ELBO and KL-Divergence

Kunal Gupta, Tianyu Wang

University of California, San Diego

k5gupta@eng.ucsd.edu

December 4, 2018
Problem Formulation

Consider a dataset \( \mathbf{X} = \{ \mathbf{x}_i \}_{i=1}^N \) of \( N \) i.i.d. samples of some continuous/discrete random variable \( x \).

We assume that the random variable \( x \) is generated from some unobserved/latent continuous random variable \( z \), as shown in Fig 1.

**Figure:** Graphical model to be considered for the latent variable \( z \) and observed variable \( x \). Solid lines denote the generative (decoding) model

\[
p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z}),
\]

while the dashed lines denote the variational approximation (encoding) model \( q_\phi(\mathbf{z}|\mathbf{x}) \).
Problem Formulation

Consider a dataset $\mathbf{X} = \{x_i\}_{i=1}^{N}$ of $N$ i.i.d. samples of some continuous/discrete random variable $x$.

Figure: Graphical model to be considered for the latent variable $z$ and observed variable $x$. Solid lines denote the generative (decoding) model $p_\theta(x, z) = p_\theta(x|z)p_\theta(z)$, while the dashed lines denote the variational approximation (encoding) model $q_\phi(z|x)$. 
Consider a dataset $\mathbf{X} = \{\mathbf{x}_i\}^N_{i=1}$ of $N$ i.i.d. samples of some continuous/discrete random variable $\mathbf{x}$.

We assume that the random variable $\mathbf{x}$ is generated from some unobserved/latent continuous random variable $\mathbf{z}$, as shown in Fig 1.

**Figure:** Graphical model to be considered for the latent variable $\mathbf{z}$ and observed variable $\mathbf{x}$. Solid lines denote the generative (decoding) model $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$, while the dashed lines denote the variational approximation (encoding) model $q_\phi(\mathbf{z}|\mathbf{x})$. 
Each sample $x_i$ is generated from the following process:

A value $z_i$ is sampled from some prior distribution $p_{\theta}(z_i)$.

A value $x_i$ is sampled from some likelihood distribution $p_{\theta}(x_i | z_i)$.

We wish to calculate the posterior distribution $p_{\theta}(z_i | x_i)$.

Calculating $p_{\theta}(x_i)$ is hard.

Although we can assume that $p_{\theta}(z_i)$ and $p_{\theta}(x_i | z_i)$ are from some parametric family, getting the posterior distribution $p_{\theta}(z_i | x_i)$ is generally intractable due to the integration of the marginal $p_{\theta}(x_i) = \int p_{\theta}(x_i | z_i) p_{\theta}(z_i) dz_i$. 

Kunal Gupta, Tianyu Wang (UC San Diego)
Each sample $x_i$ is generated in from the following process:
Each sample $x_i$ is generated in the following process:

- A value $z_i$ is sampled from some prior distribution $p_\theta(z)$
Each sample $x_i$ is generated in from the following process:
- A value $z_i$ is sampled from some prior distribution $p_\theta(z)$
- A value $x_i$ is sampled from some likelihood distribution $p_\theta(x|z)$
Each sample $x_i$ is generated in from the following process:
- A value $z_i$ is sampled from some prior distribution $p_\theta(z)$
- A value $x_i$ is sampled from some likelihood distribution $p_\theta(x|z)$

We wish to calculate the posterior distribution $p_\theta(z|x)$.
Each sample $x_i$ is generated in from the following process:
- A value $z_i$ is sampled from some prior distribution $p_\theta(z)$
- A value $x_i$ is sampled from some likelihood distribution $p_\theta(x|z)$

We wish to calculate the posterior distribution $p_\theta(z|x)$.

Calculating $p_\theta(x)$ is hard

Although we can assume that $p_\theta(z)$ and $p_\theta(x|z)$ are from some parametric family, getting the posterior distribution $p_\theta(z|x)$ is generally intractable due to the integration of the marginal $p_\theta(x) = \int p_\theta(x|z)p_\theta(z)dz$
Proposed solution

In variational inference, we propose a posterior $q_{\phi}(z|x)$ of some parametric form to approximate the generally intractable true posterior $p_{\theta}(z|x)$. 

Kunal Gupta, Tianyu Wang (UC San Diego)

ERL Discussions

December 4, 2018
In variational inference, we propose a posterior $q_{\phi}(z|x)$ of some parametric form to approximate the generally intractable true posterior $p_{\theta}(z|x)$. 
Methodology

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

(1)
Methodology

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

(1)

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

(2)
\[
\log p_\theta(x) = \mathbb{E}_{z \sim q_\phi(z|x_i)} \left[ \log p_\theta(x) \right] \tag{1}
\]

\[
= \mathbb{E}_{z \sim q_\phi(z|x_i)} \left[ \log \frac{p_\theta(x, z)}{p_\theta(z|x)} \right] \tag{2}
\]

\[
= \mathbb{E}_{z \sim q_\phi(z|x_i)} \left[ \log p_\theta(x, z) - \log p_\theta(z|x) - \log q_\phi(z|x) + \log q_\phi(z|x) \right] \tag{3}
\]
Methodology

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

(1)

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

(2)

\[
= \mathbb{E}_{z \sim q_\phi(z|x_i)} [\log p_\theta(x, z) - \log p_\theta(z|x) - \log q_\phi(z|x) + \log q_\phi(z|x)] 
\]

(3)

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

(4)
Methodology

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

(1)

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

(2)

\[
= \mathbb{E}_{z \sim q_\phi(z|x_i)} \left[ \log p_\theta(x, z) - \log p_\theta(z|x) - \log q_\phi(z|x) + \log q_\phi(z|x) \right]
\]

(3)

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

(4)

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

(5)
Methodology

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

(1)

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

(2)

\[
= \mathbb{E}_{z \sim q_\phi(z|x_i)} [\log p_\theta(x, z) - \log p_\theta(z|x) - \log q_\phi(z|x) + \log q_\phi(z|x)]
\]

(3)

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

(4)

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

(5)

\[
\geq \mathbb{E}_{z \sim q_\phi(z|x_i)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right] = ELBO(\phi, \theta; x_i)
\]

(6)
Note:

Maximum value of $ELBO(\phi, \theta; x_i)$ is the best possible estimate of $\log p_{\theta}(x)$ with variational posterior.
Methodology

Note:
Maximum value of \( ELBO(\phi, \theta; x_i) \) is the best possible estimate of \( \log p_\theta(x) \) with variational posterior.

Alternatively, KL - ELBO relation

\[
KL[q_\phi(z|x) || p_\theta(z|x)] = \log p_\theta(x) - ELBO(\phi, \theta; x_i) \]

Thus, maximizing \( ELBO(\phi, \theta; x_i) \) is same as minimizing \( KL[q_\phi(z|x) || p_\theta(z|x)] \)
Note:

Maximum value of $ELBO(\phi, \theta; x_i)$ is the best possible estimate of $\log p_\theta(x)$ with variational posterior.

Alternatively, KL - ELBO relation

$$KL[q_\phi(z|x)||p_\theta(z|x)] = \log p_\theta(x) - ELBO(\phi, \theta; x_i)$$

Thus, maximizing $ELBO(\phi, \theta; x_i)$ is same as minimizing $KL[q_\phi(z|x)||p_\theta(z|x)]$

$$ELBO(\phi, \theta; x_i) = \mathbb{E}_{z \sim q_\phi(z|x_i)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right]$$  \hspace{1cm} (7)
Methodology

Note:

Maximum value of $ELBO(\phi, \theta; x_i)$ is the best possible estimate of $\log p_{\theta}(x)$ with variational posterior.

Alternatively, KL - ELBO relation

$KL [q_\phi(z|x)||p_\theta(z|x)] = \log p_{\theta}(x) - ELBO(\phi, \theta; x_i)$ Thus, maximizing $ELBO(\phi, \theta; x_i)$ is same as minimizing $KL [q_\phi(z|x)||p_\theta(z|x)]$

\[
ELBO(\phi, \theta; x_i) = \mathbb{E}_{z \sim q_\phi(z|x_i)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right] 
\] (7)

\[
= \mathbb{E}_{z \sim q_\phi(z|x_i)} \left[ \log p_\theta(x|z) + \log \frac{p_\theta(z)}{q_\phi(z|x)} \right] 
\] (8)
Methodology

Note:

Maximum value of $ELBO(\phi, \theta; x_i)$ is the best possible estimate of $\log p_\theta(x)$ with variational posterior.

Alternatively, KL - ELBO relation

$$KL [q_\phi(z|x) \| p_\theta(z|x)] = \log p_\theta(x) - ELBO(\phi, \theta; x_i)$$

Thus, maximizing $ELBO(\phi, \theta; x_i)$ is same as minimizing $KL [q_\phi(z|x) \| p_\theta(z|x)]$

$$ELBO(\phi, \theta; x_i) = \mathbb{E}_{z \sim q_\phi(z|x_i)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right]$$ (7)

$$= \mathbb{E}_{z \sim q_\phi(z|x_i)} \left[ \log p_\theta(x|z) + \log \frac{p_\theta(z)}{q_\phi(z|x)} \right]$$ (8)

$$= \mathbb{E}_{z \sim q_\phi(z|x_i)} \left[ \log p_\theta(x|z) \right] - KL [q_\phi(z|x) \| p_\theta(z)]$$ (9)
In Eqn. 9, we will differentiate and optimize the ELBO w.r.t. the encoder parameter $\phi$ and decoder parameter $\theta$. Note $KL[q_\phi(z|x)||p_\theta(z)]$ will be easy to solve for simple distributions. As the expression is available in closed form (Gaussian). While $\nabla_\theta \text{ELBO}$ is trivial, $\nabla_\phi \text{ELBO}$ is problematic due to the expected value over $z$. To estimate the gradient of the form $\nabla_\phi E_{z \sim q_\phi(z)}[f(z)]$, we derive a score function $\hat{I}_1(\phi)$. 
In Eqn. 9, we will differentiate and optimize the ELBO w.r.t. the encoder parameter $\phi$ and decoder parameter $\theta$.

**Note**

$KL[q_\phi(z|x) || p_\theta(z)]$ will be easy to solve for simple distributions. As the expression is available in closed form (Gaussian).
Maximizing ELBO

- In Eqn.9, we will differentiate and optimize the ELBO w.r.t. the encoder parameter $\phi$ and decoder parameter $\theta$.

**Note**

$KL[q_\phi(z|x)||p_\theta(z)]$ will be easy to solve for simple distributions. As the expression is available in closed form (Gaussian).

- While $\nabla_\theta ELBO$ is trivial, $\nabla_\phi ELBO$ is problematic due to the expected value over $z$. 
In Eqn. 9, we will differentiate and optimize the ELBO w.r.t. the encoder parameter $\phi$ and decoder parameter $\theta$.

**Note**

$KL[q_\phi(z|x) \| p_\theta(z)]$ will be easy to solve for simple distributions. As the expression is available in closed form (Gaussian).

While $\nabla_\theta ELBO$ is trivial, $\nabla_\phi ELBO$ is problematic due to the expected value over $z$.

To estimate the gradient of the form $\nabla_\phi \mathbb{E}_{z \sim q_\phi(z)}[f(z)]$, we derive a score function $\hat{I}_1(\phi)$. 
Maximizing ELBO

\[ \nabla_{\phi} \mathbb{E}_{z \sim q_{\phi}(z)}[f(z)] = \int \nabla_{\phi} q_{\phi}(z) f(z) dz \] (10)
Maximizing ELBO

\[ \nabla_\phi \mathbb{E}_{z \sim q_\phi(z)}[f(z)] = \int \nabla_\phi q_\phi(z)f(z) dz \] 

(10)

\[ = \int \frac{q_\phi(z)}{q_\phi(z)} \nabla_\phi q_\phi(z)f(z) dz \] 

(11)
Maximizing ELBO

\[
\nabla_{\phi} \mathbb{E}_{z \sim q_{\phi}(z)}[f(z)] = \int \nabla_{\phi} q_{\phi}(z) f(z) dz
\]

(10)

\[= \int q_{\phi}(z) \frac{\nabla_{\phi} q_{\phi}(z)}{q_{\phi}(z)} f(z) dz\]

(11)

\[= \int q_{\phi}(z) \nabla_{\phi} \log q_{\phi}(z) f(z) dz\]

(12)
Maximizing ELBO

\[ \nabla_\phi \mathbb{E}_{z \sim q_\phi(z)}[f(z)] = \int \nabla_\phi q_\phi(z)f(z)dz \quad (10) \]

\[ = \int \frac{q_\phi(z)}{q_\phi(z)} \nabla_\phi q_\phi(z)f(z)dz \quad (11) \]

\[ = \int q_\phi(z)\nabla_\phi \log q_\phi(z)f(z)dz \quad (12) \]

\[ = \mathbb{E}_{z \sim q_\phi(z)}[\nabla_\phi \log q_\phi(z)f(z)] = \mathbb{E}_{z \sim q_\phi(z)}[\hat{I}_1(\phi)] \quad (13) \]
\[ \nabla_{\phi} E_{z \sim q_{\phi}(z)}[f(z)] = \int \nabla_{\phi} q_{\phi}(z)f(z)dz \quad (10) \]

\[ = \int \frac{q_{\phi}(z)}{q_{\phi}(z)} \nabla_{\phi} q_{\phi}(z)f(z)dz \quad (11) \]

\[ = \int q_{\phi}(z) \nabla_{\phi} \log q_{\phi}(z)f(z)dz \quad (12) \]

\[ = E_{z \sim q_{\phi}(z)}[\nabla_{\phi} \log q_{\phi}(z)f(z)] = E_{z \sim q_{\phi}(z)}[\hat{I}_1(\phi)] \quad (13) \]

\[ \hat{I}_1(\phi) = f(z) \frac{\partial \log q_{\phi}(z)}{\partial \phi}, \quad (14) \]
Maximizing ELBO

\[ \nabla_{\phi} \mathbb{E}_{z \sim q_{\phi}(z)}[f(z)] = \int \nabla_{\phi} q_{\phi}(z)f(z)dz \]  
(10)

\[ = \int \frac{q_{\phi}(z)}{q_{\phi}(z)} \nabla_{\phi} q_{\phi}(z)f(z)dz \]  
(11)

\[ = \int q_{\phi}(z) \nabla_{\phi} \log q_{\phi}(z)f(z)dz \]  
(12)

\[ = \mathbb{E}_{z \sim q_{\phi}(z)}[\nabla_{\phi} \log q_{\phi}(z)f(z)] = \mathbb{E}_{z \sim q_{\phi}(z)}[\hat{I}_1(\phi)] \]  
(13)

\[ \hat{I}_1(\phi) = f(z) \frac{\partial \log q_{\phi}(z)}{\partial \phi}, \]  
(14)

The gradient can be approximated by MC Sampling from \( z_i \sim q_{\phi}(z) \).

\[ \nabla_{\phi} \mathbb{E}_{z \sim q_{\phi}(z)}[f(z)] \approx \frac{1}{M} \sum_{l=1}^{M} f(z_l) \frac{\partial \log q_{\phi}(z_l)}{\partial \phi} \]  
(15)
The score function estimator is simple but suffers from high variance so in practice, the re-parametrization trick is used.
Maximizing ELBO

- The score function estimator is simple but suffers from high variance so in practice, the re-parametrization trick is used.
- Assuming that we can re-parameterize the random variable $z \sim q_\phi(z|x_i)$ with a deterministic differentiable transformation $(g_\phi)$ of some parameter-free auxiliary variable $\epsilon$:
Maximizing ELBO

- The score function estimator is simple but suffers from high variance so in practice, the re-parametrization trick is used.
- Assuming that we can re-parameterize the random variable \( z \sim q_\phi(z|x_i) \) with a deterministic differentiable transformation \((g_\phi)\) of some parameter-free auxiliary variable \( \epsilon \):

\[
z = g_\phi(x_i, \epsilon) \text{ with } \epsilon \sim p(\epsilon),
\]

(16)
The score function estimator is simple but suffers from high variance so in practice, the re-parametrization trick is used.

Assuming that we can re-parameterize the random variable \( z \sim q_\phi(z|x_i) \) with a deterministic differentiable transformation \((g_\phi)\) of some parameter-free auxiliary variable \( \epsilon \):

\[
z = g_\phi(x_i, \epsilon) \quad \text{with} \quad \epsilon \sim p(\epsilon),
\]

(16)

We can estimate with the gradient with the pathwise derivative estimator

\[
\hat{I}_2(\phi) = f'(g_\phi(x_i, \epsilon)) \frac{\partial g_\phi(x_i, \epsilon)}{\partial \phi}
\]

(17)
Maximizing ELBO

- The score function estimator is simple but suffers from high variance so in practice, the re-parametrization trick is used.
- Assuming that we can re-parameterize the random variable \( z \sim q_\phi(z|x_i) \) with a deterministic differentiable transformation \((g_\phi)\) of some parameter-free auxiliary variable \( \epsilon \):

\[
z = g_\phi(x_i, \epsilon) \quad \text{with} \quad \epsilon \sim p(\epsilon), \quad (16)
\]

We can estimate with the gradient with the pathwise derivative estimator

\[
\hat{I}_2(\phi) = f'(g_\phi(x_i, \epsilon)) \frac{\partial g_\phi(x_i, \epsilon)}{\partial \phi} \quad (17)
\]

and the gradient can be approximated by

\[
\nabla_\phi \mathbb{E}_{z \sim q_\phi(z|x_i)}[f(z)] \approx \frac{1}{M} \sum_{l=1}^{M} f'(g_\phi(x_i, \epsilon_l)) \frac{\partial g_\phi(x_i, \epsilon_l)}{\partial \phi} \quad (18)
\]

with \( \epsilon_l \sim p(\epsilon) \).
The End