ELBO and KL-Divergence

Kunal Gupta, Tianyu Wang

University of California, San Diego

k5gupta@eng.ucsd.edu

December 4, 2018

Problem Formulation

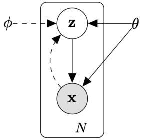


Figure: Graphical model to be considered for the latent variable $\mathbf{z}$ and observed variable $\mathbf{x}$. Solid lines denote the generative (decoding) model $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$, while the dashed lines denote the variational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

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 $\mathbf{x}$.

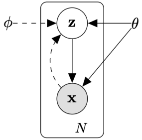


Figure: Graphical model to be considered for the latent variable $\mathbf{z}$ and observed variable $\mathbf{x}$. Solid lines denote the generative (decoding) model $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$, while the dashed lines denote the variational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

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

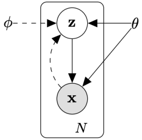
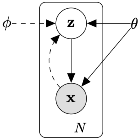
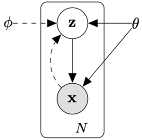


Figure: Graphical model to be considered for the latent variable $\mathbf{z}$ and observed variable $\mathbf{x}$. Solid lines denote the generative (decoding) model $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$, while the dashed lines denote the variational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

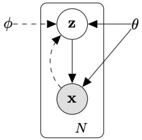


ϕ
$\mathbf{z}$
θ
$\mathbf{x}$
N

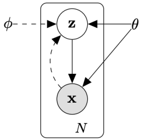
- Each sample $\mathbf{x}_i$ is generated in from the following process:



- Each sample $\mathbf{x}_i$ is generated in from the following process:
  - A value $\mathbf{z}_i$ is sampled from some prior distribution $p_\theta(\mathbf{z})$

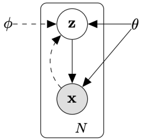


- Each sample $\mathbf{x}_i$ is generated in from the following process:
  - A value $\mathbf{z}_i$ is sampled from some prior distribution $p_\theta(\mathbf{z})$
  - A value $\mathbf{x}_i$ is sampled from some likelihood distribution $p_\theta(\mathbf{x}|\mathbf{z})$



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

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

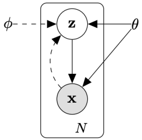


- Each sample $\mathbf{x}_i$ is generated in from the following process:
  - A value $\mathbf{z}_i$ is sampled from some prior distribution $p_\theta(\mathbf{z})$
  - A value $\mathbf{x}_i$ is sampled from some likelihood distribution $p_\theta(\mathbf{x}|\mathbf{z})$

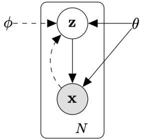
We wish to calculate the posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$.

Calculating $p_\theta(\mathbf{x})$ is hard

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

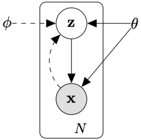


Proposed solution



Proposed solution

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



Methodology

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

Methodology

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

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

Methodology

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

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

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)}[\log p_\theta(x,z) - \log p_\theta(z|x) - \log q_\phi(z|x) + \log q_\phi(z|x)] \tag{3}$$

Methodology

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

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

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)}[\log p_\theta(x,z) - \log p_\theta(z|x) - \log q_\phi(z|x) + \log q_\phi(z|x)] \tag{3}$$

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)} + \log \frac{q_\phi(z|x)}{p_\theta(z|x)}\right] \tag{4}$$

Methodology

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

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

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)}[\log p_\theta(x,z) - \log p_\theta(z|x) - \log q_\phi(z|x) + \log q_\phi(z|x)] \tag{3}$$

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)} + \log \frac{q_\phi(z|x)}{p_\theta(z|x)}\right] \tag{4}$$

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] + KL\left[q_\phi(z|x)||p_\theta(z|x)\right] \tag{5}$$

Methodology

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

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

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)}[\log p_\theta(x,z) - \log p_\theta(z|x) - \log q_\phi(z|x) + \log q_\phi(z|x)] \tag{3}$$

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)} + \log \frac{q_\phi(z|x)}{p_\theta(z|x)}\right] \tag{4}$$

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] + KL\left[q_\phi(z|x)||p_\theta(z|x)\right] \tag{5}$$

$$\geq \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = ELBO(\phi,\theta;x_i) \tag{6}$$

Methodology

Note:

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

Methodology

Note:

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

Alternatively, KL - ELBO relation

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

Methodology

Note:

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

Alternatively, KL - ELBO relation

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

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

Methodology

Note:

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

Alternatively, KL - ELBO relation

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

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

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

Methodology

Note:

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

Alternatively, KL - ELBO relation

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

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

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

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)}\left[\log p_\theta(x|z)\right] - KL\left[q_\phi(z|x)||p_\theta(z)\right] \tag{9}$$

Maximizing ELBO

- In Eqn.9, we will differentiate and optimize the ELBO w.r.t. the encoder parameter ϕ and decoder parameter θ.

Maximizing ELBO

- In Eqn.9, we will differentiate and optimize the ELBO w.r.t. the encoder parameter ϕ and decoder parameter θ.

Note

$KL\left[q_\phi(z|x)||p_\theta(z)\right]$ 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 ϕ and decoder parameter θ.

Note

$KL\left[q_\phi(z|x)||p_\theta(z)\right]$ 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 $\mathbf{z}$.

Maximizing ELBO

- In Eqn.9, we will differentiate and optimize the ELBO w.r.t. the encoder parameter ϕ and decoder parameter θ.

Note

$KL\left[q_\phi(z|x)||p_\theta(z)\right]$ 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 $\mathbf{z}$.

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

Maximizing ELBO

$$\nabla_\phi \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z})}[f(\mathbf{z})] = \int \nabla_\phi q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \tag{10}$$

Maximizing ELBO

$$\nabla_\phi \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z})}[f(\mathbf{z})] = \int \nabla_\phi q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \tag{10}$$

$$= \int \frac{q_\phi(\mathbf{z})}{q_\phi(\mathbf{z})} \nabla_\phi q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \tag{11}$$

Maximizing ELBO

$$\nabla_\phi \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z})}[f(\mathbf{z})] = \int \nabla_\phi q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \tag{10}$$

$$= \int \frac{q_\phi(\mathbf{z})}{q_\phi(\mathbf{z})} \nabla_\phi q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \tag{11}$$

$$= \int q_\phi(\mathbf{z}) \nabla_\phi \log q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \tag{12}$$

Maximizing ELBO

$$\nabla_\phi \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z})}[f(\mathbf{z})] = \int \nabla_\phi q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \tag{10}$$

$$= \int \frac{q_\phi(\mathbf{z})}{q_\phi(\mathbf{z})} \nabla_\phi q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \tag{11}$$

$$= \int q_\phi(\mathbf{z}) \nabla_\phi \log q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \tag{12}$$

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

Maximizing ELBO

$$\nabla_\phi \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z})}[f(\mathbf{z})] = \int \nabla_\phi q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \tag{10}$$

$$= \int \frac{q_\phi(\mathbf{z})}{q_\phi(\mathbf{z})} \nabla_\phi q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \tag{11}$$

$$= \int q_\phi(\mathbf{z}) \nabla_\phi \log q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \tag{12}$$

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

$$\hat{l}_1(\phi) = f(\mathbf{z}) \frac{\partial \log q_\phi(\mathbf{z})}{\partial \phi}, \tag{14}$$

Maximizing ELBO

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

$$= \int \frac{q_\phi(\mathbf{z})}{q_\phi(\mathbf{z})} \nabla_\phi q_\phi(\mathbf{z}) f(\mathbf{z}) dz \tag{11}$$

$$= \int q_\phi(\mathbf{z}) \nabla_\phi \log q_\phi(\mathbf{z}) f(\mathbf{z}) dz \tag{12}$$

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

$$\hat{l}_1(\phi) = f(\mathbf{z}) \frac{\partial \log q_\phi(\mathbf{z})}{\partial \phi}, \tag{14}$$

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

$$\nabla_\phi \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z})}[f(\mathbf{z})] \approx \frac{1}{M} \sum_{l=1}^{M} f(\mathbf{z}_l) \frac{\partial \log q_\phi(\mathbf{z}_l)}{\partial \phi} \tag{15}$$

Maximizing ELBO

- 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 $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}_i)$ with a deterministic differentiable transformation (g_ϕ) of some parameter-free auxiliary variable ϵ:

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 $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}_i)$ with a deterministic differentiable transformation (g_ϕ) of some parameter-free auxiliary variable ϵ:

$$z = g_\phi(\mathbf{x}_i, \epsilon) \;\; \text{with} \quad \epsilon \sim p(\epsilon), \tag{16}$$

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 $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}_i)$ with a deterministic differentiable transformation (g_ϕ) of some parameter-free auxiliary variable ϵ:

$$z = g_\phi(\mathbf{x}_i, \epsilon) \ \text{ with } \ \ \epsilon \sim p(\epsilon), \tag{16}$$

We can estimate with the gradient with the pathwise derivative estimator

$$\hat{I}_2(\phi) = f'(g_\phi(\mathbf{x}_i, \epsilon))\frac{\partial g_\phi(\mathbf{x}_i, \epsilon)}{\partial \phi} \tag{17}$$

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 $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}_i)$ with a deterministic differentiable transformation (g_ϕ) of some parameter-free auxiliary variable ϵ:

$$z = g_\phi(\mathbf{x}_i, \epsilon) \text{ with } \epsilon \sim p(\epsilon), \tag{16}$$

We can estimate with the gradient with the pathwise derivative estimator

$$\hat{l}_2(\phi) = f'(g_\phi(\mathbf{x}_i, \epsilon)) \frac{\partial g_\phi(\mathbf{x}_i, \epsilon)}{\partial \phi} \tag{17}$$

and the gradient can be approximated by

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

with $\epsilon_l \sim p(\epsilon)$.

The End