A ~~Short~~ Tutorial on Variational Auto Encoders

Ehsan Shareghi

July 2018

Inference Problem

Imagine the following inference problem, where a latent variable of interest, z, needs to be inferred given an input data x:

$$\underbrace{P(z|x)}_{\text{Posterior}} = \frac{\overbrace{P(x|z)}^{\text{Likelihood}}\ \overbrace{p(z)}^{\text{Prior}}}{\underbrace{P(x)}_{\text{Evidence or Marginal Likelihood}}} \quad \text{where} \quad P(x) = \int P(x|z)P(z)dz$$

Analytically intractable! Solution? Approximate $P(z|x)$.

Posterior Approximation

Two possibilities:

Sample Markov Chain Monte Carlo (MCMC) to simulate samples from the posterior. Beautiful but quite tricky!

Posterior Approximation

Two possibilities:

Sample
: Markov Chain Monte Carlo (MCMC) to simulate samples from the posterior. Beautiful but quite tricky!

Optimize
: Variational technique to start from a family of distributions and searching for the parameters that results in the closest approximation of the posterior.

Distributions Closeness

Several possibilities (active research area)! Widely used is `KL-divergence`:

$$\mathtt{KL}(\underbrace{Q(z|x)}_{\text{approx.}} \,||\, \underbrace{P(z|x)}_{\text{true}}) \quad \text{and} \quad \mathtt{KL}(\underbrace{P(z|x)}_{\text{true}} \,||\, \underbrace{Q(z|x)}_{\text{approx.}})$$

`KL` is not symmetric! Which one to choose?

Distributions Closeness

Several possibilities (active research area)! Widely used is `KL-divergence`:

$$\texttt{KL}(\underbrace{Q(z|x)}_{\text{approx.}} \,||\, \underbrace{P(z|x)}_{\text{true}}) \quad \text{and} \quad \texttt{KL}(\underbrace{P(z|x)}_{\text{true}} \,||\, \underbrace{Q(z|x)}_{\text{approx.}})$$

`KL` is not symmetric! Which one to choose?

$$\texttt{KL}(Q(z|x)||P(z|x)) = \int Q(z|x) \log \frac{Q(z|x)}{P(z|x)} dz \qquad \text{high when high } Q \text{ and low } P$$

$$\texttt{KL}(P(z|x)||Q(z|x)) = \int P(z|x) \log \frac{P(z|x)}{Q(z|x)} dz \qquad \text{high when low } Q \text{ and high } P$$

Distributions Closeness

Several possibilities (active research area)! Widely used is `KL-divergence`:

$$\mathtt{KL}(\underbrace{Q(z|x)}_{\text{approx.}}\,||\,\underbrace{P(z|x)}_{\text{true}}) \quad \text{and} \quad \mathtt{KL}(\underbrace{P(z|x)}_{\text{true}}\,||\,\underbrace{Q(z|x)}_{\text{approx.}})$$

`KL` is not symmetric! Which one to choose?

$$\mathtt{KL}(Q(z|x)||P(z|x)) = \int Q(z|x)\log\frac{Q(z|x)}{P(z|x)}dz \qquad \text{high when high } Q \text{ and low } P$$

$$\mathtt{KL}(P(z|x)||Q(z|x)) = \int P(z|x)\log\frac{P(z|x)}{Q(z|x)}dz \qquad \text{high when low } Q \text{ and high } P$$

Writing `KL` as `Expectation`:

$$\mathtt{KL}(Q(z|x)||P(z|x)) = \int Q(z|x)\log\frac{Q(z|x)}{P(z|x)}dz = \left\langle \log\frac{Q(z|x)}{P(z|x)} \right\rangle_{Q(z|x)}$$

From KL-divergence to ELBO

$$\mathtt{KL}(Q(z|x)||P(z|x)) = \left\langle \log \frac{Q(z|x)}{\frac{P(x,z)}{P(x)}} \right\rangle_{Q(z|x)} = \left\langle \log \frac{Q(z|x)}{P(x,z)} P(x) \right\rangle_{Q(z|x)}$$

$$= \left\langle \log \frac{Q(z|x)}{P(x,z)} \right\rangle_{Q(z|x)} + \left\langle \log P(x) \right\rangle_{Q(z|x)}$$

From KL-divergence to ELBO

$$\mathtt{KL}(Q(z|x)||P(z|x)) = \left\langle \log \frac{Q(z|x)}{\frac{P(x,z)}{P(x)}} \right\rangle_{Q(z|x)} = \left\langle \log \frac{Q(z|x)}{P(x,z)} P(x) \right\rangle_{Q(z|x)}$$

$$= \left\langle \log \frac{Q(z|x)}{P(x,z)} \right\rangle_{Q(z|x)} + \left\langle \log P(x) \right\rangle_{Q(z|x)}$$

Where,

$$\left\langle \log P(x) \right\rangle_{Q(z|x)} = \int Q(z|x) \log P(x) dz = \log P(x) \int Q(z|x) dz = \log P(x)$$

From KL-divergence to ELBO

$$\texttt{KL}(Q(z|x)||P(z|x)) = \left\langle \log \frac{Q(z|x)}{\frac{P(x,z)}{P(x)}} \right\rangle_{Q(z|x)} = \left\langle \log \frac{Q(z|x)}{P(x,z)} P(x) \right\rangle_{Q(z|x)}$$

$$= \left\langle \log \frac{Q(z|x)}{P(x,z)} \right\rangle_{Q(z|x)} + \left\langle \log P(x) \right\rangle_{Q(z|x)}$$

Where,

$$\left\langle \log P(x) \right\rangle_{Q(z|x)} = \int Q(z|x) \log P(x) dz = \log P(x) \int Q(z|x) dz = \log P(x)$$

So,

$$\texttt{KL}(Q(z|x)||P(z|x)) = \left\langle \log \frac{Q(z|x)}{P(x,z)} \right\rangle_{Q(z|x)} + \log P(x)$$

From KL-divergence to ELBO

$$\mathtt{KL}(Q(z|x)||P(z|x)) - \left\langle \log \frac{Q(z|x)}{P(x,z)} \right\rangle_{Q(z|x)} = \underbrace{\log P(x)}_{\text{constant}}$$

From KL-divergence to ELBO

$$\mathtt{KL}(Q(z|x)||P(z|x)) - \left\langle \log \frac{Q(z|x)}{P(x,z)} \right\rangle_{Q(z|x)} = \underbrace{\log P(x)}_{\text{constant}}$$

$$\underbrace{\mathtt{KL}(Q(z|x)||P(z|x))}_{\text{always non-negative}} + \left\langle \log \frac{P(x,z)}{Q(z|x)} \right\rangle_{Q(z|x)} = \underbrace{\log P(x)}_{\text{constant}}$$

From KL-divergence to ELBO

$$\mathtt{KL}(Q(z|x)||P(z|x)) - \left\langle \log \frac{Q(z|x)}{P(x,z)} \right\rangle_{Q(z|x)} = \underbrace{\log P(x)}_{\text{constant}}$$

$$\underbrace{\mathtt{KL}(Q(z|x)||P(z|x))}_{\text{always non-negative}} + \left\langle \log \frac{P(x,z)}{Q(z|x)} \right\rangle_{Q(z|x)} = \underbrace{\log P(x)}_{\text{constant}}$$

So,

$$\text{minimizing } \mathtt{KL}(Q(z|x)||P(z|x)) \equiv \text{maximizing } \underbrace{\left\langle \log \frac{P(x,z)}{Q(z|x)} \right\rangle_{Q(z|x)}}_{\mathtt{ELBO}}$$

Maximizing ELBO

Two possibilities:

- Monte Carlo Gradient has high variance of gradient (active research area)
- Autoencoder, under certain assumptions, reduces the amount of variance

Maximizing ELBO via Monte Carlo Gradient

Objective function is : $\mathcal{L} = \left\langle \log \frac{P_\psi(x,z)}{Q_\theta(z|x)} \right\rangle_{Q_\theta(z|x)}$

Maximizing ELBO via Monte Carlo Gradient

Objective function is : $\mathcal{L} = \left\langle \log \frac{P_\psi(x,z)}{Q_\theta(z|x)} \right\rangle_{Q_\theta(z|x)}$

Relies on a very strong assumption: $\nabla_\theta \left\langle . \right\rangle_{Q_\theta(z)} = \left\langle \nabla_\theta . \right\rangle_{Q_\theta(z)}$

Maximizing ELBO via Monte Carlo Gradient

Objective function is : $\mathcal{L} = \left\langle \log \frac{P_\psi(x,z)}{Q_\theta(z|x)} \right\rangle_{Q_\theta(z|x)}$

Relies on a very strong assumption: $\nabla_\theta \left\langle . \right\rangle_{Q_\theta(z)} = \left\langle \nabla_\theta . \right\rangle_{Q_\theta(z)}$

Derivative w.r.t. variational distribution parameter θ,

$$\nabla_\theta \mathcal{L} \approx \frac{1}{K} \sum_{z_k \sim Q_\theta(z|x)} \nabla_\theta \log Q_\theta(z_k|x) \left(\log \frac{P_\psi(x, z_k)}{Q_\theta(z_k|x)} \right)$$

Maximizing ELBO via Monte Carlo Gradient

Objective function is : $\mathcal{L} = \left\langle \log \frac{P_\psi(x,z)}{Q_\theta(z|x)} \right\rangle_{Q_\theta(z|x)}$

Relies on a very strong assumption: $\nabla_\theta \left\langle \, . \, \right\rangle_{Q_\theta(z)} = \left\langle \nabla_\theta . \right\rangle_{Q_\theta(z)}$

Derivative w.r.t. variational distribution parameter θ,

$$\nabla_\theta \mathcal{L} \approx \frac{1}{K} \sum_{z_k \sim Q_\theta(z|x)} \nabla_\theta \log Q_\theta(z_k|x) \left(\log \frac{P_\psi(x, z_k)}{Q_\theta(z_k|x)} \right)$$

Derivative w.r.t. true posterior distribution parameter ψ,

$$\nabla_\psi \mathcal{L} \approx \frac{1}{K} \sum_{z_k \sim Q_\theta(z|x)} \nabla_\psi \log P_\psi(x, z_k)$$

Maximizing ELBO via Monte Carlo Gradient

Objective function is : $\mathcal{L} = \left\langle \log \frac{P_\psi(x,z)}{Q_\theta(z|x)} \right\rangle_{Q_\theta(z|x)}$

Relies on a very strong assumption: $\nabla_\theta \left\langle . \right\rangle_{Q_\theta(z)} = \left\langle \nabla_\theta . \right\rangle_{Q_\theta(z)}$

Derivative w.r.t. variational distribution parameter θ,

$$\nabla_\theta \mathcal{L} \approx \frac{1}{K} \sum_{z_k \sim Q_\theta(z|x)} \nabla_\theta \log Q_\theta(z_k|x) \left(\log \frac{P_\psi(x, z_k)}{Q_\theta(z_k|x)} \right)$$

Derivative w.r.t. true posterior distribution parameter ψ,

$$\nabla_\psi \mathcal{L} \approx \frac{1}{K} \sum_{z_k \sim Q_\theta(z|x)} \nabla_\psi \log P_\psi(x, z_k)$$

It turns out that optimizing the parameters using the samples exhibits high variance (active research area!)

Maximizing ELBO - Variance Reduction

Let us expand ELBO as follows,

$$
\begin{aligned}
\left\langle \log \frac{P(x,z)}{Q(z|x)} \right\rangle_{Q(z|x)} &= \left\langle \log \frac{P(x|z)P(z)}{Q(z|x)} \right\rangle_{Q(z|x)} \\
&= \left\langle \log P(x|z) \right\rangle_{Q(z|x)} + \left\langle \log \frac{P(z)}{Q(z|x)} \right\rangle_{Q(z|x)} \\
&= \left\langle \log P(x|z) \right\rangle_{Q(z|x)} - \texttt{KL-divergence}(Q(z|x)||P(z))
\end{aligned}
$$

Maximizing ELBO - Variance Reduction

Let us take another path to expand ELBO as follows,

$$
\begin{aligned}
\left\langle \log \frac{P(x,z)}{Q(z|x)} \right\rangle_{Q(z|x)} &= \left\langle \log \frac{P(x|z)P(z)}{Q(z|x)} \right\rangle_{Q(z|x)} \\
&= \left\langle \log P(x|z) \right\rangle_{Q(z|x)} + \left\langle \log \frac{P(z)}{Q(z|x)} \right\rangle_{Q(z|x)} \\
&= \left\langle \log P(x|z) \right\rangle_{Q(z|x)} - \texttt{KL-divergence}(Q(z|x)||P(z))
\end{aligned}
$$

If we assume $P(z)$ and $Q(z|x)$ to be Gaussian distributions, KL-divergence$(Q(z|x)||P(z))$ could be solved analytically. Hence, approximation is required only to be applied to first term, which consequently reduces the variance during the optimization.

Maximizing ELBO via Autoencoder

$$\left\langle \log P(x|z) \right\rangle_{Q(z|x)} - \texttt{KL-divergence}(Q(z|x)||P(z))$$

- The first term includes two inference components, $Q(z|x)$ inferring z from x, and $P(z|x)$ inferring x from z. Maximizing the expectation is essentially finding the optimal parameter of Q and P, such that z could be successfully inferred from x, and then x could be successfully inferred from z. Similar to autoencoders!

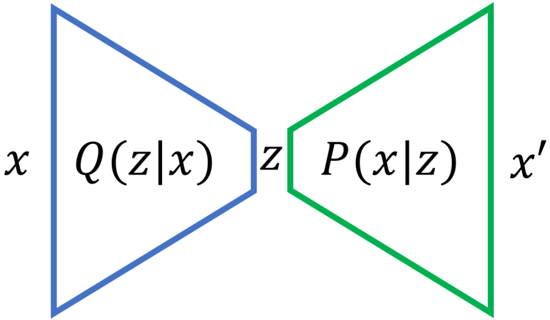
Maximizing ELBO via Autoencoder

$$\left\langle \log P(x|z) \right\rangle_{Q(z|x)} - \texttt{KL-divergence}(Q(z|x)||P(z))$$

- The first term includes two inference components, $Q(z|x)$ inferring z from x, and $P(z|x)$ inferring x from z. Maximizing the expectation is essentially finding the optimal parameter of Q and P, such that z could be successfully inferred from x, and then x could be successfully inferred from z. Similar to autoencoders!
- $\texttt{KL-divergence}(Q(z|x)||P(z))$ avoids copy-paste trivial solution.

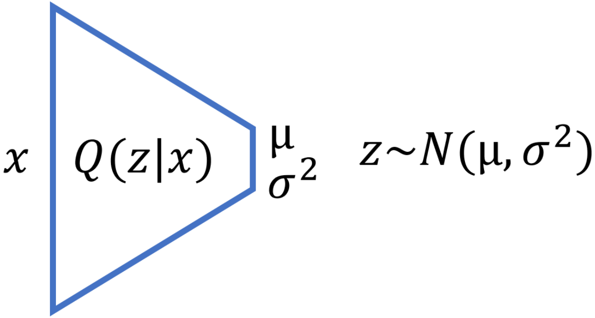
Architecture

- The functions by which encoding and decoding happens can have any neural architecture



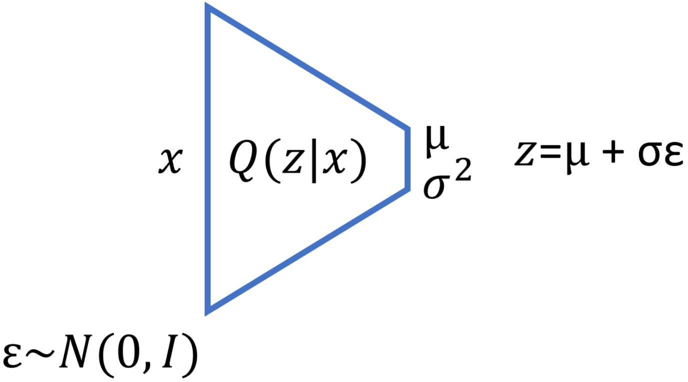
Architecture

- The functions by which encoding and decoding happens can have any neural architecture
- For the encoder, we assumed $Q(z|x)$ to be a Multivariate Gaussian with diagonal covariance.



Architecture

- The functions by which encoding and decoding happens can have any neural architecture
- For the encoder, we assume $Q(z|x)$ to be a Multivariate Gaussian with diagonal covariance.



- This reparametrization, results in a well-behaved gradient estimator,

$$\nabla_\theta \langle f_\theta(z) \rangle_{Q_\theta(z)} = \langle \nabla_\theta f_\theta(z) \rangle_{P_\gamma(\epsilon)} \quad \text{where} \quad z = g(\theta, \epsilon) \quad \text{and} \quad \epsilon \sim P_\gamma(\epsilon)$$

Refined ELBO

$$\left\langle \log P(x|z) \right\rangle_{Q(z|x)} - \texttt{KL-divergence}(Q(z|x)||P(z))$$

Is approximated as,

$$\frac{1}{K} \sum_{z_i \in \{\mu_1 + \Sigma_1 \times \epsilon : \epsilon_k \sim \mathcal{N}(0, \mathcal{I})\}_{k=1}^{K}} \log P(x|z_i) - \frac{1}{2} \sum_{i=1}^{N} \log \sigma_{i,1}^2 + 1 - \sigma_{i,1}^2 - \mu_{i,1}^2$$

This results in a well-behaved gradient estimator which avoids the high variance issue of Monte Carlo Gradients, without making any strong assumption about exchangeability of expectation and gradient.