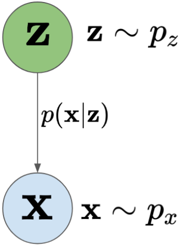
Machine Learning II: Fractal 3

Rajendra Nagar

Assistant Professor
Department of Electircal Engineering
Indian Institute of Technology Jodhpur
http://home.iitj.ac.in/~rn/

Generative Process



$$p(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

- The latent variables $\mathbf{z}$ encode semantically meaningful information about $\mathbf{x}$. Therefore, it is natural to view this generative process as first generating the "high-level" semantic information about $\mathbf{x}$ before fully generating $\mathbf{x}$.
- One way to measure how closely $p(\mathbf{x}, \mathbf{z})$ fits the observed dataset is to measure the Kullback-Leibler (KL) divergence between the data distribution (which we denote as $p_{\text{data}}(\mathbf{x})$) and the model's marginal distribution $\int_z p(\mathbf{x}, \mathbf{z})d\mathbf{z}$. The distribution that "best" fits the data is thus obtained by minimizing the KL divergence.

$$\min D_{\text{KL}}(p_{\text{data}}(\mathbf{x})||p(\mathbf{x}))$$

Optimizing the KL divergence is equivalent to maximizing the marginal log-likelihood $\log p(x)$ over the input dataset.

$$\begin{aligned} D_{\text{KL}}(p_{\text{data}}(\mathbf{x})||p(\mathbf{x})) &= \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\log p_{\text{data}}(\mathbf{x}) - \log(p(\mathbf{x}))] \\ \min D_{\text{KL}}(p_{\text{data}}(\mathbf{x})||p(\mathbf{x})) &\Leftrightarrow \max \log p(\mathbf{x}). \end{aligned}$$

Generative Process

$$
\begin{aligned}
\log p(\mathbf{x}) &= \log\left(\sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x},\mathbf{z})\right) \\
&= \log\left(\sum_{\mathbf{z}\in\mathcal{Z}} \frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z})} q(\mathbf{z})\right) \\
&= \log\left(\mathbb{E}_{\mathbf{z}\sim q_z}\left[\frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]\right) \\
&\geq \mathbb{E}_{\mathbf{z}\sim q_z}\left[\log\left(\frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right)\right] = \text{ELBO} \text{ Evidence Lower Bound} \\
\text{ELBO} &= \mathbb{E}_{\mathbf{z}\sim q_z}\left[\log\left(p(\mathbf{x},\mathbf{z})\right) - \log\left(q(\mathbf{z})\right)\right] \\
&= \mathbb{E}_{\mathbf{z}\sim q_z}\left[\log\left(p(\mathbf{x},\mathbf{z})\right)\right] - \sum_{\mathbf{z}\in\mathcal{Z}} q(\mathbf{z}) \log\left(q(\mathbf{z})\right) \\
&= \mathbb{E}_{\mathbf{z}\sim q_z}\left[\log\left(p(\mathbf{x},\mathbf{z})\right)\right] + \mathcal{H}(q)
\end{aligned}
$$

Generative Process

$$
\begin{aligned}
D_{\mathsf{KL}}(q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x})) &= \mathbb{E}_{\mathbf{z}\sim q_z}\left[\log\left(\frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})}\right)\right] \\
&= \sum_{\mathbf{z}\in\mathcal{Z}} q(\mathbf{z}) \log\left(\frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})}\right) \\
&= \sum_{\mathbf{z}\in\mathcal{Z}} q(\mathbf{z}) \log\left(q(\mathbf{z})\right) - \sum_{\mathbf{z}\in\mathcal{Z}} q(\mathbf{z}) \log\left(p(\mathbf{z}|\mathbf{x})\right) \\
&= -\mathcal{H}(q) - \sum_{\mathbf{z}\in\mathcal{Z}} q(\mathbf{z}) \log\left(p(\mathbf{z}|\mathbf{x})\right) \\
&= -\mathcal{H}(q) - \sum_{\mathbf{z}\in\mathcal{Z}} q(\mathbf{z}) \log\left(\frac{p(\mathbf{z},\mathbf{x})}{p(\mathbf{x})}\right) \\
&= -\mathcal{H}(q) - \sum_{\mathbf{z}\in\mathcal{Z}} q(\mathbf{z}) \log\left(p(\mathbf{z},\mathbf{x})\right) + \sum_{\mathbf{z}\in\mathcal{Z}} q(\mathbf{z}) \log\left(p(\mathbf{x})\right) \\
&= -\mathcal{H}(q) - \mathbb{E}_{\mathbf{z}\sim q_{\mathbf{z}}}\left[\log\left(p(\mathbf{z},\mathbf{x})\right)\right] + \log\left(p(\mathbf{x})\right)
\end{aligned}
$$

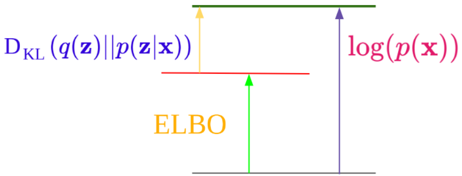
Generative Process

$$
\begin{aligned}
D_{\mathsf{KL}}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})) &= -\mathcal{H}(q) - \mathbb{E}_{\mathbf{z}\sim q_{\mathbf{z}}}\left[\log\left(p(\mathbf{z},\mathbf{x})\right)\right] + \log\left(p(\mathbf{x})\right) \\
\log\left(p(\mathbf{x})\right) &= \mathbb{E}_{\mathbf{z}\sim q_{\mathbf{z}}}\left[\log\left(p(\mathbf{z},\mathbf{x})\right)\right] + \mathcal{H}(q) + D_{\mathsf{KL}}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})) \\
&= \mathsf{ELBO} + D_{\mathsf{KL}}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})) \\
\log p(\mathbf{x}) &\geq \mathsf{ELBO} \\
\log p(\mathbf{x}) &\stackrel{?}{=} \mathsf{ELBO} \\
D_{\mathsf{KL}}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})) &= 0 \\
q(\mathbf{z}) &= p(\mathbf{z}|\mathbf{x}).
\end{aligned}
$$

Therefore, closer $q(z)$ is to $p(\mathbf{z}|\mathbf{x})$, the closer the ELBO is to the true log-likelihood. What if the posterior $p(\mathbf{z}|\mathbf{x})$ is intractable to compute? Suppose $q_{\boldsymbol{\phi}}(\mathbf{z})$ is a (tractable) probability distribution over the hidden variables $\mathbf{z}$ parameterized by ϕ (variational parameters). Pick ϕ so that $q_{\boldsymbol{\phi}}(\mathbf{z})$ is as close as possible to $p(\mathbf{z}|\mathbf{x})$.

Variational Auto-encoders

$$
\begin{aligned}
\log\left(p(\mathbf{x})\right) &= \text{ELBO} + D_{\text{KL}}(q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x})) \\
\text{ELBO} &= \mathbb{E}_{\mathbf{z}\sim q_z}\left[\log\left(\frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right)\right] = \mathbb{E}_{\mathbf{z}\sim q_z}\left[\log\left(p(\mathbf{x}|\mathbf{z})\right)\right] \\
\log p(\mathbf{x}) &\geq \text{ELBO} \\
\log p(\mathbf{x}) &\stackrel{?}{=} \text{ELBO} \\
D_{\text{KL}}(q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x})) &= 0 \Rightarrow q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}).
\end{aligned}
$$



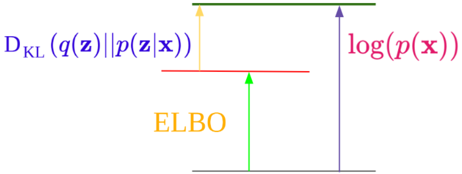
Since $\log\left(p(\mathbf{x})\right)$ is intractable we maximize the ELBO. To make the bound tighter, we should make $q(z)$ as close as possible to $p(\mathbf{z}|\mathbf{x})$, i.e., minimize $D_{\text{KL}}(q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x}))$.

Variational Auto-encoders

$$\begin{aligned}\log\left(p(\mathbf{x})\right) &= \text{ELBO} + D_{\text{KL}}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))\\ \text{ELBO} &= \mathbb{E}_{\mathbf{z}\sim q_z}\left[\log\left(p(\mathbf{x}|\mathbf{z})\right)\right]\end{aligned}$$

Maximize ELBO such that $q(z)$ is as close as possible to $p(\mathbf{z}|\mathbf{x})$. This can be posed as an optimization problem as below.

$$\max \mathbb{E}_{\mathbf{z}\sim q_z}\left[\log\left(p(\mathbf{x}|\mathbf{z})\right)\right] \text{ such that } q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}).$$



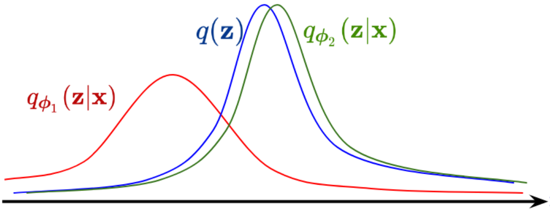
Variational Auto-encoders

Maximize ELBO such that $q(z)$ is as close as possible to $p(\mathbf{z}|\mathbf{x})$. This can be posed as an optimization problem as below.

$$\max \mathbb{E}_{\mathbf{z}\sim q_z}\left[\log\left(p(\mathbf{x}|\mathbf{z})\right)\right] \text{ such that } q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}).$$

Variational Inference

What if the posterior $p(\mathbf{z}|\mathbf{x})$ is intractable? (This can be seen using Bayes theorem $p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}$). Suppose $q_{\phi}(\mathbf{z}|\mathbf{x})$ is a tractable probability distribution over the hidden variables $\mathbf{z}$ parameterized by ϕ (variational parameters). Then, pick ϕ so that $q_{\phi}(\mathbf{z}|\mathbf{x})$ is as close as possible to $q(\mathbf{z})$.



Realization of Variational Auto-encoders using NNs

So far we have considered only abstract representations of $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, and $q(\mathbf{z})$. Let us assume that $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{z}})$, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{z}|\mathbf{x}})$ and $q(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.