Variational Inference

Rafael Rafailov

October 9, 2021

Kullback–Leibler divergence

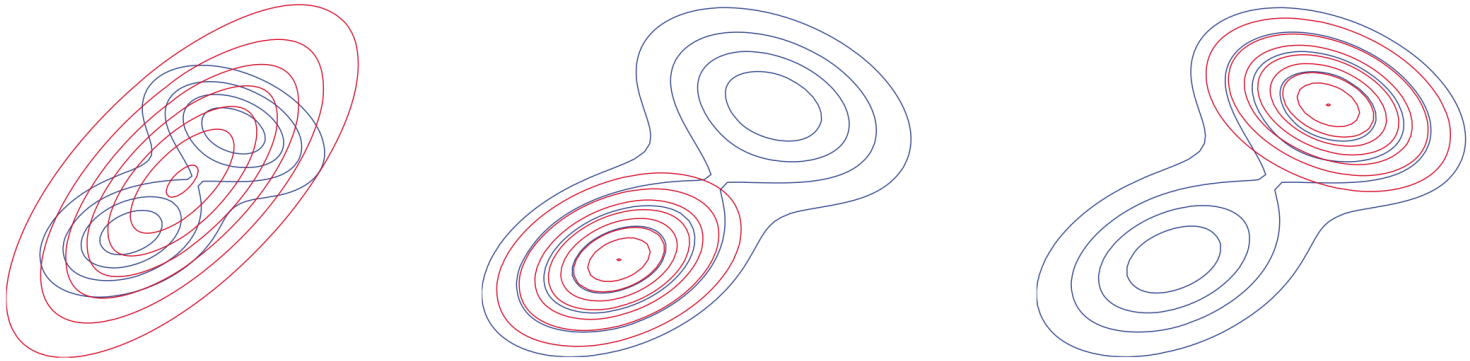
Definition

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

$$D_{KL}[p||q] = \int_X p(x) \log \frac{q(x)}{p(x)} dx$$
$$= E_{x \sim p(x)} \left[\log \frac{p(x)}{q(x)} \right]$$

Properties of the KL divergence

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



(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

$$D_{KL}[q||p] = E_{x \sim p(x)}\left[\log \frac{q(x)}{p(x)}\right] \qquad \log E_{x \sim p(x)}\left[\frac{q(x)}{p(x)}\right]$$

$$= \int_x p(x)\frac{q(x)}{p(x)}dx =$$

$$\log \int_x q(x)dx = \log 1 = 0$$

   where the above follows from Jensen's inequality.

2. From the previous point $D_{KL}[q||p]$ if and only if $p = q$ (module some measure-theoretic considerations).

Generative models over data

$$\mathbf{x}_i \sim P_{\text{data}}$$
$$i = 1, 2, \ldots, n$$

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

[1] Image credit: CS 236

Generative models over data

$$\mathbf{x}_i \sim P_{\text{data}}$$
$$i = 1, 2, \ldots, n$$

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

$$\arg\min_p \mathrm{D}_{KL}(p_{\text{data}} \| p) = \max_p \frac{1}{|D|} \sum_{x_i \in D} \log p(x_i) \qquad (0.1)$$

[1] Image credit: CS 236

Generative models over data

$$\mathbf{x}_i \sim P_{\text{data}}$$
$$i = 1, 2, \ldots, n$$

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

$$\arg\min_p \mathrm{D}_{KL}(p_{\text{data}} \| p) = \max_p \frac{1}{|D|} \sum_{x_i \in D} \log p(x_i) \tag{0.1}$$

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.

Latent variable models

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

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

Latent variable models

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

Data likelihood under latent variable models

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

Data likelihood under latent variable models

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

Data likelihood under latent variable models

General solution: Introduce inference distribution $q(zjx)$.
Intuition:

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

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

Intuition:

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

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

Intuition:

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

$$\log p\ (x) =$$

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

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

$$\log \int_z \frac{q_\phi(z|x)}{q_\phi(z|x)} p_\theta(x|z) p(z) dz =$$

$$\underbrace{\int_z}_{=1}$$

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

$$\log \int \frac{q(z|x)}{q(z|x)} p(x|z)p(z)dz = \log \int q(z|x)\frac{p(x;z)}{q(z|x)}dz =$$

$$\underbrace{\int q(z|x)}_{=1}$$

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

$$\log \int_z \frac{q_\phi(z|x)}{q_\phi(z|x)} p_\theta(x|z)p(z)dz = \log \int_z q_\phi(z|x)\frac{p_\theta(x;z)}{q_\phi(z|x)}dz =$$

$$\log \underbrace{E_{z \sim q_\phi(z|x)}}_{= 1} \left[\frac{p_\theta(x;z)}{q_\phi(z|x)}\right]$$

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

$$\log \int \frac{q(z|x)}{q(z|x)} p(x|z)p(z)dz = \log \int q(z|x)\frac{p(x;z)}{q(z|x)}dz =$$

$$\log \underbrace{E_{z\sim q(z|x)}\left[\frac{p(x;z)}{q(z|x)}\right]}_{=1} \geq \underbrace{E_{z\sim q(z|x)}\left[\log \frac{p(x;z)}{q(z|x)}\right]}_{\text{Jensen}} =$$

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

$$\log \int_z \frac{q_\phi(z|x)}{q_\phi(z|x)} p_\theta(x|z)p(z)dz = \log \int_z q_\phi(z|x)\frac{p_\theta(x;z)}{q_\phi(z|x)}dz =$$

$$\log \underbrace{\int_z}_{=1} \log E_{z \sim q_\phi(z|x)}\left[\frac{p_\theta(x;z)}{q_\phi(z|x)}\right] \underbrace{\geq}_{\text{Jensen}} E_{z \sim q_\phi(z|x)}\left[\log \frac{p_\theta(x;z)}{q_\phi(z|x)}\right] =$$

$$E_{z \sim q_\phi(z|x)}[\log p_\theta(x|z)] + \underbrace{E_{z \sim q_\phi(z|x)}[\log p(z) - \log q_\phi(z|x)]}_{= -D_{KL}(q_\phi(z|x)||p(z))} =: L_{\theta,\phi}(x)$$

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

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

Proposition

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

$$\log p_\theta(x) - D_{KL}(q(z)||p(z|x)) = E_{z \sim q(z)}[\log p_\theta(x|z)] - D_{KL}(q(z)||p(z))$$

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

Proposition

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

$$\log p_\theta(x) - D_{KL}(q(z)||p(z|x)) = E_{z \sim q(z)}[\log p_\theta(x|z)] - D_{KL}(q(z)||p(z))$$

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

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

Proposition

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

$$\log p_\theta(x) - D_{KL}(q(z)||p(z|x)) = E_{z \sim q(z)}[\log p_\theta(x|z)] - D_{KL}(q(z)||p(z))$$

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

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:

$$\log p_\theta(x) - D_{KL}(q(z)||p(z|x)) = E_{z \sim q(z)}[\log p_\theta(x|z)] - D_{KL}(q(z)||p(z))$$

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

From the properties of KL-divergences, equality is achieved only when $q(z) = p(z|x)$.

To minimize the ELBO gap we choose $q_\phi(z) = q(z|x)$.

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

$$L_{;} = \underbrace{E_{z \quad q \ (zjx)}[\log p \ (xjz)]}_{\text{reconstruction}} \quad \underbrace{D_{KL}(q \ (zjx)jjp(z))}_{\text{KL regularization}}$$

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

$$L_{;} = \underbrace{E_{z \sim q(z|x)}[\log p(x|z)]}_{\text{reconstruction}} \underbrace{D_{KL}(q(z|x)||p(z))}_{\text{KL regularization}}$$

Set $q(z|x)$ to be multivariate normal distribution parameterized by a neural network, i.e. $q(z|x) = N(z; (x); (x))$, where $(x) = \text{diag}({}^2_1(x); : : : ; {}^2_K(x))$ is a diagonal matrix.

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

$$L_{;} = \underbrace{E_{z \; q \; (zjx)}[\log p \; (xjz)]}_{\text{reconstruction}} \quad \underbrace{D_{KL}(q \; (zjx) jj \, p(z))}_{\text{KL regularization}}$$

Set $q \; (zjx)$ to be multivariate normal distribution parameterized by a neural network, i.e. $q \; (zjx) = N(z; \quad (x); \quad (x))$, where $(x) = \text{diag}(\; {}^2_1(x); : : : ; \; {}^2_K(x))$ is a diagonal matrix.

Proposition

Let $q = N(\; _1; \; _1)$ and $p = N(\; _2; \; _2)$, then

$$D_{KL}(qjj \, p) = \frac{1}{2} \Big[\text{tr}(\; _2{}^{1} \; _1) + (\; _2 \quad _1) \; _2{}^{1}(\; _2 \quad _1) \quad k + \log \frac{\det \; _2}{\det \; _1} \Big]$$

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

$$L_{;} = \underbrace{E_{z \quad q \ (zjx)}[\log p \ (xjz)]}_{\text{reconstruction}} \quad \underbrace{D_{KL}(q \ (zjx)jj\,p(z))}_{\text{KL regularization}}$$

Set $q\ (zjx)$ to be multivariate normal distribution parameterized by a neural network, i.e. $q\ (zjx) = N(z;\ (x);\ (x))$, where $(x) = \text{diag}(\ _1^2(x); : : : ;\ _K^2(x))$ is a diagonal matrix.

Proposition

Let $q = N(\ _1;\ _1)$ and $p = N(\ _2;\ _2)$, then

$$D_{KL}(qjj\,p) = \frac{1}{2}\Big[\text{tr}(\ _2^{\ 1}\ _1) + (\ _2 \quad _1)\ _2^{\ 1}(\ _2 \quad _1) \quad k + \log\frac{\det\ _2}{\det\ _1}\Big]$$

This is a di erentiable function!

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

$$L_{;} = \underbrace{E_{z \quad q \ (z|x)}[\log p \ (x|z)]}_{\text{reconstruction}} \quad \underbrace{D_{KL}(q \ (z|x)||p(z))}_{\text{KL regularization}}$$

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

$$L_{;} = \underbrace{E_{z \quad q \ (z|x)}[\log p \ (x|z)]}_{\text{reconstruction}} \quad \underbrace{D_{KL}(q \ (z|x)||p(z))}_{\text{KL regularization}}$$

Optimizing the ELBO

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

$$L_{;} = \underbrace{\mathrm{E}_{z \quad q \ (z|x)}[\log p \ (x|z)]}_{\text{reconstruction}} \quad \underbrace{\mathrm{D}_{KL}(q \ (z|x)||p(z))}_{\text{KL regularization}}$$

Need to compute

$$r_{;} \mathrm{E}_{z \quad q \ (z|x)}[\log p \ (x|z)]$$

Optimizing the ELBO

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

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

Need to compute

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

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:

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

Need to compute

$$r(x; \theta) = \mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(x|z)] \qquad r(x; \theta) = \hat{\mathbb{E}}_{z \sim q_\phi(z|x)}[\log p_\theta(x|z)]$$

Optimizing the ELBO

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

$$L_{;} = \underbrace{E_{z \; q \; (zjx)}[\log p \; (xjz)]}_{\text{reconstruction}} \underbrace{D_{KL}(q \; (zjx) jj p(z))}_{\text{KL regularization}}$$

Need to compute

$$r_{;} E_{z \; q \; (zjx)}[\log p \; (xjz)] \qquad r_{;} \hat{E}_{z \; q \; (zjx)}[\log p \; (xjz)]$$

Reparameterization Trick:

$$\tilde{z} = \quad (x) + \quad (x); \text{ where } \quad N(0; I)$$

Optimizing the ELBO

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

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

Need to compute

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

Reparameterization Trick:

$$\tilde{z} = \mu(x) + \sigma(x)\epsilon; \text{ where } \epsilon \sim N(0; I)$$

$$r(x; \theta, \phi) = \mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(x|z)] \qquad r(x; \theta, \phi) \approx \frac{1}{M} \sum_{j=1}^{M} \log p_\theta(x|\underbrace{\mu(x) + \tilde{z}^{(j)} \sigma(x)}_{\tilde{z}^{(j)} \sim N(\mu(x); \sigma(x))})$$

Optimizing the ELBO

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

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

Need to compute

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

Reparameterization Trick:

$$\tilde{z} = \mu(x) + \sigma(x)\epsilon; \text{ where } \epsilon \sim N(0; I)$$

$$r(x; \theta, \phi) = \mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(x|z)] \qquad r(x; \theta, \phi) \approx \frac{1}{M} \sum_{j=1}^{M} \log p_\theta(x|\underbrace{\mu(x) + \sigma(x)\epsilon^{(j)}}_{z^{(j)} \sim N(\mu(x); \sigma(x))})$$

In practice usually $M = 1$.

Variational Auto Encoder (VAE)

Variational Auto Encoder (VAE)

Variational Auto Encoder (VAE)

$$\max_{;} \frac{1}{N} \sum_{i=1}^{N} \log p\ (x_i j \quad (x) + \quad (x)) \quad D_{KL}[N(\quad (x_i);\quad (x_i))jjN(0; I)]$$

What does the VAE actually do

$$\max_{;} = \underbrace{E_{z \quad q \ (zjx)}[\log p \ (xjz)]}_{\text{reconstruction}} \quad \underbrace{D_{KL}(q \ (zjx)jjp(z))}_{\text{KL regularization}}$$

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

What does the VAE actually do

$$\max_{;} = \underbrace{E_{z \quad q \ (zjx)}[\log p \ (xjz)]}_{\text{reconstruction}} \quad \underbrace{D_{KL}(q \ (zjx)jjp(z))}_{\text{KL regularization}}$$

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

What does the VAE actually do

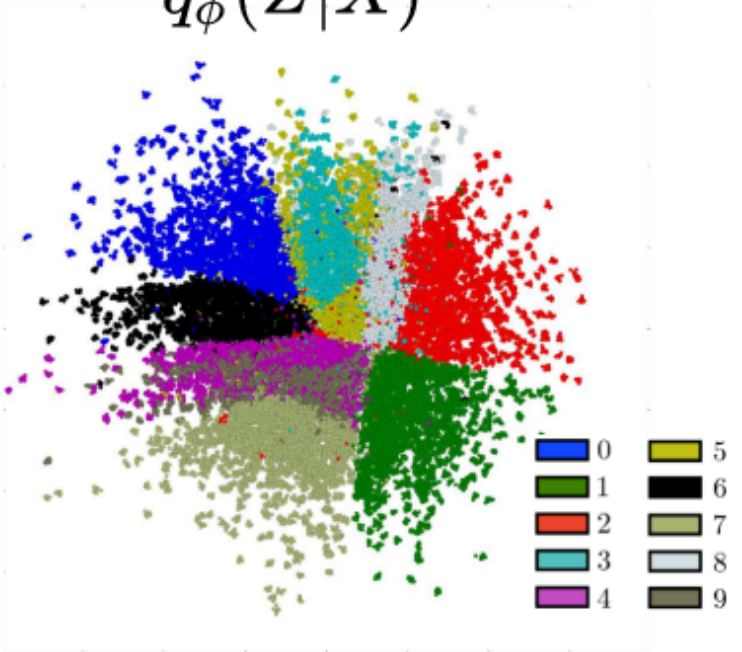
$$\max_{;} = \underbrace{E_{z \quad q \ (zjx)}[\log p \ (xjz)]}_{\text{reconstruction}} \quad \underbrace{D_{KL}(q \ (zjx)jjp(z))}_{\text{KL regularization}}$$

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

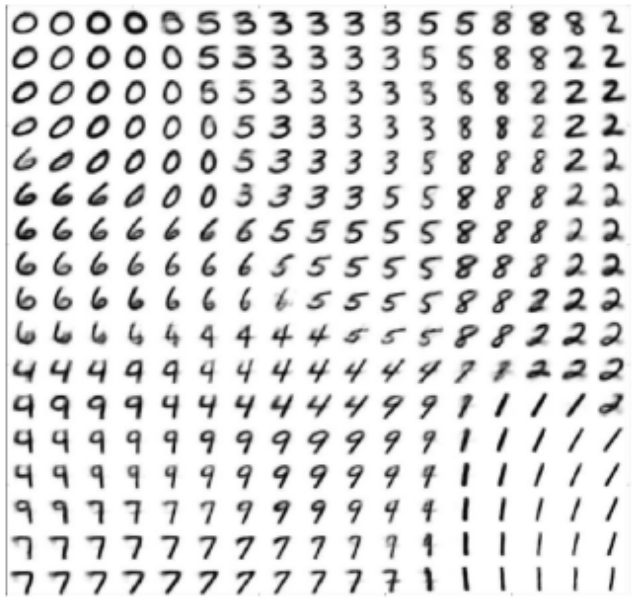
[1] Image credit: Jeremy Jordan's blog post.

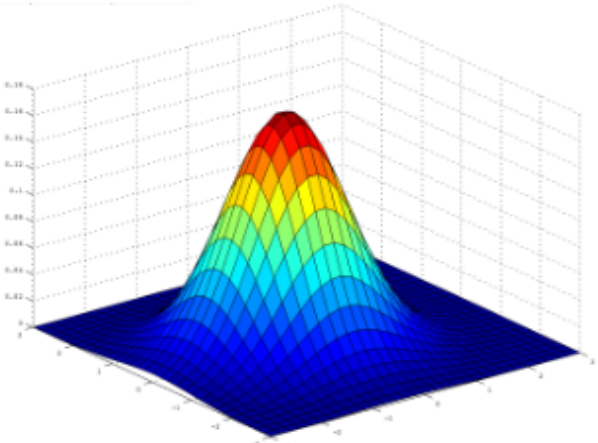
Example: MNIST VAE

$q_\phi(Z|X)$



$p_\theta(X|Z)$





$p(Z)$

Conditional VAEs

Conditional VAEs

Figure: Samples from CVAE trained on SVHN

Conditional VAEs

Figure: Samples from CVAE trained on SVHN

$$\max_{;} \mathrm{E}_{z \quad q \ (z|x;y)}[\log p \ (x|z;y)] \quad \mathrm{D}_{KL}(q \ (z|x;y)||p \ (z|y))$$

Conditional VAEs

Figure: Samples from CVAE trained on SVHN

$$\max_{;} \mathsf{E}_{z \quad q \ (z|x;y)}[\log p \ (x|z;y)] \quad \mathsf{D}_{KL}(q \ (z|x;y)||p \ (z|y))$$

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

Final Questions

Questions?