Variational Inference

Rafael Rafailov

October 9, 2021

Kullback–Leibler divergence

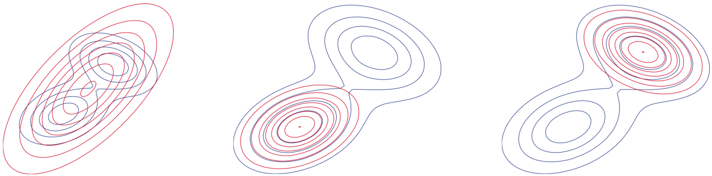
Definition

Consider two distributions p and q over a set $\mathcal{X}$. The KL-divergence $\mathbb{D}_{KL}[p||q]$ is defined as

$$\begin{aligned}\mathbb{D}_{KL}[p||q] &= -\int_{\mathcal{X}} p(x) \log \frac{q(x)}{p(x)} dx \\ &= \mathbb{E}_{x\sim p(x)}\left[\log \frac{p(x)}{q(x)}\right]\end{aligned}$$

Properties of the KL divergence

1. The KL divergence is not symmetric $\mathbb{D}_{KL}[p||q] \neq \mathbb{D}_{KL}[q||p]$



(a) (b) (c)

Figure: Fitting a unimodal approximating distribution q (red) to a multimodal p (blue). Using KL(p||q) leads to **mode-covering** (a). However, using KL(q||p) forces q to be **mode-seeking** (b, c)

[1] Image credit: CS 236

Properties of the KL divergence

1. The KL-divergence between two distributions (when it is defined) is always non-negative.
   Proof: We have

$$
\begin{aligned}
\mathbb{D}_{KL}[q||p] &= \mathbb{E}_{x\sim p(x)}\left[-\log\frac{q(x)}{p(x)}\right] \geq -\log\mathbb{E}_{x\sim p(x)}\left[\frac{q(x)}{p(x)}\right] \\
&= -\int_{\mathcal{X}} p(x)\frac{q(x)}{p(x)}dx = \\
&-\log\int_{\mathcal{X}} q(x)dx = -\log 1 = 0
\end{aligned}
$$

   where the above follows from Jensen's inequality.
2. From the previous point $\mathbb{D}_{KL}[q||p]$ if and only if $p = q$ (module some measure-theoretic considerations).

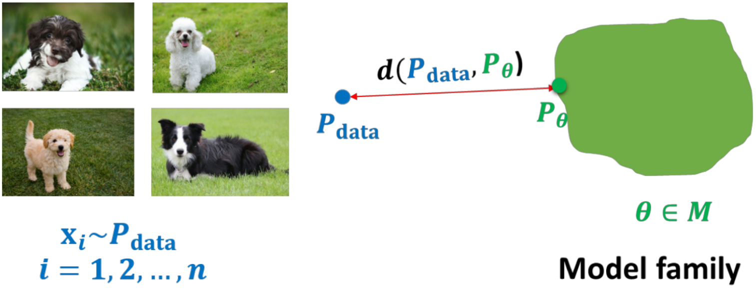


Figure: We want to fit the probability distribution of the data

[1]Image credit: CS 236

Generative models over data

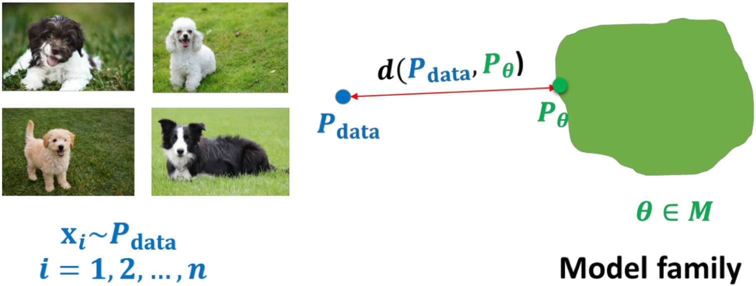


Figure: We want to fit the probability distribution of the data

$$\arg\min_{p_\theta} \mathbb{D}_{KL}(p_{\text{data}} || p_\theta) = \max_{p_\theta} \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \log p_\theta(x_i) \tag{0.1}$$

[1] Image credit: CS 236

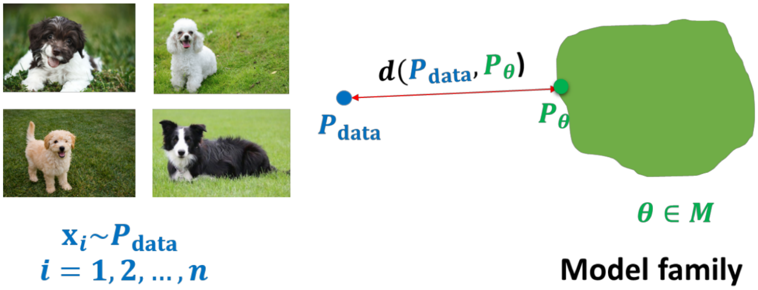


Figure: We want to fit the probability distribution of the data

$$\arg\min_{p_\theta} \mathbb{D}_{KL}(p_{\text{data}} || p_\theta) = \max_{p_\theta} \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \log p_\theta(x_i) \qquad (0.1)$$

We just need to train a maximum likelihood model!

[1] Image credit: CS 236

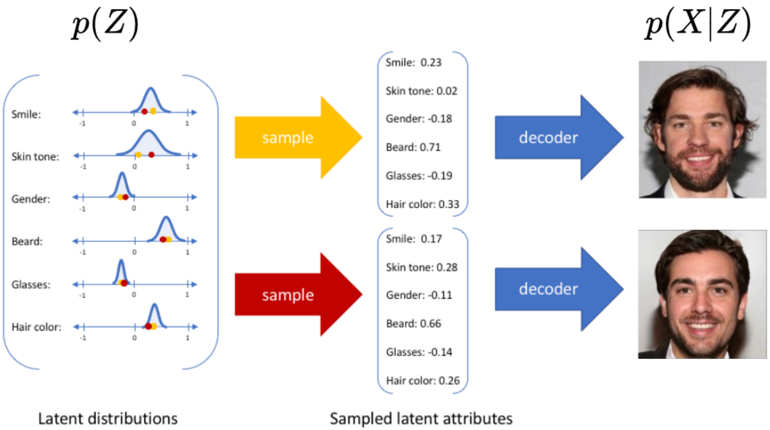
Generative models over data

Evaluating likelihoods over high dimensions is hard!

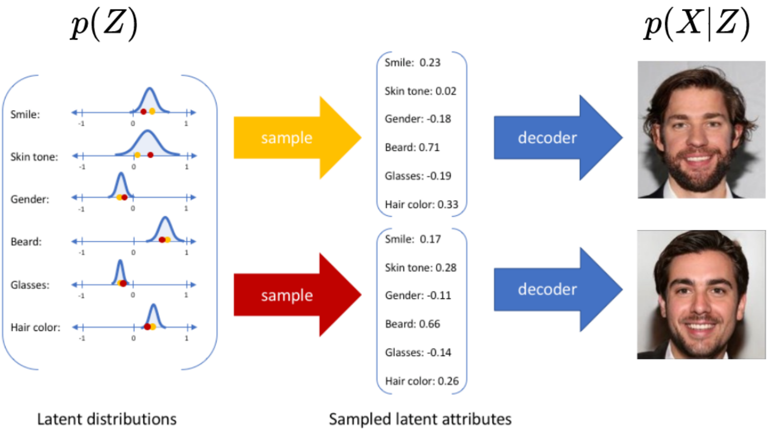
Latent variable models

[1] Image credit: Jeremy Jordan's blog post.

Latent variable models



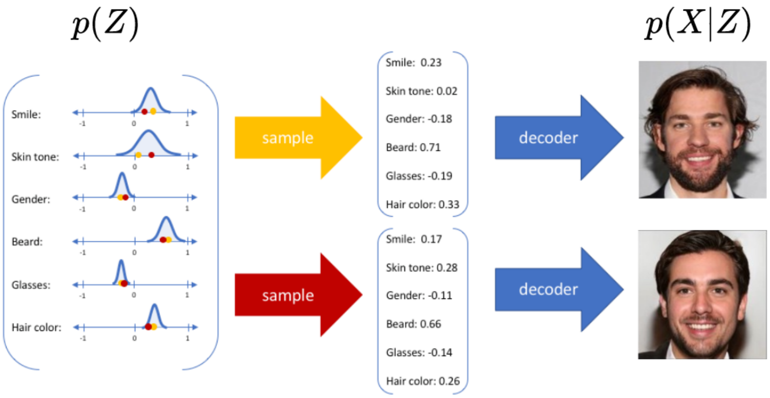
[1] Image credit: Jeremy Jordan's blog post.



1. Data is governed by a simple **latent** distribution $p(Z)$.

[1] Image credit: Jeremy Jordan's blog post.

Latent variable models



Latent distributions

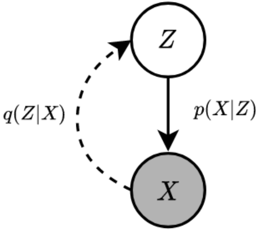
Sampled latent attributes

1. Data is governed by a simple **latent** distribution $p(Z)$.
2. The **observed** data X is generated by a conditional distribution $p(X|Z)$.

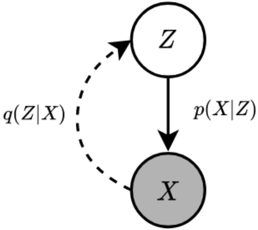
[1] Image credit: Jeremy Jordan's blog post.

Data likelihood under latent variable models

General solution: Introduce inference distribution $q(z|x)$.



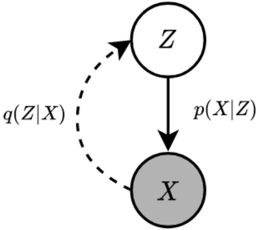
General solution: Introduce inference distribution $q(z|x)$.
Intuition:



Data likelihood under latent variable models

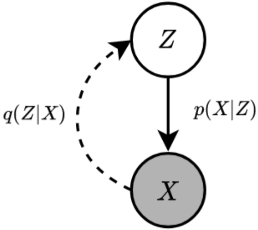
General solution: Introduce inference distribution $q(z|x)$.
Intuition:

1. Guess the likely z given x_i and use those to compute likelihood.



General solution: Introduce inference distribution $q(z|x)$.
Intuition:

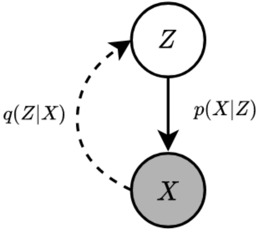
1. Guess the likely z given x_i and use those to compute likelihood.
2. Evaluate uncertainty through a distribution over z - $q(z|x)$.



Data likelihood under latent variable models

General solution: Introduce inference distribution $q(z|x)$.
Intuition:

1. Guess the likely z given x_i and use those to compute likelihood.
2. Evaluate uncertainty through a distribution over z - $q(z|x)$.
3. Approach is similar to the EM algorithm.



Variational Inference

$$\log p_\theta(x) =$$

$$\log p_\theta(x) = \log \int p_\theta(x|z)p(z)dz =$$

$$\log p_\theta(x) = \log \int p_\theta(x|z)p(z)dz =$$

$$\log \int \underbrace{\frac{q_\phi(z|x)}{q_\phi(z|x)}}_{=1} p_\theta(x|z)p(z)dz =$$

Variational Inference

$$\log p_\theta(x) = \log \int p_\theta(x|z)p(z)dz =$$

$$\log \int \underbrace{\frac{q_\phi(z|x)}{q_\phi(z|x)}}_{=1} p_\theta(x|z)p(z)dz = \log \int q_\phi(z|x)\frac{p_\theta(x,z)}{q_\phi(z|x)}dz =$$

$$\log p_\theta(x) = \log \int p_\theta(x|z)p(z)dz =$$

$$\log \int \underbrace{\frac{q_\phi(z|x)}{q_\phi(z|x)}}_{=1} p_\theta(x|z)p(z)dz = \log \int q_\phi(z|x)\frac{p_\theta(x,z)}{q_\phi(z|x)}dz =$$

$$\log \mathbb{E}_{z\sim q_\phi(z|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right]$$

$$\log p_\theta(x) = \log \int p_\theta(x|z)p(z)dz =$$

$$\log \int \underbrace{\frac{q_\phi(z|x)}{q_\phi(z|x)}}_{=1} p_\theta(x|z)p(z)dz = \log \int q_\phi(z|x)\frac{p_\theta(x,z)}{q_\phi(z|x)}dz =$$

$$\log \mathbb{E}_{z\sim q_\phi(z|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \underbrace{\geq}_{\text{Jensen}} \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] =$$

Variational Inference

$$\log p_\theta(x) = \log \int p_\theta(x|z)p(z)dz =$$

$$\log \int \underbrace{\frac{q_\phi(z|x)}{q_\phi(z|x)}}_{=1} p_\theta(x|z)p(z)dz = \log \int q_\phi(z|x) \frac{p_\theta(x,z)}{q_\phi(z|x)} dz =$$

$$\log \mathbb{E}_{z\sim q_\phi(z|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \underbrace{\geq}_{\text{Jensen}} \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] =$$

$$\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] + \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p(z) - \log q_\phi(z|x)]}_{=-\mathbb{D}_{KL}(q_\phi(z|x)||p(z))} =: \mathcal{L}_{\theta,\phi}(x)$$

Why exactly do we sample from $q_\phi(z|x)$?

Variational Inference

Why exactly do we sample from $q_\phi(z|x)$?

Proposition

For any distribution $q(z)$ we have:

$$\log p_\theta(x) - \mathbb{D}_{KL}(q(z)||p(z|x)) =$$
$$\mathbb{E}_{z\sim q(z)}[\log p_\theta(x|z)] - \mathbb{D}_{KL}(q(z)||p(z))$$

Variational Inference

Why exactly do we sample from $q_\phi(z|x)$?

Proposition

For any distribution $q(z)$ we have:

$$\log p_\theta(x) - \mathbb{D}_{KL}(q(z)||p(z|x)) = \\ \mathbb{E}_{z\sim q(z)}[\log p_\theta(x|z)] - \mathbb{D}_{KL}(q(z)||p(z))$$

1. The Evidence Lower Bound (ELBO) is a lower bound on the data log-likelihood under any sampling distribution $q(z)$.

Variational Inference

Why exactly do we sample from $q_\phi(z|x)$?

Proposition

For any distribution $q(z)$ we have:

$$\log p_\theta(x) - \mathbb{D}_{KL}(q(z)||p(z|x)) = \mathbb{E}_{z\sim q(z)}[\log p_\theta(x|z)] - \mathbb{D}_{KL}(q(z)||p(z))$$

1. The Evidence Lower Bound (ELBO) is a lower bound on the data log-likelihood under any sampling distribution $q(z)$.
2. From the properties of KL-divergences, equality is achieved only when $q(z) = p(z|x)$.

Why exactly do we sample from $q_\phi(z|x)$?

Proposition

For any distribution $q(z)$ we have:

$$\log p_\theta(x) - \mathbb{D}_{KL}(q(z)||p(z|x)) =$$
$$\mathbb{E}_{z\sim q(z)}[\log p_\theta(x|z)] - \mathbb{D}_{KL}(q(z)||p(z))$$

1. The Evidence Lower Bound (ELBO) is a lower bound on the data log-likelihood under any sampling distribution $q(z)$.
2. From the properties of KL-divergences, equality is achieved only when $q(z) = p(z|x)$.
3. To minimize the ELBO gap we choose $q(z) = q(z|x)$.

Optimizing the ELBO

The ELBO consists of two terms: reconstruction and a KL regularization:

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

Optimizing the ELBO

The ELBO consists of two terms: reconstruction and a KL regularization:

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

- Set $q_\phi(z|x)$ to be multivariate normal distribution parameterized by a neural network, i.e $q_\phi(z|x) = \mathcal{N}(z; \mu_\theta(x), \sigma_\theta(x))$, where $\sigma_\theta(x) = \text{diag}(\sigma_1^2(x), \ldots, \sigma_K^2(x))$ is a diagonal matrix.

Optimizing the ELBO

The ELBO consists of two terms: reconstruction and a KL regularization:

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

- Set $q_\phi(z|x)$ to be multivariate normal distribution parameterized by a neural network, i.e $q_\phi(z|x) = \mathcal{N}(z; \mu_\theta(x), \sigma_\theta(x))$, where $\sigma_\theta(x) = \text{diag}(\sigma_1^2(x), \ldots, \sigma_K^2(x))$ is a diagonal matrix.

Proposition

Let $q = \mathcal{N}(\mu_1, \Sigma_1)$ *and* $p = \mathcal{N}(\mu_2, \Sigma_2)$, *then*

$$\mathbb{D}_{KL}(q||p) = \frac{1}{2}\Big[tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)\Sigma_2^{-1}(\mu_2 - \mu_1) - k + \log\left(\frac{\det \Sigma_2}{\det \Sigma_1}\right)\Big]$$

Optimizing the ELBO

The ELBO consists of two terms: reconstruction and a KL regularization:

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

- Set $q_\phi(z|x)$ to be multivariate normal distribution parameterized by a neural network, i.e $q_\phi(z|x) = \mathcal{N}(z; \mu_\theta(x), \sigma_\theta(x))$, where $\sigma_\theta(x) = \text{diag}(\sigma_1^2(x), \ldots, \sigma_K^2(x))$ is a diagonal matrix.

Proposition

Let $q = \mathcal{N}(\mu_1, \Sigma_1)$ *and* $p = \mathcal{N}(\mu_2, \Sigma_2)$, *then*

$$\mathbb{D}_{KL}(q||p) = \frac{1}{2}\Big[tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)\Sigma_2^{-1}(\mu_2 - \mu_1) - k + \log\left(\frac{\det \Sigma_2}{\det \Sigma_1}\right)\Big]$$

This is a differentiable function!

Optimizing the ELBO

The ELBO consists of two terms: reconstruction and a KL regularization:

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

Optimizing the ELBO

The ELBO consists of two terms: reconstruction and a KL regularization:

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

Optimizing the ELBO

The ELBO consists of two terms: reconstruction and a KL regularization:

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

Need to compute

$$\nabla_{\phi,\theta}\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]$$

Optimizing the ELBO

The ELBO consists of two terms: reconstruction and a KL regularization:

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

Need to compute

$$\nabla_{\phi,\theta}\mathbb{E}_{z\sim q_{\color{red}\phi}(z|x)}[\log p_{\color{red}\theta}(x|z)]$$

Both the expectation and the likelihood are functions of model parameters!

Optimizing the ELBO

The ELBO consists of two terms: reconstruction and a KL regularization:

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

Need to compute

$$\nabla_{\phi,\theta}\mathbb{E}_{z\sim q_{\color{red}\phi}(z|x)}[\log p_{\color{red}\theta}(x|z)] \approx \nabla_{\phi,\theta}\hat{\mathbb{E}}_{z\sim q_{\color{red}\phi}(z|x)}[\log p_{\color{red}\theta}(x|z)]$$

Optimizing the ELBO

The ELBO consists of two terms: reconstruction and a KL regularization:

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

Need to compute

$$\nabla_{\phi,\theta}\mathbb{E}_{z\sim q_{\phi}(z|x)}[\log p_{\theta}(x|z)] \approx \nabla_{\phi,\theta}\hat{\mathbb{E}}_{z\sim q_{\phi}(z|x)}[\log p_{\theta}(x|z)]$$

Reparameterization Trick:

$$\tilde{z} = \mu_\phi(x) + \epsilon\sigma_\phi(x), \text{ where } \epsilon \sim \mathcal{N}(0, I)$$

Optimizing the ELBO

The ELBO consists of two terms: reconstruction and a KL regularization:

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

Need to compute

$$\nabla_{\phi,\theta}\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] \approx \nabla_{\phi,\theta}\hat{\mathbb{E}}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]$$

Reparameterization Trick:

$$\tilde{z} = \mu_\phi(x) + \epsilon\sigma_\phi(x), \text{ where } \epsilon \sim \mathcal{N}(0, I)$$

$$\nabla_{\phi,\theta}\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] \approx \nabla_{\phi,\theta}\frac{1}{M}\sum_{j=1}^{M}\log p_\theta(x|\underbrace{\mu_\phi(x) + \epsilon^{(j)}\sigma_\phi(x)}_{\tilde{z}^{(j)}\sim\mathcal{N}(\mu_\phi(x),\sigma_\phi(x))})$$

The ELBO consists of two terms: reconstruction and a KL regularization:

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

Need to compute

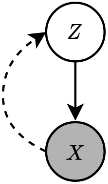
$$\nabla_{\phi,\theta}\mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(x|z)] \approx \nabla_{\phi,\theta}\hat{\mathbb{E}}_{z \sim q_\phi(z|x)}[\log p_\theta(x|z)]$$

Reparameterization Trick:

$$\tilde{z} = \mu_\phi(x) + \epsilon\sigma_\phi(x), \text{ where } \epsilon \sim \mathcal{N}(0, I)$$

$$\nabla_{\phi,\theta}\mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(x|z)] \approx \nabla_{\phi,\theta}\frac{1}{M}\sum_{j=1}^{M} \log p_\theta(x|\underbrace{\mu_\phi(x) + \epsilon^{(j)}\sigma_\phi(x)}_{\tilde{z}^{(j)} \sim \mathcal{N}(\mu_\phi(x), \sigma_\phi(x))})$$

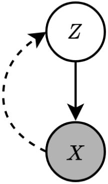
In practice usually $M = 1$.



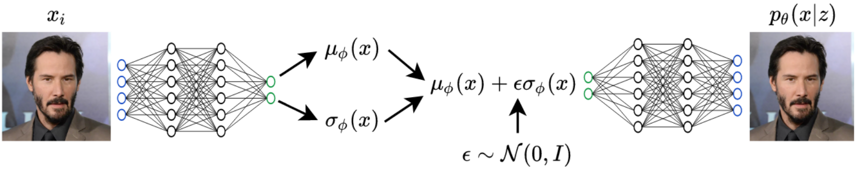
$$q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \sigma_\phi(x))$$

$$p_\theta(x|z) = \mathcal{N}(\mu_\theta(z), \sigma_\theta(z))$$

$q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \sigma_\phi(x))$

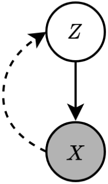


$p_\theta(x|z) = \mathcal{N}(\mu_\theta(z), \sigma_\theta(z))$

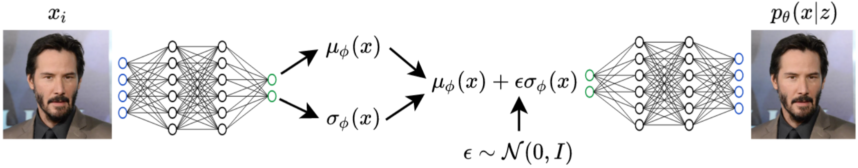


Variational Auto Encoder (VAE)

$q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \sigma_\phi(x))$



$p_\theta(x|z) = \mathcal{N}(\mu_\theta(z), \sigma_\theta(z))$



$$\max_{\phi,\theta} \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(x_i|\mu_\phi(x) + \epsilon\sigma_\phi(x)) - \mathbb{D}_{KL}[\mathcal{N}(\mu_\phi(x_i), \sigma_\phi(x_i))||\mathcal{N}(0, I)]$$

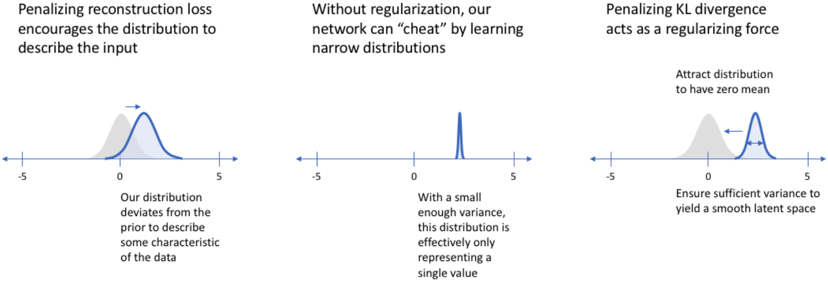
What does the VAE actually do

$$\max_{\phi,\theta} = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

[1] Image credit: Jeremy Jordan's blog post.

What does the VAE actually do

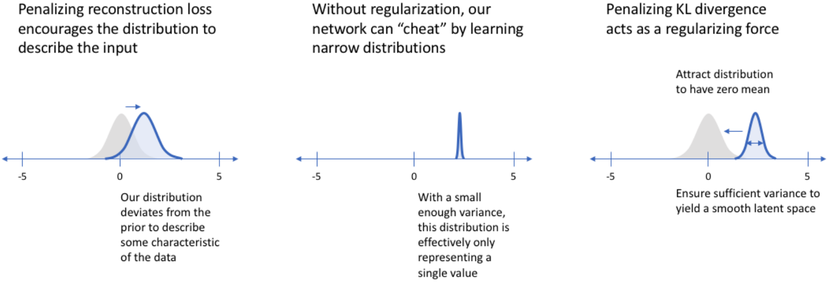
$$\max_{\phi,\theta} = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$



[1]Image credit: Jeremy Jordan's blog post.

What does the VAE actually do

$$\max_{\phi,\theta} = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}(q_\phi(z|x)||p(z))}_{\text{KL regularization}}$$

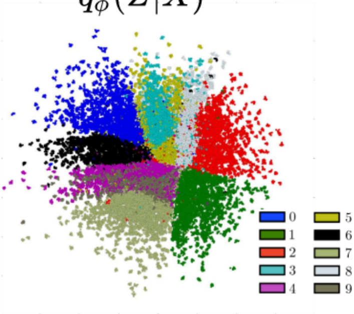


The VAE objectives arranges data on a compact manifold (we can sample from) in a continuous smooth way.

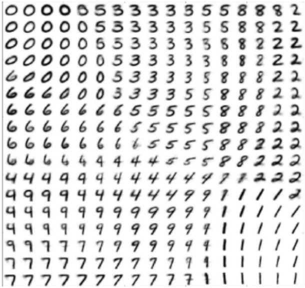
[1] Image credit: Jeremy Jordan's blog post.

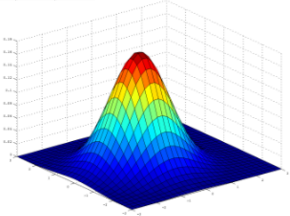
Example: MNIST VAE

$q_\phi(Z|X)$



$p_\theta(X|Z)$

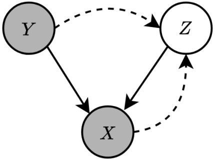




$p(Z)$

Conditional VAEs

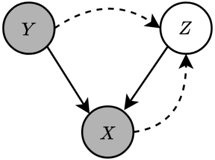
$$q_\phi(z|x,y) = \mathcal{N}(\mu_\phi(x,y), \sigma_\phi(x,y))$$



$$p_\theta(x|z,y) = \mathcal{N}(\mu_\theta(z,y), \sigma_\theta(z,y))$$

Conditional VAEs

$$q_\phi(z|x,y) = \mathcal{N}(\mu_\phi(x,y), \sigma_\phi(x,y))$$

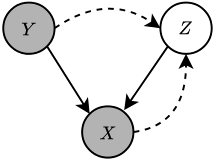


$$p_\theta(x|z,y) = \mathcal{N}(\mu_\theta(z,y), \sigma_\theta(z,y))$$

Figure: Samples from CVAE trained on SVHN

Conditional VAEs

$q_\phi(z|x,y) = \mathcal{N}(\mu_\phi(x,y), \sigma_\phi(x,y))$



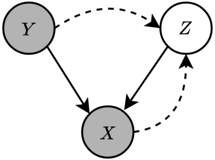
$p_\theta(x|z,y) = \mathcal{N}(\mu_\theta(z,y), \sigma_\theta(z,y))$

Figure: Samples from CVAE trained on SVHN

$$\max_{\theta,\phi} \mathbb{E}_{z\sim q_\phi(z|x,y)}[\log p_\theta(x|z,y)] - \mathbb{D}_{KL}(q_\phi(z|x,y)||p_\theta(z|y))$$

Conditional VAEs

$$q_\phi(z|x,y) = \mathcal{N}(\mu_\phi(x,y), \sigma_\phi(x,y))$$



$$p_\theta(x|z,y) = \mathcal{N}(\mu_\theta(z,y), \sigma_\theta(z,y))$$

Figure: Samples from CVAE trained on SVHN

$$\max_{\theta,\phi} \mathbb{E}_{z\sim q_\phi(z|x,y)}[\log p_\theta(x|z,y)] - \mathbb{D}_{KL}(q_\phi(z|x,y)||p_\theta(z|y))$$

We have (optional) additional conditional-specific prior $p_\theta(z|y)$.

Final Questions

Questions?