Deep Generative Models

Lecture 3

Roman Isachenko

Moscow Institute of Physics and Technology
Yandex School of Data Analysis

2025, Autumn

Recap of Previous Lecture

Jacobian Matrix

Given a differentiable function $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^m$,

$$\mathbf{z} = \mathbf{f}(\mathbf{x}), \quad \mathbf{J} = \frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial z_1}{\partial x_1} & \cdots & \frac{\partial z_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_m}{\partial x_1} & \cdots & \frac{\partial z_m}{\partial x_m} \end{pmatrix} \in \mathbb{R}^{m \times m}$$

Change of Variables Theorem (CoV)

Let $\mathbf{x} \in \mathbb{R}^m$ be a random vector with density $p(\mathbf{x})$, and let $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^m$ be a C^1-diffeomorphism ($\mathbf{f}$ and $\mathbf{f}^{-1}$ are continuously differentiable mappings). If $\mathbf{z} = \mathbf{f}(\mathbf{x})$, then

$$p(\mathbf{x}) = p(\mathbf{z}) |\det(\mathbf{J}_\mathbf{f})| = p(\mathbf{z}) \left| \det \left(\frac{\partial \mathbf{z}}{\partial \mathbf{x}} \right) \right| = p(\mathbf{f}(\mathbf{x})) \left| \det \left(\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \right) \right|$$

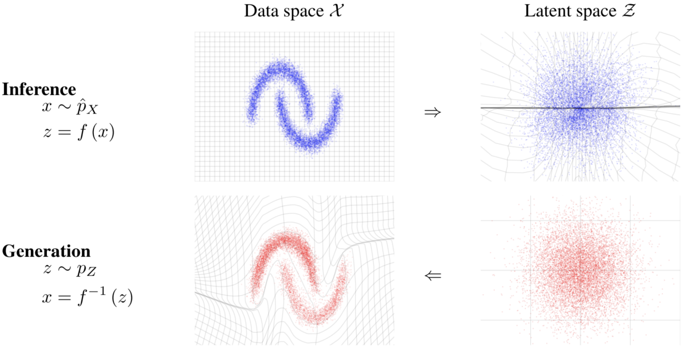
$$p(\mathbf{z}) = p(\mathbf{x}) |\det(\mathbf{J}_{\mathbf{f}^{-1}})| = p(\mathbf{x}) \left| \det \left(\frac{\partial \mathbf{x}}{\partial \mathbf{z}} \right) \right| = p(\mathbf{f}^{-1}(\mathbf{z})) \left| \det \left(\frac{\partial \mathbf{f}^{-1}(\mathbf{z})}{\partial \mathbf{z}} \right) \right|$$

Dinh L., Sohl-Dickstein J., Bengio S. Density Estimation Using Real NVP, 2016

Recap of Previous Lecture

Definition

A normalizing flow is a *differentiable*, *invertible* transformation that maps data $\mathbf{x}$ to noise $\mathbf{z}$.



Log-Likelihood

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log p(\mathbf{f}_K \circ \cdots \circ \mathbf{f}_1(\mathbf{x})) + \sum_{k=1}^{K} \log |\det(\mathbf{J}_{\mathbf{f}_k})|$$

Dinh L., Sohl-Dickstein J., Bengio S. Density Estimation Using Real NVP, 2016

Recap of Previous Lecture

Flow Log-Likelihood

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log p(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})) + \log |\det(\mathbf{J}_\mathbf{f})|$$

One significant challenge is efficiently computing the Jacobian determinant.

Linear Flows

$$\mathbf{z} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{W}\mathbf{x}, \quad \mathbf{W} \in \mathbb{R}^{m \times m}, \quad \boldsymbol{\theta} = \mathbf{W}, \quad \mathbf{J}_\mathbf{f} = \mathbf{W}^T$$

- LU Decomposition:
$$\mathbf{W} = \mathbf{P}\mathbf{L}\mathbf{U}.$$
- QR Decomposition:
$$\mathbf{W} = \mathbf{Q}\mathbf{R}.$$

Decomposition is performed only once during initialization. Then the decomposed matrices ($\mathbf{P}, \mathbf{L}, \mathbf{U}$ or $\mathbf{Q}, \mathbf{R}$) are optimized.

Kingma D. P., et al. Glow: Generative Flow With Invertible 1x1 Convolutions, 2018
Hoogeboom E., et al. Emerging Convolutions for Generative Normalizing Flows, 2019

Recap of Previous Lecture

Consider an autoregressive model:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{j=1}^{m} p_{\boldsymbol{\theta}}(x_j|\mathbf{x}_{1:j-1}), \quad p_{\boldsymbol{\theta}}(x_j|\mathbf{x}_{1:j-1}) = \mathcal{N}\left(\mu_{j,\boldsymbol{\theta}}(\mathbf{x}_{1:j-1}), \sigma^2_{j,\boldsymbol{\theta}}(\mathbf{x}_{1:j-1})\right).$$

Gaussian Autoregressive Normalizing Flow

$$\mathbf{x} = \mathbf{f}^{-1}_{\boldsymbol{\theta}}(\mathbf{z}) \quad \Rightarrow \quad x_j = \sigma_{j,\boldsymbol{\theta}}(\mathbf{x}_{1:j-1}) \cdot z_j + \mu_{j,\boldsymbol{\theta}}(\mathbf{x}_{1:j-1}).$$

$$\mathbf{z} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}) \quad \Rightarrow \quad z_j = (x_j - \mu_{j,\boldsymbol{\theta}}(\mathbf{x}_{1:j-1})) \cdot \frac{1}{\sigma_{j,\boldsymbol{\theta}}(\mathbf{x}_{1:j-1})}.$$

- This transformation is both **invertible** and **differentiable**, mapping $p(\mathbf{z})$ to $p_{\boldsymbol{\theta}}(\mathbf{x})$.
- The Jacobian matrix for this transformation is triangular.

The generative function $\mathbf{f}^{-1}_{\boldsymbol{\theta}}(\mathbf{z})$ is **sequential**, while the inference function $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$ is **not sequential**.

Papamakarios G., Pavlakou T., Murray I. Masked Autoregressive Flow for Density Estimation, 2017

Recap of Previous Lecture

Let us partition $\mathbf{x}$ and $\mathbf{z}$ into two groups:

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2] = [\mathbf{x}_{1:d}, \mathbf{x}_{d+1:m}]; \quad \mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2] = [\mathbf{z}_{1:d}, \mathbf{z}_{d+1:m}].$$

Coupling Layer

$$\begin{cases} \mathbf{x}_1 = \mathbf{z}_1; \\ \mathbf{x}_2 = \mathbf{z}_2 \odot \boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{z}_1) + \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_1). \end{cases} \quad \begin{cases} \mathbf{z}_1 = \mathbf{x}_1; \\ \mathbf{z}_2 = (\mathbf{x}_2 - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}_1)) \odot \frac{1}{\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x}_1)}. \end{cases}$$

Both density estimation and sampling require just a single pass!

Jacobian

$$\det \left(\frac{\partial \mathbf{z}}{\partial \mathbf{x}} \right) = \det \begin{pmatrix} \mathbf{I}_d & 0_{d \times (m-d)} \\ \frac{\partial \mathbf{z}_2}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{z}_2}{\partial \mathbf{x}_2} \end{pmatrix} = \prod_{j=1}^{m-d} \frac{1}{\sigma_{j,\boldsymbol{\theta}}(\mathbf{x}_1)}.$$

A coupling layer is a special instance of an gaussian autoregressive normalizing flow.

Dinh L., Sohl-Dickstein J., Bengio S. Density Estimation Using Real NVP, 2016

Outline

1. Latent Variable Models (LVM)
2. Variational Evidence Lower Bound (ELBO)
3. EM-Algorithm
4. Amortized Inference

Outline

1. Latent Variable Models (LVM)
2. Variational Evidence Lower Bound (ELBO)
3. EM-Algorithm
4. Amortized Inference

Bayesian Framework

Bayes' Theorem

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

- $\mathbf{x}$: observed variables;
- $\boldsymbol{\theta}$: unknown latent variables/parameters;
- $p_{\boldsymbol{\theta}}(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta})$: likelihood;
- $p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$: evidence;
- $p(\boldsymbol{\theta})$: prior distribution;
- $p(\boldsymbol{\theta}|\mathbf{x})$: posterior distribution.

Bayesian Framework

Bayes' Theorem

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

- $\mathbf{x}$: observed variables;
- $\boldsymbol{\theta}$: unknown latent variables/parameters;
- $p_{\boldsymbol{\theta}}(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta})$: likelihood;
- $p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$: evidence;
- $p(\boldsymbol{\theta})$: prior distribution;
- $p(\boldsymbol{\theta}|\mathbf{x})$: posterior distribution.

Interpretation

- We begin with unknown variables $\boldsymbol{\theta}$ and a prior belief $p(\boldsymbol{\theta})$.
- Once data $\mathbf{x}$ is observed, the posterior $p(\boldsymbol{\theta}|\mathbf{x})$ incorporates both prior beliefs and evidence from the data.

Bayesian Framework

Consider the case where the unobserved variables $\boldsymbol{\theta}$ are model parameters (i.e., $\boldsymbol{\theta}$ are random variables).

- $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$: observed samples;
- $p(\boldsymbol{\theta})$: prior distribution.

Bayesian Framework

Consider the case where the unobserved variables $\boldsymbol{\theta}$ are model parameters (i.e., $\boldsymbol{\theta}$ are random variables).

- $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$: observed samples;
- $p(\boldsymbol{\theta})$: prior distribution.

Posterior Distribution

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

Bayesian Framework

Consider the case where the unobserved variables $\boldsymbol{\theta}$ are model parameters (i.e., $\boldsymbol{\theta}$ are random variables).

- $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$: observed samples;
- $p(\boldsymbol{\theta})$: prior distribution.

Posterior Distribution

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

If the evidence $p(\mathbf{X})$ is intractable (due to high-dimensional integration), the posterior cannot be computed exactly.

Bayesian Framework

Consider the case where the unobserved variables $\boldsymbol{\theta}$ are model parameters (i.e., $\boldsymbol{\theta}$ are random variables).

- $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n}$: observed samples;
- $p(\boldsymbol{\theta})$: prior distribution.

Posterior Distribution

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

If the evidence $p(\mathbf{X})$ is intractable (due to high-dimensional integration), the posterior cannot be computed exactly.

Maximum a Posteriori (MAP) Estimation

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}) = \arg\max_{\boldsymbol{\theta}} (\log p(\mathbf{X}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}))$$

Latent Variable Models (LVM)

Maximum Likelihood Extimation (MLE) Problem

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{X}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{x}_i) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i).$$

Latent Variable Models (LVM)

Maximum Likelihood Extimation (MLE) Problem

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{X}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{x}_i) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i).$$

The distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ can be highly complex and often intractable (just like the true data distribution $p_{\text{data}}(\mathbf{x})$).

Latent Variable Models (LVM)

Maximum Likelihood Extimation (MLE) Problem

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{X}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{x}_i) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i).$$

The distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ can be highly complex and often intractable (just like the true data distribution $p_{\text{data}}(\mathbf{x})$).

Extended Probabilistic Model

Introduce a latent variable $\mathbf{z}$ for each observed sample $\mathbf{x}$:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z}); \quad \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}).$$

Latent Variable Models (LVM)

Maximum Likelihood Extimation (MLE) Problem

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{X}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{x}_i) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i).$$

The distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ can be highly complex and often intractable (just like the true data distribution $p_{\text{data}}(\mathbf{x})$).

Extended Probabilistic Model

Introduce a latent variable $\mathbf{z}$ for each observed sample $\mathbf{x}$:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z}); \quad \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}).$$

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z}) d\mathbf{z}.$$

Latent Variable Models (LVM)

Maximum Likelihood Extimation (MLE) Problem

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{X}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{x}_i) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i).$$

The distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ can be highly complex and often intractable (just like the true data distribution $p_{\text{data}}(\mathbf{x})$).

Extended Probabilistic Model

Introduce a latent variable $\mathbf{z}$ for each observed sample $\mathbf{x}$:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z}); \quad \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}).$$

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z}) d\mathbf{z}.$$

Motivation

Both $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ and $p(\mathbf{z})$ are usually much simpler than $p_{\boldsymbol{\theta}}(\mathbf{x})$.

Latent Variable Models (LVM)

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} \rightarrow \max_{\boldsymbol{\theta}}$$

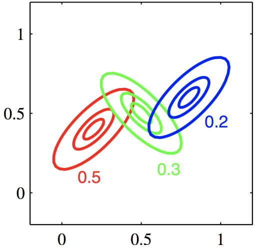
Bishop C. Pattern Recognition and Machine Learning, 2006

Latent Variable Models (LVM)

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} \to \max_{\boldsymbol{\theta}}$$

Examples

Mixture of Gaussians



- $p_{\boldsymbol{\theta}}(\mathbf{x}|z) = \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$
- $p(z) = \text{Categorical}(\boldsymbol{\pi})$

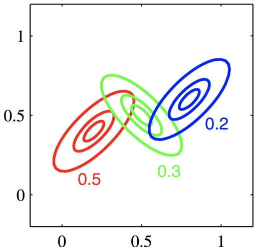
Bishop C. Pattern Recognition and Machine Learning, 2006

Latent Variable Models (LVM)

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} \to \max_{\boldsymbol{\theta}}$$

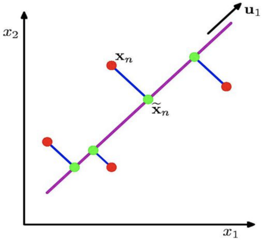
Examples

Mixture of Gaussians



- $p_{\boldsymbol{\theta}}(\mathbf{x}|z) = \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$
- $p(z) = \mathrm{Categorical}(\boldsymbol{\pi})$

PCA Model



- $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$
- $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$

Bishop C. Pattern Recognition and Machine Learning, 2006

MLE for LVM

$$\sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i) = \sum_{i=1}^{n} \log \int p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_i)p(\mathbf{z}_i)d\mathbf{z}_i \rightarrow \max_{\boldsymbol{\theta}}.$$

image credit: https://jmtomczak.github.io/blog/4/4_VAE.html

MLE for LVM

$$\sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i) = \sum_{i=1}^{n} \log \int p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_i)p(\mathbf{z}_i)d\mathbf{z}_i \rightarrow \max_{\boldsymbol{\theta}}.$$

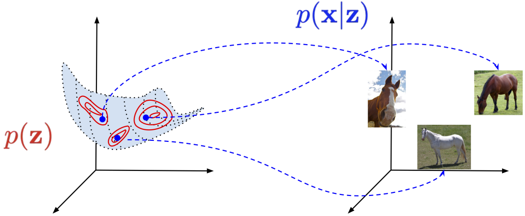
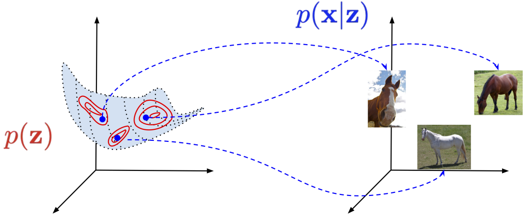


image credit: https://jmtomczak.github.io/blog/4/4_VAE.html

MLE for LVM

$$\sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i) = \sum_{i=1}^{n} \log \int p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_i)p(\mathbf{z}_i)d\mathbf{z}_i \to \max_{\boldsymbol{\theta}}.$$



Naive Approach

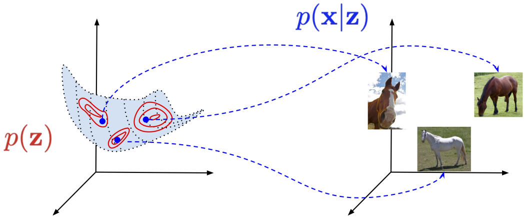
$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \mathbb{E}_{p(\mathbf{z})}p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) \approx \frac{1}{K}\sum_{k=1}^{K} p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}_k),$$

where $\mathbf{z}_k \sim p(\mathbf{z})$.

image credit: https://jmtomczak.github.io/blog/4/4_VAE.html

MLE for LVM

$$\sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i) = \sum_{i=1}^{n} \log \int p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_i)p(\mathbf{z}_i)d\mathbf{z}_i \to \max_{\boldsymbol{\theta}}.$$



Naive Approach

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \mathbb{E}_{p(\mathbf{z})} p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) \approx \frac{1}{K}\sum_{k=1}^{K} p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}_k),$$

where $\mathbf{z}_k \sim p(\mathbf{z})$.

Challenge: As the dimensionality of $\mathbf{z}$ increases, the number of samples needed to adequately cover the latent space grows exponentially.

image credit: https://jmtomczak.github.io/blog/4/4_VAE.html

Outline

ELBO Derivation I

Inequality Derivation

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

ELBO Derivation I

Inequality Derivation

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int \frac{q(\mathbf{z})}{q(\mathbf{z})} p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

ELBO Derivation I

Inequality Derivation

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int \frac{q(\mathbf{z})}{q(\mathbf{z})} p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} =$$

$$= \log \mathbb{E}_q \left[\frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right]$$

ELBO Derivation I

Inequality Derivation

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int \frac{q(\mathbf{z})}{q(\mathbf{z})} p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} =$$

$$= \log \mathbb{E}_q \left[\frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right] \geq \mathbb{E}_q \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} = \mathcal{L}_{q, \boldsymbol{\theta}}(\mathbf{x})$$

ELBO Derivation I

Inequality Derivation

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int \frac{q(\mathbf{z})}{q(\mathbf{z})} p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} =$$

$$= \log \mathbb{E}_q \left[\frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right] \geq \mathbb{E}_q \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} = \mathcal{L}_{q, \boldsymbol{\theta}}(\mathbf{x})$$

Here, $q(\mathbf{z})$ is any distribution such that $\int q(\mathbf{z}) d\mathbf{z} = 1$.

ELBO Derivation I

Inequality Derivation

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int \frac{q(\mathbf{z})}{q(\mathbf{z})} p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} =$$

$$= \log \mathbb{E}_q \left[\frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right] \geq \mathbb{E}_q \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} = \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$$

Here, $q(\mathbf{z})$ is any distribution such that $\int q(\mathbf{z}) d\mathbf{z} = 1$.

Variational Evidence Lower Bound (ELBO)

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_q \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \leq \log p_{\boldsymbol{\theta}}(\mathbf{x})$$

ELBO Derivation I

Inequality Derivation

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int \frac{q(\mathbf{z})}{q(\mathbf{z})} p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} =$$

$$= \log \mathbb{E}_q \left[\frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right] \geq \mathbb{E}_q \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} = \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$$

Here, $q(\mathbf{z})$ is any distribution such that $\int q(\mathbf{z}) d\mathbf{z} = 1$.

Variational Evidence Lower Bound (ELBO)

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_q \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \leq \log p_{\boldsymbol{\theta}}(\mathbf{x})$$

This inequality holds for any choice of $q(\mathbf{z})$.

ELBO Derivation II

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})}$$

Equality Derivation

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z}$$

ELBO Derivation II

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})}$$

Equality Derivation

$$\begin{aligned}\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} = \\ = \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) p_{\boldsymbol{\theta}}(\mathbf{x})}{q(\mathbf{z})} d\mathbf{z}\end{aligned}$$

ELBO Derivation II

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})}$$

Equality Derivation

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} =$$

$$= \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) p_{\boldsymbol{\theta}}(\mathbf{x})}{q(\mathbf{z})} d\mathbf{z} =$$

$$= \int q(\mathbf{z}) \log p_{\boldsymbol{\theta}}(\mathbf{x}) d\mathbf{z} + \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})} d\mathbf{z}$$

ELBO Derivation II

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})}$$

Equality Derivation

$$\begin{aligned}
\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} = \\
= \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) p_{\boldsymbol{\theta}}(\mathbf{x})}{q(\mathbf{z})} d\mathbf{z} = \\
= \int q(\mathbf{z}) \log p_{\boldsymbol{\theta}}(\mathbf{x}) d\mathbf{z} + \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})} d\mathbf{z} = \\
= \log p_{\boldsymbol{\theta}}(\mathbf{x}) - \mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))
\end{aligned}$$

ELBO Derivation II

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})}$$

Equality Derivation

$$\begin{aligned}
\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} = \\
= \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) p_{\boldsymbol{\theta}}(\mathbf{x})}{q(\mathbf{z})} d\mathbf{z} = \\
= \int q(\mathbf{z}) \log p_{\boldsymbol{\theta}}(\mathbf{x}) d\mathbf{z} + \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})} d\mathbf{z} = \\
= \log p_{\boldsymbol{\theta}}(\mathbf{x}) - \mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))
\end{aligned}$$

Variational Decomposition

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) + \mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) \geq \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}).$$

ELBO Derivation II

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})}$$

Equality Derivation

$$\begin{aligned}
\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} = \\
= \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) p_{\boldsymbol{\theta}}(\mathbf{x})}{q(\mathbf{z})} d\mathbf{z} = \\
= \int q(\mathbf{z}) \log p_{\boldsymbol{\theta}}(\mathbf{x}) d\mathbf{z} + \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})} d\mathbf{z} = \\
= \log p_{\boldsymbol{\theta}}(\mathbf{x}) - \mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))
\end{aligned}$$

Variational Decomposition

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) + \mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) \geq \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}).$$

Here, $\mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) \geq 0$.

Variational Evidence Lower Bound (ELBO)

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z}$$

Variational Evidence Lower Bound (ELBO)

$$\begin{aligned}\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) &= \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\ &= \int q(\mathbf{z}) \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) d\mathbf{z} + \int q(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z}\end{aligned}$$

Variational Evidence Lower Bound (ELBO)

$$
\begin{aligned}
\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) &= \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\
&= \int q(\mathbf{z}) \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) d\mathbf{z} + \int q(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\
&= \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z}))
\end{aligned}
$$

Variational Evidence Lower Bound (ELBO)

$$\begin{aligned}
\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) &= \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\
&= \int q(\mathbf{z}) \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) d\mathbf{z} + \int q(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\
&= \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z}))
\end{aligned}$$

Log-Likelihood Decomposition

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) + \mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$

Variational Evidence Lower Bound (ELBO)

$$
\begin{aligned}
\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) &= \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\
&= \int q(\mathbf{z}) \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) d\mathbf{z} + \int q(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\
&= \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z}))
\end{aligned}
$$

Log-Likelihood Decomposition

$$
\begin{aligned}
\log p_{\boldsymbol{\theta}}(\mathbf{x}) &= \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) + \mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) = \\
&= \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z})) + \mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})).
\end{aligned}
$$

Variational Evidence Lower Bound (ELBO)

$$
\begin{aligned}
\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) &= \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\
&= \int q(\mathbf{z}) \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) d\mathbf{z} + \int q(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\
&= \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z}))
\end{aligned}
$$

Log-Likelihood Decomposition

$$
\begin{aligned}
\log p_{\boldsymbol{\theta}}(\mathbf{x}) &= \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) + \mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) = \\
&= \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z})) + \mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})).
\end{aligned}
$$

- Instead of maximizing the likelihood, maximize the ELBO:

$$
\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{x}) \quad \rightarrow \quad \max_{q,\boldsymbol{\theta}} \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})
$$

Variational Evidence Lower Bound (ELBO)

$$\begin{aligned}\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) &= \int q(\mathbf{z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\ &= \int q(\mathbf{z}) \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) d\mathbf{z} + \int q(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\ &= \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z}))\end{aligned}$$

Log-Likelihood Decomposition

$$\begin{aligned}\log p_{\boldsymbol{\theta}}(\mathbf{x}) &= \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) + \mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) = \\ &= \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z})) + \mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})).\end{aligned}$$

- Instead of maximizing the likelihood, maximize the ELBO:

$$\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{x}) \quad \rightarrow \quad \max_{q,\boldsymbol{\theta}} \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$$

- Maximizing the ELBO with respect to the **variational** distribution q is equivalent to minimizing the KL divergence:

$$\arg\max_{q} \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) \equiv \arg\min_{q} \mathrm{KL}(q(\mathbf{z}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})).$$

Outline

EM-Algorithm

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z})) =$$

$$= \mathbb{E}_q \left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \log \frac{q(\mathbf{z})}{p(\mathbf{z})} \right] d\mathbf{z} \to \max_{q,\boldsymbol{\theta}}.$$

EM-Algorithm

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z})) =$$
$$= \mathbb{E}_q \left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \log \frac{q(\mathbf{z})}{p(\mathbf{z})} \right] d\mathbf{z} \rightarrow \max_{q,\boldsymbol{\theta}}.$$

Block-Coordinate Optimization

- Initialize $\boldsymbol{\theta}^*$;

EM-Algorithm

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z})) =$$

$$= \mathbb{E}_q \left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \log \frac{q(\mathbf{z})}{p(\mathbf{z})} \right] d\mathbf{z} \to \max_{q,\boldsymbol{\theta}} .$$

Block-Coordinate Optimization

- Initialize $\boldsymbol{\theta}^*$;
- **E-step** (optimize $\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$ over q):

$$q^*(\mathbf{z}) = \arg\max_q \mathcal{L}_{q,\boldsymbol{\theta}^*}(\mathbf{x}) =$$

$$= \arg\min_q \mathrm{KL}(q(\mathbf{z})\|p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x});$$

EM-Algorithm

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z})) =$$
$$= \mathbb{E}_q \left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \log \frac{q(\mathbf{z})}{p(\mathbf{z})} \right] d\mathbf{z} \to \max_{q,\boldsymbol{\theta}} .$$

Block-Coordinate Optimization

- Initialize $\boldsymbol{\theta}^*$;
- **E-step** (optimize $\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$ over q):
  $$q^*(\mathbf{z}) = \arg\max_q \mathcal{L}_{q,\boldsymbol{\theta}^*}(\mathbf{x}) =$$
  $$= \arg\min_q \mathrm{KL}(q(\mathbf{z})\|p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x});$$
- **M-step** (optimize $\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$ over $\boldsymbol{\theta}$):
  $$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathcal{L}_{q^*,\boldsymbol{\theta}}(\mathbf{x});$$

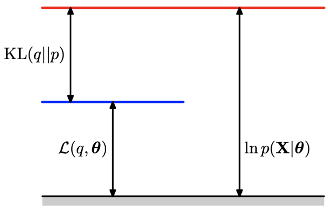
EM-Algorithm

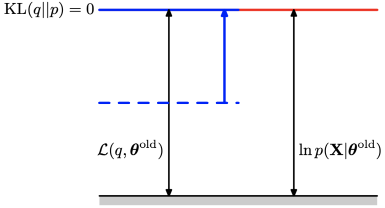
$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z})) =$$
$$= \mathbb{E}_q \left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \log \frac{q(\mathbf{z})}{p(\mathbf{z})} \right] d\mathbf{z} \to \max_{q,\boldsymbol{\theta}}.$$

Block-Coordinate Optimization

- Initialize $\boldsymbol{\theta}^*$;
- **E-step** (optimize $\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$ over q):
  $$q^*(\mathbf{z}) = \arg\max_q \mathcal{L}_{q,\boldsymbol{\theta}^*}(\mathbf{x}) =$$
  $$= \arg\min_q \mathrm{KL}(q(\mathbf{z})\|p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x});$$
- **M-step** (optimize $\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$ over $\boldsymbol{\theta}$):
  $$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathcal{L}_{q^*,\boldsymbol{\theta}}(\mathbf{x});$$
- Repeat the E-step and M-step until convergence.

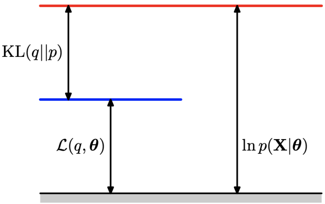
EM-Algorithm Illustration

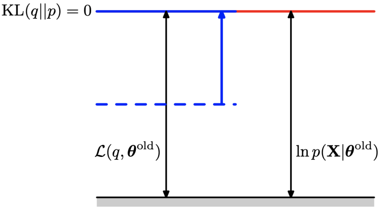


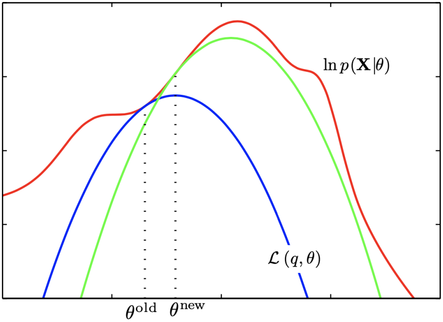


Bishop C. Pattern Recognition and Machine Learning, 2006

EM-Algorithm Illustration







Bishop C. Pattern Recognition and Machine Learning, 2006

Outline

Amortized Variational Inference

E-step

$$q(\mathbf{z}) = \arg\max_q \mathcal{L}_{q,\boldsymbol{\theta}^*}(\mathbf{x}) = \arg\min_q \mathrm{KL}(q\|p) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x}).$$

Amortized Variational Inference

E-step

$$q(\mathbf{z}) = \arg\max_q \mathcal{L}_{q,\boldsymbol{\theta}^*}(\mathbf{x}) = \arg\min_q \mathrm{KL}(q\|p) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x}).$$

$q(\mathbf{z})$ approximates the true posterior $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$, hence it is called **variational posterior**.

Amortized Variational Inference

E-step

$$q(\mathbf{z}) = \arg\max_q \mathcal{L}_{q,\boldsymbol{\theta}^*}(\mathbf{x}) = \arg\min_q \mathrm{KL}(q \| p) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x}).$$

$q(\mathbf{z})$ approximates the true posterior $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$, hence it is called **variational posterior**.

- $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ may be **intractable**;
- $q(\mathbf{z})$ is individual for each data point $\mathbf{x}$.

Amortized Variational Inference

E-step

$$q(\mathbf{z}) = \arg\max_q \mathcal{L}_{q,\boldsymbol{\theta}^*}(\mathbf{x}) = \arg\min_q \mathrm{KL}(q \| p) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x}).$$

$q(\mathbf{z})$ approximates the true posterior $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$, hence it is called **variational posterior**.

- $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ may be **intractable**;
- $q(\mathbf{z})$ is individual for each data point $\mathbf{x}$.

Variational Bayes

We restrict the family of possible distributions $q(\mathbf{z})$ to a parametric class $q_{\phi}(\mathbf{z}|\mathbf{x})$, conditioned on data $\mathbf{x}$ and parameterized by ϕ.

Amortized Variational Inference

E-step

$$q(\mathbf{z}) = \arg\max_q \mathcal{L}_{q,\boldsymbol{\theta}^*}(\mathbf{x}) = \arg\min_q \mathrm{KL}(q\|p) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x}).$$

$q(\mathbf{z})$ approximates the true posterior $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$, hence it is called **variational posterior**.

- $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ may be **intractable**;
- $q(\mathbf{z})$ is individual for each data point $\mathbf{x}$.

Variational Bayes

We restrict the family of possible distributions $q(\mathbf{z})$ to a parametric class $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, conditioned on data $\mathbf{x}$ and parameterized by $\boldsymbol{\phi}$.

- E-step

$$\boldsymbol{\phi}_k = \boldsymbol{\phi}_{k-1} + \eta \cdot \nabla_{\boldsymbol{\phi}} \mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\theta}_{k-1}}(\mathbf{x})\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_{k-1}}$$

- M-step

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \eta \cdot \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\phi}_k,\boldsymbol{\theta}}(\mathbf{x})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{k-1}}$$

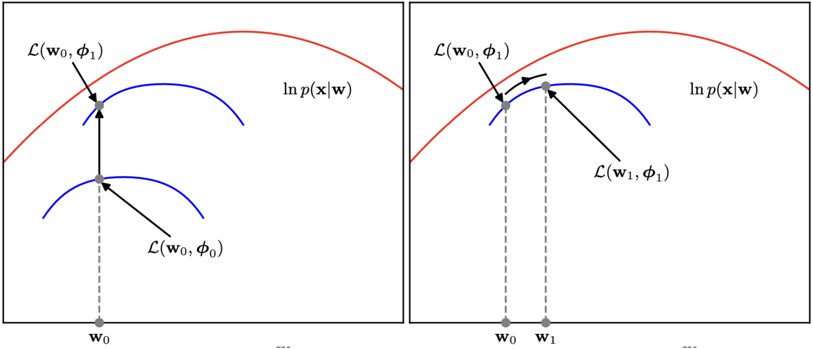
Variational EM Illustration

- E-step:

$$\phi_k = \phi_{k-1} + \eta \cdot \nabla_{\phi} \mathcal{L}_{\phi, \boldsymbol{\theta}_{k-1}}(\mathbf{x})\Big|_{\phi=\phi_{k-1}}$$

- M-step:

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \eta \cdot \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\phi_k, \boldsymbol{\theta}}(\mathbf{x})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{k-1}}$$



Bishop C., Deep Learning: Foundations and Concepts, 2024

Variational EM Algorithm

ELBO

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{x}) + \mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) \geq \mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{x}).$$

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))$$

Variational EM Algorithm

ELBO

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{x}) + \mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) \geq \mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{x}).$$

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_q \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))$$

- **E-step:**

$$\boldsymbol{\phi}_k = \boldsymbol{\phi}_{k-1} + \eta \cdot \nabla_{\boldsymbol{\phi}} \mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\theta}_{k-1}}(\mathbf{x})\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_{k-1}},$$

  where ϕ denotes the parameters of the variational posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$.

- **M-step:**

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \eta \cdot \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\phi}_k,\boldsymbol{\theta}}(\mathbf{x})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{k-1}},$$

  where $\boldsymbol{\theta}$ represents the parameters of the generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$.

Variational EM Algorithm

ELBO

$$\log p_{\boldsymbol\theta}(\mathbf{x}) = \mathcal{L}_{\boldsymbol\phi,\boldsymbol\theta}(\mathbf{x}) + \mathrm{KL}(q_{\boldsymbol\phi}(\mathbf{z}|\mathbf{x}) \| p_{\boldsymbol\theta}(\mathbf{z}|\mathbf{x})) \geq \mathcal{L}_{\boldsymbol\phi,\boldsymbol\theta}(\mathbf{x}).$$

$$\mathcal{L}_{q,\boldsymbol\theta}(\mathbf{x}) = \mathbb{E}_q \log p_{\boldsymbol\theta}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q_{\boldsymbol\phi}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))$$

- **E-step:**

$$\boldsymbol\phi_k = \boldsymbol\phi_{k-1} + \eta \cdot \nabla_{\boldsymbol\phi} \mathcal{L}_{\boldsymbol\phi,\boldsymbol\theta_{k-1}}(\mathbf{x})\Big|_{\boldsymbol\phi=\boldsymbol\phi_{k-1}},$$

  where $\boldsymbol\phi$ denotes the parameters of the variational posterior $q_{\boldsymbol\phi}(\mathbf{z}|\mathbf{x})$.

- **M-step:**

$$\boldsymbol\theta_k = \boldsymbol\theta_{k-1} + \eta \cdot \nabla_{\boldsymbol\theta} \mathcal{L}_{\boldsymbol\phi_k,\boldsymbol\theta}(\mathbf{x})\Big|_{\boldsymbol\theta=\boldsymbol\theta_{k-1}},$$

  where $\boldsymbol\theta$ represents the parameters of the generative model $p_{\boldsymbol\theta}(\mathbf{x}|\mathbf{z})$.

The remaining step is to obtain **unbiased** Monte Carlo estimates of the gradients: $\nabla_{\boldsymbol\phi} \mathcal{L}_{\boldsymbol\phi,\boldsymbol\theta}(\mathbf{x})$ and $\nabla_{\boldsymbol\theta} \mathcal{L}_{\boldsymbol\phi,\boldsymbol\theta}(\mathbf{x})$.

Summary

- The Bayesian framework generalizes nearly all standard machine learning methods.
- LVMs introduce latent representations for observed data, enabling more interpretable models.
- LVMs maximize the variational evidence lower bound (ELBO) to obtain maximum likelihood estimates for the parameters.
- The general variational EM algorithm optimizes the ELBO within LVMs to recover the MLE for the parameters $\boldsymbol{\theta}$.
- Amortized variational inference enables efficient estimation of the ELBO via Monte Carlo estimation.