positive contribution to the KL divergence from regions of **Z** space in which:

- p(**Z**) is near zero,
- unless q(**Z**) is also close to zero.

• Minimizing KL(q||p) leads to distributions q(**Z**) that avoid regions in which p(**Z**) is small.

KL(q||p)



(a)

Comparison of two KLs

- Similar arguments apply for the alternative KL divergence:

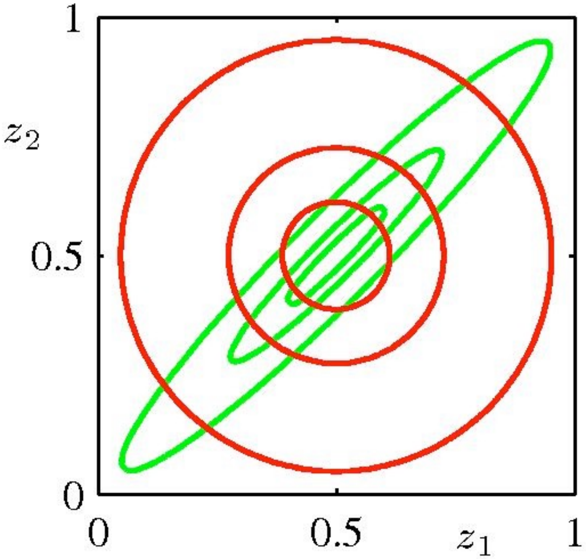
$$\mathrm{KL}(p||q) = -\int p(\mathbf{Z}) \ln \frac{q(\mathbf{Z})}{p(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

- There is a large positive contribution to the KL divergence from regions of **Z** space in which:

  - q(**Z**) is near zero,
  - unless p(**Z**) is also close to zero.

- Minimizing KL(p||q) leads to distributions q(**Z**) that are nonzero in regions where p(**Z**) is nonzero.

KL(p||q)

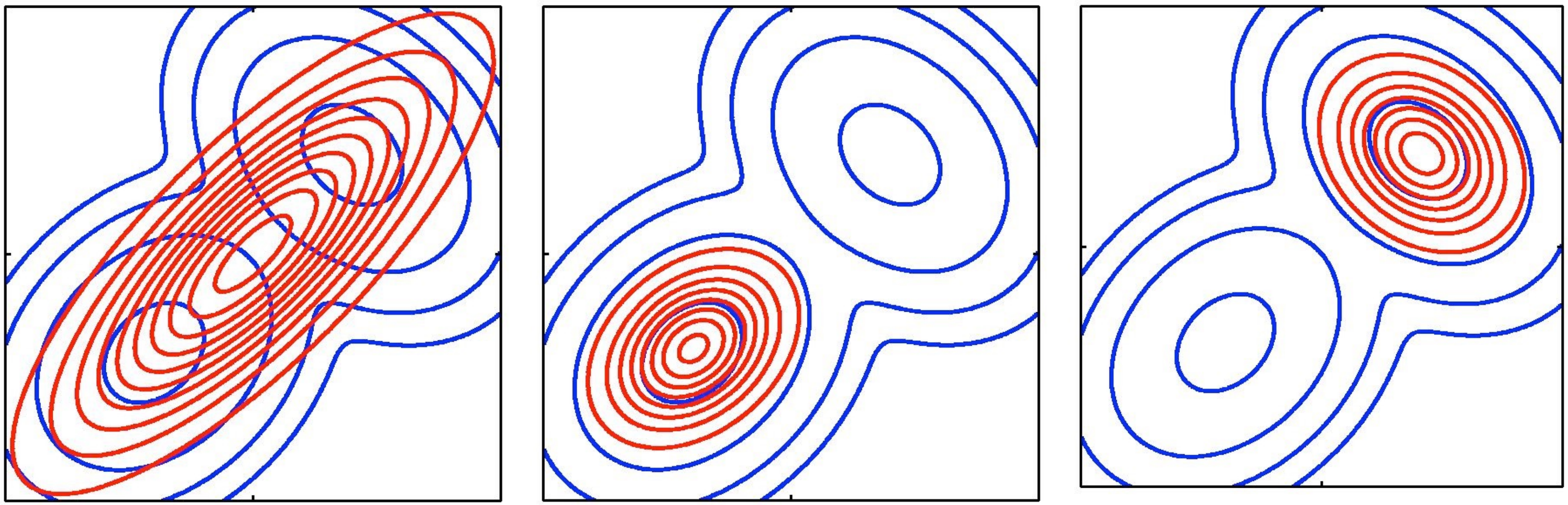


(b)

Approximating Multimodal Distribution

• Consider approximating multimodal distribution with a unimodal one.

• Blue contours show bimodal distribution p(**Z**), red contours show a single Gaussian distribution that best approximates q(**Z**) that best approximates p(**Z**).



KL(p||q) KL(q||p) KL(q||p)

• In practice, the true posterior will often be mutlimodal.

• KL(q||p) will tend to find a single mode, whereas KL(p||q) will average across all of the modes.

Alpha-family of Divergences

- The two forms of KL are members of the alpha-family divergences:

$$D_\alpha(p||q) = \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{(1+\alpha)/2} q(x)^{(1-\alpha)/2} \mathrm{d}x\right), \quad -\infty < \alpha < \infty.$$

- Observe three points:
  - KL(p||q) corresponds to the limit $\alpha \to 1$.
  - KL(q||p) corresponds to the limit $\alpha \to -1$.
  - $D_\alpha(p||q) \geq 0$, for all α, and $D_\alpha(p||q)=0$ iff q(x) = p(x).

- Suppose p(x) is fixed and we minimize $D_\alpha(p||q)$ with respect to q distribution.
- For α < -1, the divergence is zero-forcing: q(x) will underestimate the support of p(x).
- For α > 1, the divergence is zero-avoiding: q(x) will stretch to cover all of p(x).
- For α = 0, we obtain a symmetric divergence which is related to Hellinger Distance:

$$D_H(p||q) = \frac{1}{2}\int \left(p(x)^{1/2} - q(x)^{1/2}\right)^2 \mathrm{d}x.$$