Lecture 4: Variational AutoEncoder

Deep Generative Models

Sajjad Amini

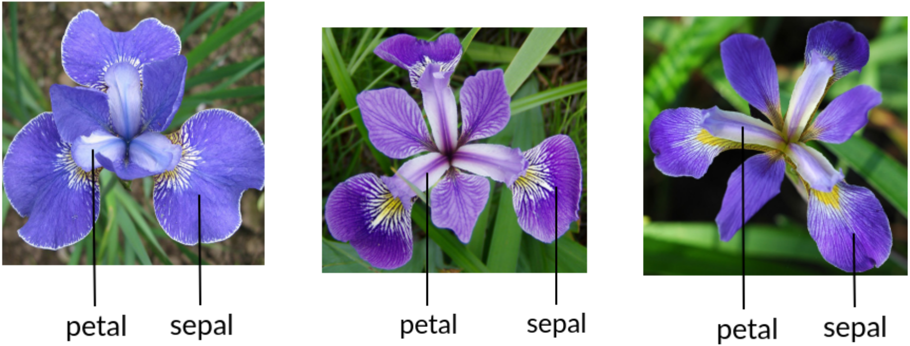
Department of Electrical Engineering
Sharif University of Technology

Contents

Section 1

Guassian Mixture Model

Example - Iris Flower



(a) Setosa (b) Versicolor (c) Virginica

Figure: Different types of Iris flower

Example - Iris Flower Data set

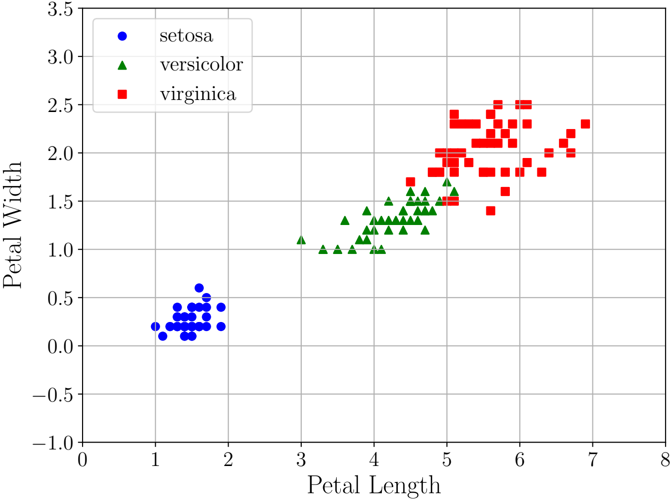


Figure: Labeled Iris dataset (in practive, you don't have the labels)

Example - Iris Flower Data set

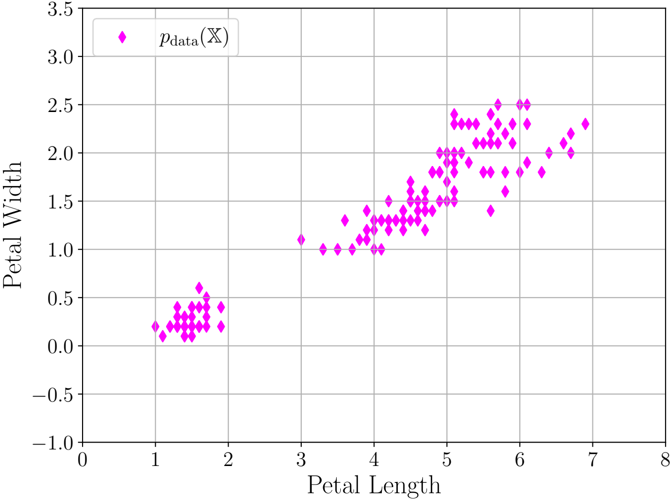


Figure: Iris dataset $\mathcal{D}$

Example - Iris Flower Data set

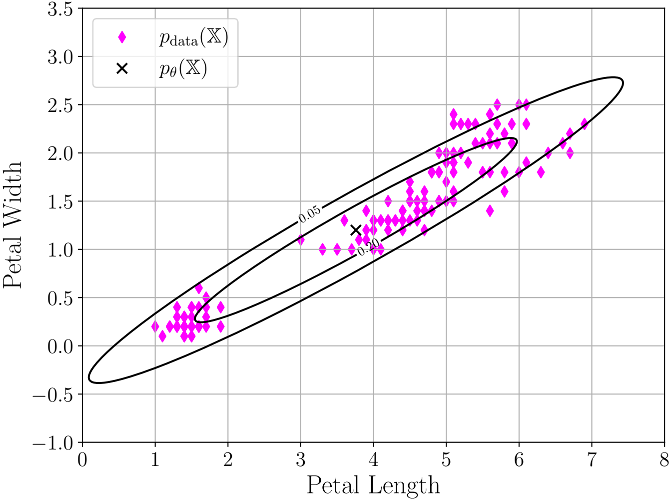


Figure: Modeling distribution using single Gaussian

Example - Iris Flower Data set

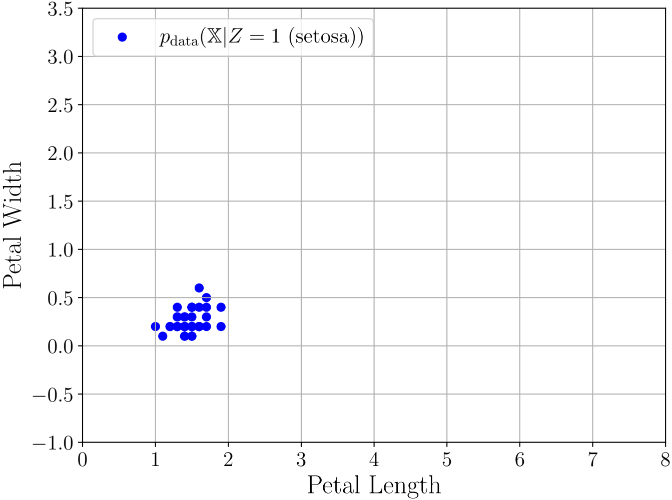


Figure: Data from $p_{\text{data}}(\mathbb{X}|Z = 1)$

Example - Iris Flower Data set

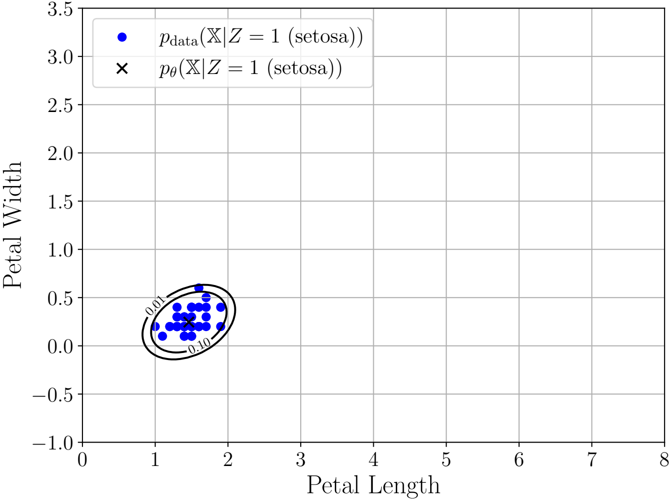


Figure: Modeling conditional distribution $p_{\text{data}}(\mathbb{X}|Z = 1)$ using signle Gaussian

Example - Iris Flower Data set

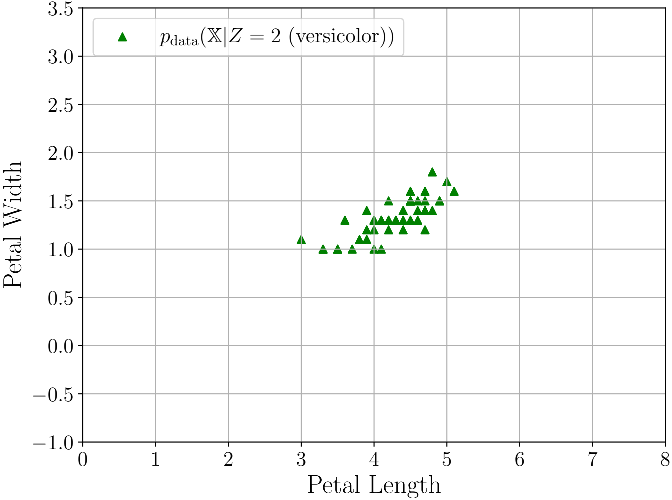


Figure: Data from $p_{\text{data}}(\mathbb{X}|Z = 2)$

Example - Iris Flower Data set

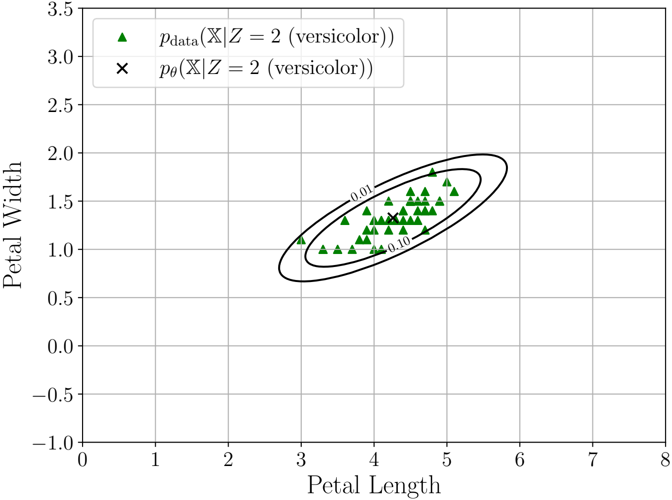


Figure: Modeling conditional distribution $p_{\text{data}}(\mathbb{X}|Z = 2)$ using signle Gaussian

Example - Iris Flower Data set

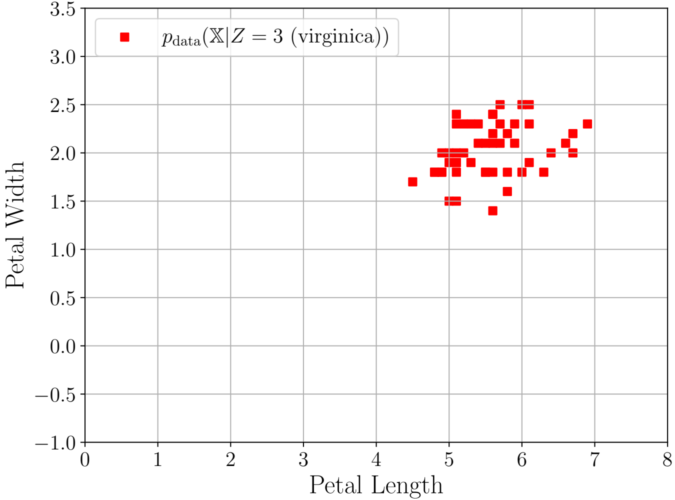


Figure: Data from $p_{\text{data}}(\mathbb{X}|Z3)$

Example - Iris Flower Data set

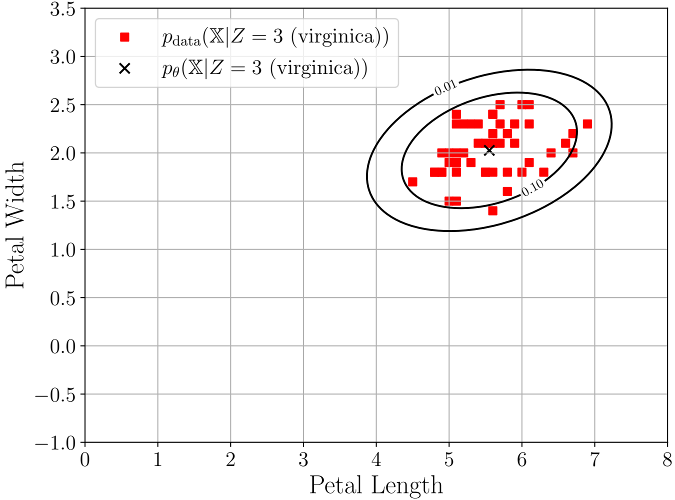


Figure: Modeling conditional distribution $p_{\text{data}}(\mathbb{X}|Z = 3)$ using signle Gaussian

Example - Iris Flower Data set

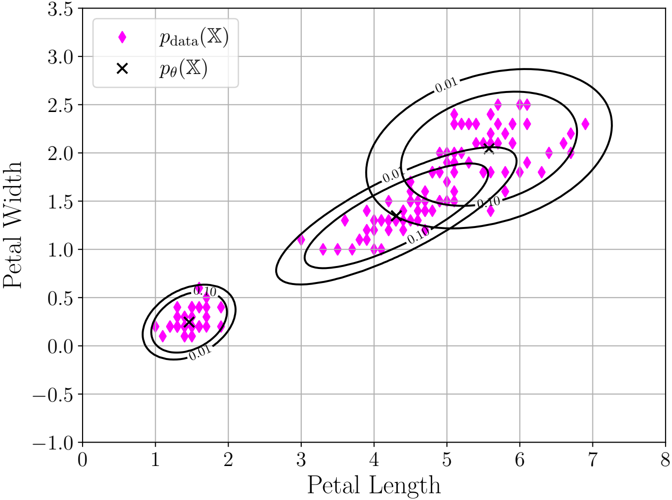


Figure: Modeling unconditional distribution using a Gaussian Mixture Model

Gaussian Mixture Model

Intuition

Intelligently combining simple distributions to create complex ones.

Gaussian Mixture Model

Intuition

Intelligently combining simple distributions to create complex ones.

GMM

Assume you have access to Categorical and Gaussian distribution, then you generate each sample as:

- Generate sample z from Categorical distribution $\mathrm{Cat}(\boldsymbol{\pi})$.

Gaussian Mixture Model

Intuition

Intelligently combining simple distributions to create complex ones.

GMM

Assume you have access to Categorical and Gaussian distribution, then you generate each sample as:

- Generate sample z from Categorical distribution $\text{Cat}(\boldsymbol{\pi})$.
- Generate sample $\boldsymbol{x}$ from conditional distribution $p(\mathbb{X}|z) = \mathcal{N}(\mathbb{X}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$

Gaussian Mixture Model

Intuition

Intelligently combining simple distributions to create complex ones.

GMM

Assume you have access to Categorical and Gaussian distribution, then you generate each sample as:

- Generate sample z from Categorical distribution $\text{Cat}(\boldsymbol{\pi})$.
- Generate sample $\boldsymbol{x}$ from conditional distribution $p(\mathbb{X}|z) = \mathcal{N}(\mathbb{X}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$

In the above process, the distribution over $\mathbb{X}$ is:

$$p(\mathbb{X}) = \sum_z p(\mathbb{X}, Z = z) = \sum_z p(\mathbb{X}|z)p(z) = \sum_z \pi_z \mathcal{N}(\mathbb{X}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$$

Gaussian Mixture Model

Intuition

Intelligently combining simple distributions to create complex ones.

GMM

Assume you have access to Categorical and Gaussian distribution, then you generate each sample as:

- Generate sample z from Categorical distribution $\mathrm{Cat}(\boldsymbol{\pi})$.
- Generate sample $\boldsymbol{x}$ from conditional distribution $p(\mathbb{X}|z) = \mathcal{N}(\mathbb{X}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$

In the above process, the distribution over $\mathbb{X}$ is:

$$p(\mathbb{X}) = \sum_z p(\mathbb{X}, Z = z) = \sum_z p(\mathbb{X}|z)p(z) = \sum_z \pi_z \mathcal{N}(\mathbb{X}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$$

while the Gaussian distribution is unimodal, the above distribution is multi-modal.

GMM

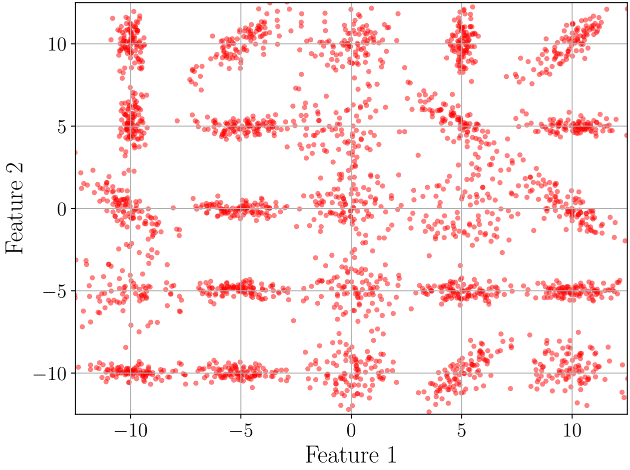


Figure: Synthesis dataset $\mathcal{D}$ samples

GMM

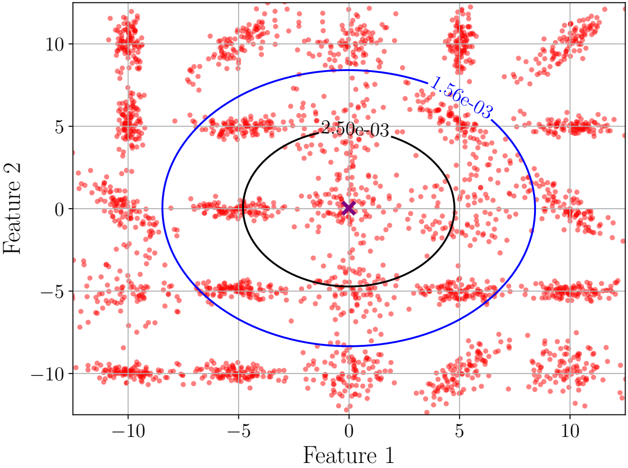


Figure: Gaussian distribution fitted to dataset $\mathcal{D}$ (the purple '×'symbol shows the mean location)

GMM

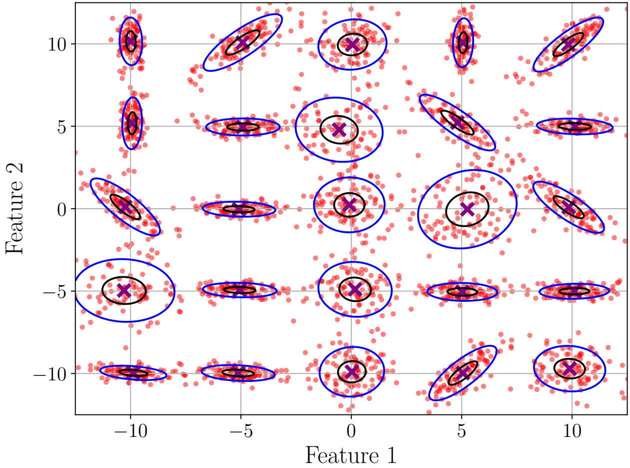


Figure: GMM distribution with four component fitted to dataset $\mathcal{D}$ with 25 components (the purple '$\times$'symbols shows each component mean location)

GMM

Pros

- A flexible distribution based on simple distribution functions

GMM

Pros

- A flexible distribution based on simple distribution functions
- Ancestral sampling based on simple distributions
  1. Sample $z \sim \mathrm{Cat}(\boldsymbol{\pi})$

GMM

Pros

- A flexible distribution based on simple distribution functions
- Ancestral sampling based on simple distributions
  1. Sample $z \sim \mathrm{Cat}(\boldsymbol{\pi})$
  2. Sample $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$

GMM

Pros

- A flexible distribution based on simple distribution functions
- Ancestral sampling based on simple distributions
  1. Sample $z \sim \mathrm{Cat}(\boldsymbol{\pi})$
  2. Sample $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$
- Latent representation $\boldsymbol{z}$ (For a given data point $\boldsymbol{x}$ in Iris dataset, if we can calculate $p(Z|\boldsymbol{x})$ then we can predict the Iris flower type)

GMM

Pros

- A flexible distribution based on simple distribution functions
- Ancestral sampling based on simple distributions
  1. Sample $z \sim \mathrm{Cat}(\boldsymbol{\pi})$
  2. Sample $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$
- Latent representation $\boldsymbol{z}$ (For a given data point $\boldsymbol{x}$ in Iris dataset, if we can calculate $p(Z|\boldsymbol{x})$ then we can predict the Iris flower type)

Cons

- No exact density estimation for new samples because of hidden variable $\boldsymbol{z}$

GMM

Pros

- A flexible distribution based on simple distribution functions
- Ancestral sampling based on simple distributions
    1. Sample $z \sim \text{Cat}(\boldsymbol{\pi})$
    2. Sample $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$
- Latent representation $\boldsymbol{z}$ (For a given data point $\boldsymbol{x}$ in Iris dataset, if we can calculate $p(Z|\boldsymbol{x})$ then we can predict the Iris flower type)

Cons

- No exact density estimation for new samples because of hidden variable $\boldsymbol{z}$
- Sensitivity to initialization

GMM

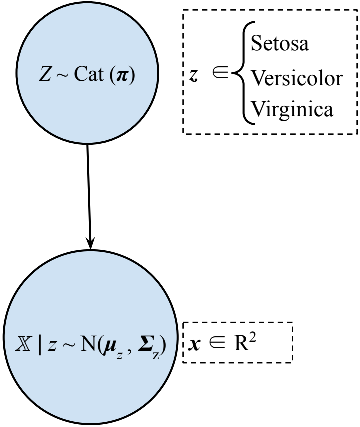
Pros

- A flexible distribution based on simple distribution functions
- Ancestral sampling based on simple distributions
  1. Sample $z \sim \mathrm{Cat}(\boldsymbol{\pi})$
  2. Sample $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$
- Latent representation $\boldsymbol{z}$ (For a given data point $\boldsymbol{x}$ in Iris dataset, if we can calculate $p(Z|\boldsymbol{x})$ then we can predict the Iris flower type)

Cons

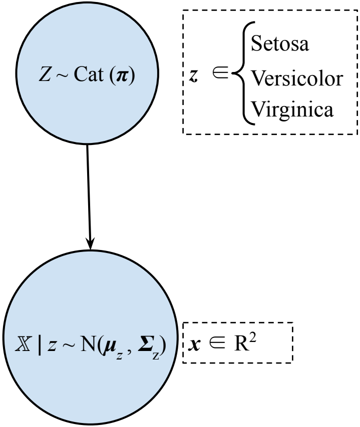
- No exact density estimation for new samples because of hidden variable $\boldsymbol{z}$
- Sensitivity to initialization
- Number of component

Extending GMM

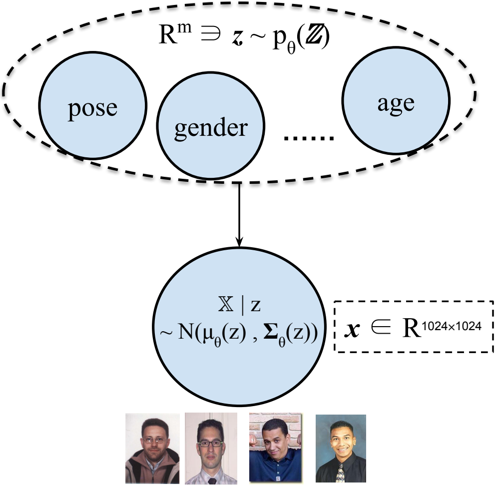


(a) GMM Concept

Extending GMM



(a) GMM Concept



(b) VAE Concept (images sourse: [1])

Section 2

Model Specification

Sample Model

Model Specification

Assume we have the following simple distributions:

- Latent variable distribution:

$$p_\theta(\mathbb{Z}) = \mathcal{N}(\mathbb{Z}|\boldsymbol{\mu}_z, \sigma_z^2\boldsymbol{I}),\ \mathbb{Z} \in \mathbb{R}^m$$

Sample Model

Model Specification

Assume we have the following simple distributions:

- Latent variable distribution:

$$p_\theta(\mathbb{Z}) = \mathcal{N}(\mathbb{Z}|\boldsymbol{\mu}_z, \sigma_z^2 \boldsymbol{I}),\ \mathbb{Z} \in \mathbb{R}^m$$

- Conditional observed variable distribution:

$$p_\theta(\mathbb{X}|\boldsymbol{z}) = \mathcal{N}(\mathbb{X}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})), \quad \begin{cases} \mathbb{X} \in \mathbb{R}^n \\ \boldsymbol{\mu}_\theta(\boldsymbol{z}) = \mathrm{NN}_{\boldsymbol{\alpha}}(\boldsymbol{z}) \\ \boldsymbol{\Sigma}_\theta(\boldsymbol{z}) = \mathrm{diag}\left(\sigma\left(\mathrm{NN}_{\boldsymbol{\beta}}(\boldsymbol{z})\right)\right) \end{cases}$$

Sample Model

Model Specification

Assume we have the following simple distributions:

- Latent variable distribution:

$$p_\theta(\mathbb{Z}) = \mathcal{N}(\mathbb{Z}|\boldsymbol{\mu}_z, \sigma_z^2 \boldsymbol{I}),\ \mathbb{Z} \in \mathbb{R}^m$$

- Conditional observed variable distribution:

$$p_\theta(\mathbb{X}|\boldsymbol{z}) = \mathcal{N}(\mathbb{X}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})), \quad \begin{cases} \mathbb{X} \in \mathbb{R}^n \\ \boldsymbol{\mu}_\theta(\boldsymbol{z}) = \mathrm{NN}_{\boldsymbol{\alpha}}(\boldsymbol{z}) \\ \boldsymbol{\Sigma}_\theta(\boldsymbol{z}) = \mathrm{diag}\left(\sigma\left(\mathrm{NN}_{\boldsymbol{\beta}}(\boldsymbol{z})\right)\right) \end{cases}$$

thus $\boldsymbol{\theta} = \{\boldsymbol{\mu}_z, \sigma_z^2, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$.

Task 1: Generation

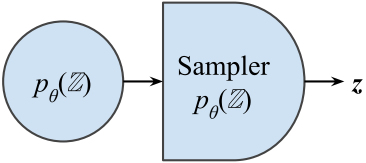


Figure: Sampling latent vector

Task 1: Generation

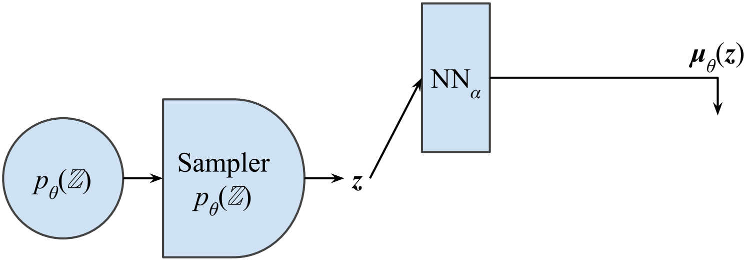


Figure: Calculating conditional distribution mean vector

Task 1: Generation

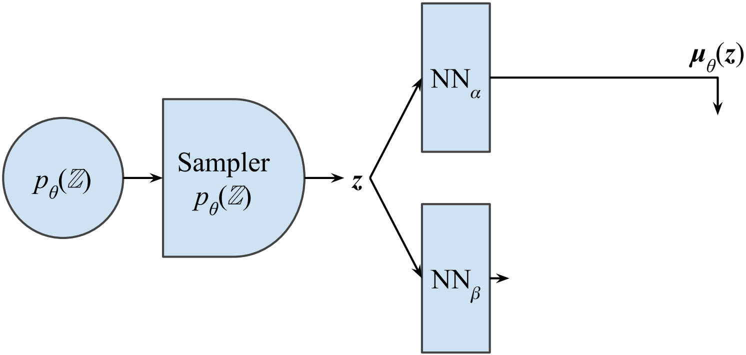


Figure: Calculating conditional distribution covariance matrix

Task 1: Generation

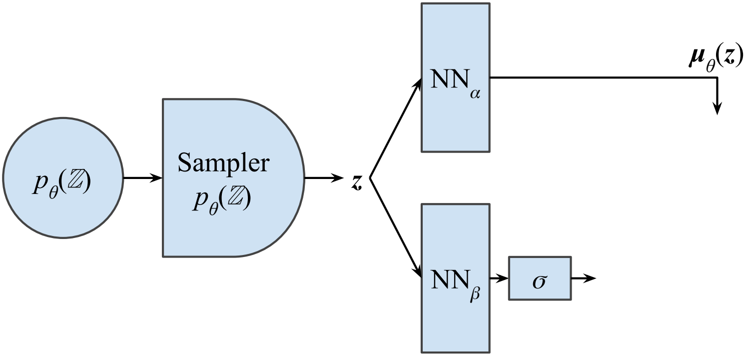


Figure: Calculating conditional distribution covariance matrix

Task 1: Generation

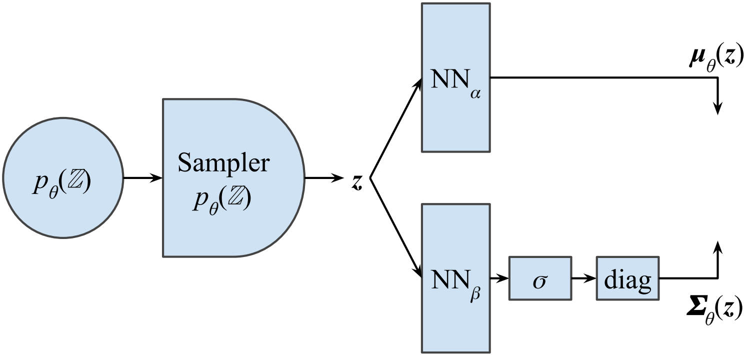


Figure: Calculating conditional distribution covariance matrix

Task 1: Generation

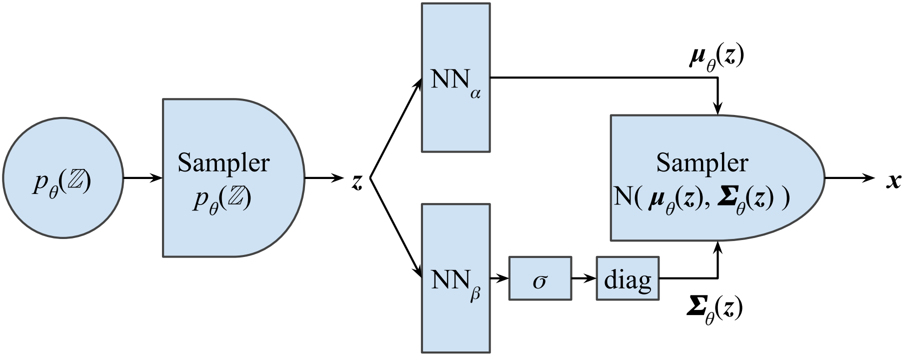


Figure: Sampling conditional distribution

Task 2: High-level Representation

High-level Representation

Remember the example of the Iris dataset. Accessing the latent random variable Z can reveal information about the type of flower. So Z can be considered as *High-level Representation*.

Task 2: High-level Representation

High-level Representation

Remember the example of the Iris dataset. Accessing the latent random variable Z can reveal information about the type of flower. So Z can be considered as *High-level Representation*.

Challenges

For estimating $\boldsymbol{z}$ corresponding to an observed vector $\boldsymbol{x}$, we have the following challeneges:

- Estimating the parameters of conditional distribution $\mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z}))$ with only one sample $\boldsymbol{x}$ has a high variance.

Task 2: High-level Representation

High-level Representation

Remember the example of the Iris dataset. Accessing the latent random variable Z can reveal information about the type of flower. So Z can be considered as *High-level Representation*.

Challenges

For estimating $\boldsymbol{z}$ corresponding to an observed vector $\boldsymbol{x}$, we have the following challeneges:

- Estimating the parameters of conditional distribution $\mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z}))$ with only one sample $\boldsymbol{x}$ has a high variance.
- Assume you have access to both $\boldsymbol{\mu}_\theta(\boldsymbol{z})$ and $\boldsymbol{\Sigma}_\theta(\boldsymbol{z})$. Then calculating $\boldsymbol{z}$ is impossible because Deep Neural Networks are non-invertible.

Task 3: Density Estimation

Density Estimation

We have access to both $p_\theta(\mathbb{Z})$ and $p_\theta(\mathbb{X}|\mathbb{Z})$, thus:

$$p_\theta(\mathbb{X}, \mathbb{Z}) = p_\theta(\mathbb{X}|\mathbb{Z})p_\theta(\mathbb{Z})$$

Task 3: Density Estimation

Density Estimation

We have access to both $p_\theta(\mathbb{Z})$ and $p_\theta(\mathbb{X}|\mathbb{Z})$, thus:

$$p_\theta(\mathbb{X}, \mathbb{Z}) = p_\theta(\mathbb{X}|\mathbb{Z})p_\theta(\mathbb{Z})$$

To calculate the density, we need to marginalize out the hidden random vector $\mathbb{Z}$, thus:

$$p_\theta(\mathbb{X}) = \int_{\boldsymbol{z}} p_\theta(\mathbb{X}, \boldsymbol{z})\mathrm{d}\boldsymbol{z}$$

Task 3: Density Estimation

Density Estimation

We have access to both $p_\theta(\mathbb{Z})$ and $p_\theta(\mathbb{X}|\mathbb{Z})$, thus:

$$p_\theta(\mathbb{X}, \mathbb{Z}) = p_\theta(\mathbb{X}|\mathbb{Z}) p_\theta(\mathbb{Z})$$

To calculate the density, we need to marginalize out the hidden random vector $\mathbb{Z}$, thus:

$$\begin{aligned} p_\theta(\mathbb{X}) &= \int_{\boldsymbol{z}} p_\theta(\mathbb{X}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} \\ &= \int_{\boldsymbol{z}} p_\theta(\mathbb{X}|\boldsymbol{z}) p_\theta(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \end{aligned}$$

Task 3: Density Estimation

Density Estimation

We have access to both $p_\theta(\mathbb{Z})$ and $p_\theta(\mathbb{X}|\mathbb{Z})$, thus:

$$p_\theta(\mathbb{X}, \mathbb{Z}) = p_\theta(\mathbb{X}|\mathbb{Z})p_\theta(\mathbb{Z})$$

To calculate the density, we need to marginalize out the hidden random vector $\mathbb{Z}$, thus:

$$\begin{aligned} p_\theta(\mathbb{X}) &= \int_{\boldsymbol{z}} p_\theta(\mathbb{X}, \boldsymbol{z})\mathrm{d}\boldsymbol{z} \\ &= \int_{\boldsymbol{z}} p_\theta(\mathbb{X}|\boldsymbol{z})p_\theta(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \\ &= \int_{\boldsymbol{z}} \mathcal{N}\left(\mathbb{X}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\right) p_\theta(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \end{aligned}$$

Task 3: Density Estimation

Density Estimation

We have access to both $p_\theta(\mathbb{Z})$ and $p_\theta(\mathbb{X}|\mathbb{Z})$, thus:

$$p_\theta(\mathbb{X}, \mathbb{Z}) = p_\theta(\mathbb{X}|\mathbb{Z})p_\theta(\mathbb{Z})$$

To calculate the density, we need to marginalize out the hidden random vector $\mathbb{Z}$, thus:

$$\begin{aligned}
p_\theta(\mathbb{X}) &= \int_{\boldsymbol{z}} p_\theta(\mathbb{X}, \boldsymbol{z})\mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}} p_\theta(\mathbb{X}|\boldsymbol{z})p_\theta(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}} \mathcal{N}\left(\mathbb{X}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\right) p_\theta(\boldsymbol{z})\mathrm{d}\boldsymbol{z}
\end{aligned}$$

Challenge 1

Calculating the likelihood and thus training the model seems to be intractable.

Section 3

Casting Likelihood Calculation as Expectation

No Free Lunch: Challenges Start

Calculating the Likelihood for Sample $\boldsymbol{x}$

As we see before, we are interested in the following optimization:

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} -\frac{1}{|\mathcal{D}|} \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}_i)$$

No Free Lunch: Challenges Start

Calculating the Likelihood for Sample $\boldsymbol{x}$

As we see before, we are interested in the following optimization:

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} -\frac{1}{|\mathcal{D}|} \sum_{i=1}^{N} \log p_{\theta}(\boldsymbol{x}_i)$$

To compute $p_{\theta}(\boldsymbol{x})$ in a latent variable model we have:

$$p_{\theta}(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_{\theta}(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z}$$

No Free Lunch: Challenges Start

Calculating the Likelihood for Sample $\boldsymbol{x}$

As we see before, we are interested in the following optimization:

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{N}\log p_\theta(\boldsymbol{x}_i)$$

To compute $p_\theta(\boldsymbol{x})$ in a latent variable model we have:

$$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}, \boldsymbol{z})\mathrm{d}\boldsymbol{z}$$

The above integral is hard to compute and in each update of the optimization problem, we want:

- Evaluate the integral as a function of $\boldsymbol{\theta}$
- Evaluate the gradient vector of the log-likelihood

Likelihood as Expectation

We can write the data likelihood as:

$$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{z}) p_\theta(\boldsymbol{x}|\boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \mathbb{E}_{\boldsymbol{z} \sim p_\theta(\mathbb{Z})} \left[p_\theta(\boldsymbol{x}|\boldsymbol{z}) \right]$$

Thus we have an expectation. How can we estimate it?

Likelihood as Expectation

Likelihood as Expectation

We can write the data likelihood as:

$$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}, \boldsymbol{z})\mathrm{d}\boldsymbol{z} = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{z})p_\theta(\boldsymbol{x}|\boldsymbol{z})\mathrm{d}\boldsymbol{z} = \mathbb{E}_{\boldsymbol{z}\sim p_\theta(\mathbb{Z})}\left[p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$$

Thus we have an expectation. How can we estimate it?

Mante Carlo Extimate

We can estimate the expectation using the Monte Carlo estimate by k samples from the distribution as:

$$p_\theta(\boldsymbol{x}) \simeq \frac{1}{k}\sum_{i=1}^{k} p_\theta(\boldsymbol{x}|\boldsymbol{z}_i), \;\; \boldsymbol{z}_i \sim p_\theta(\mathbb{Z}) \text{ for } i = 1, \ldots, k$$

The above estimate, while unbiased, exhibits significant variance. Why?

High Variance Estimation - Toy Example

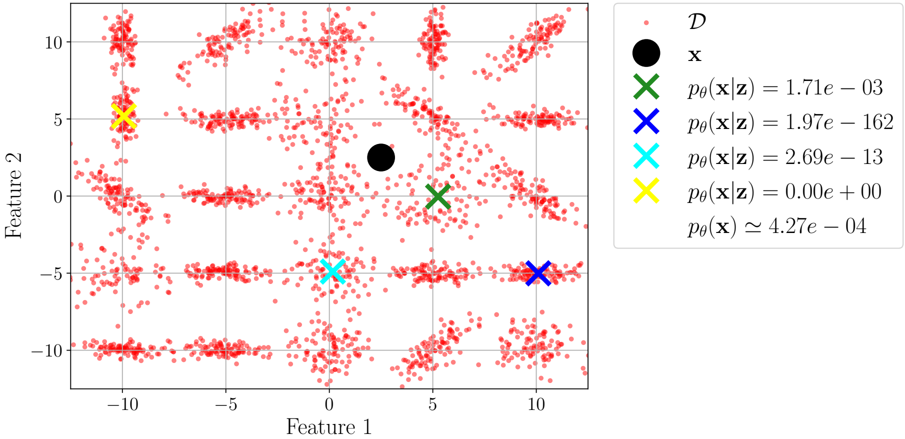


Figure: $[4.27 \times 10^{-4}]$

High Variance Estimation - Toy Example

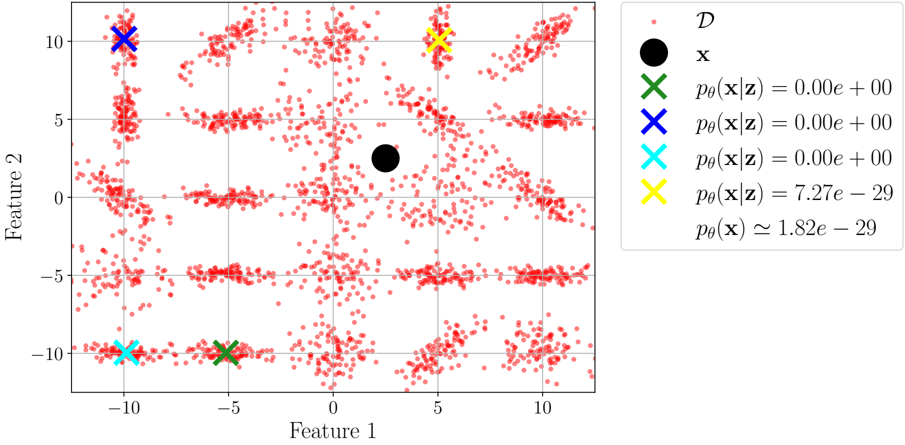


Figure: $[4.27 \times 10^{-4},\ 1.82 \times 10^{-29}]$

High Variance Estimation - Toy Example

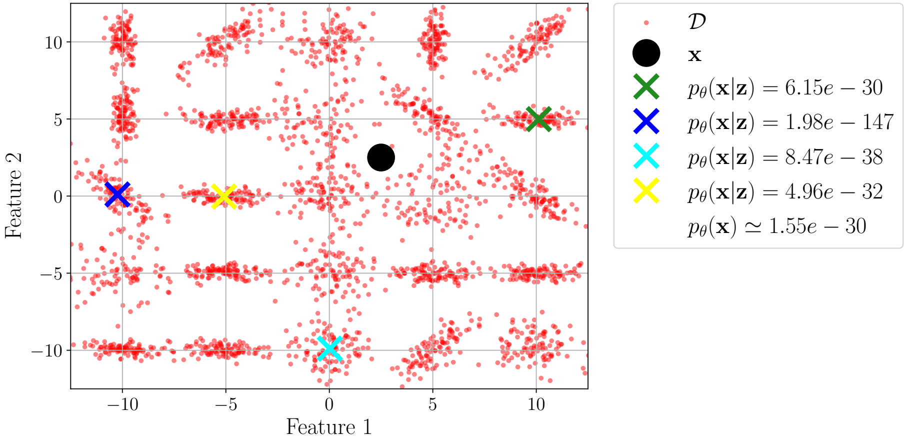


Figure: $[4.27 \times 10^{-4},\ 1.82 \times 10^{-29},\ 1.55 \times 10^{-30}]$

High Variance Estimation - Toy Example

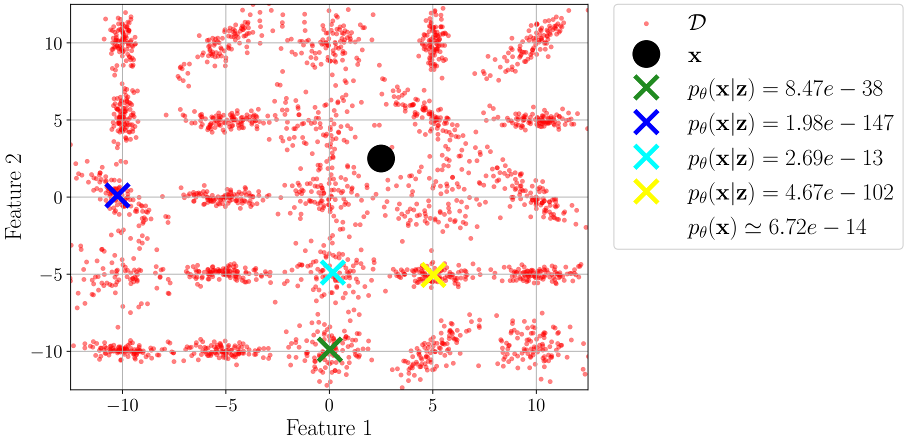


Figure: $[4.27 \times 10^{-4},\ 1.82 \times 10^{-29},\ 1.55 \times 10^{-30},\ 6.72 \times 10^{-14}]$

High Variance Estimate - Real World Example

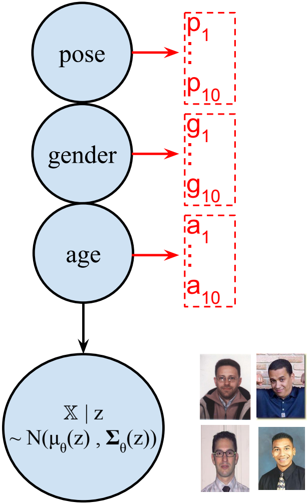


Figure: (images source: [1])

High Variance Estimate - Real World Example

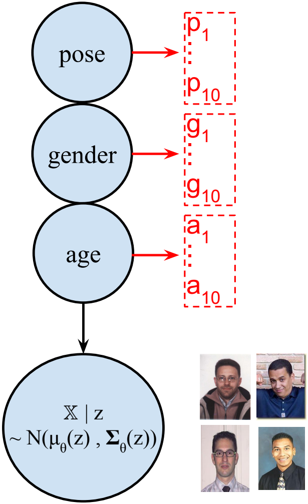


Figure: (images source: [1])

Scenario

As we can see $|\mathcal{Z}| = 10^3$. Assume we want to compute $p_\theta(\overline{\boldsymbol{x}})$, the true latent is $\overline{\boldsymbol{z}}$ and:

- $k = 10$
- $\mathcal{S}$ is the set of generated latent vectors

Then:

High Variance Estimate - Real World Example

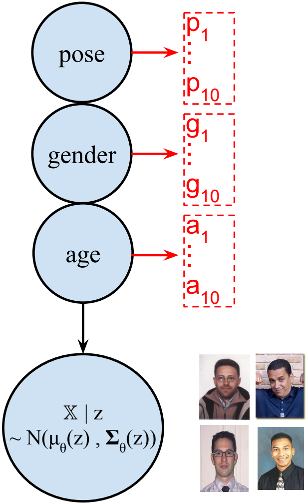


Figure: (images source: [1])

Scenario

As we can see $|\mathcal{Z}| = 10^3$. Assume we want to compute $p_\theta(\overline{\boldsymbol{x}})$, the true latent is $\overline{\boldsymbol{z}}$ and:

- $k = 10$
- $\mathcal{S}$ is the set of generated latent vectors

Then:

- $p\{\overline{\boldsymbol{z}} \in \mathcal{S}\} = 0.01 \Rightarrow p_\theta(\overline{\boldsymbol{x}}) > 0$

High Variance Estimate - Real World Example

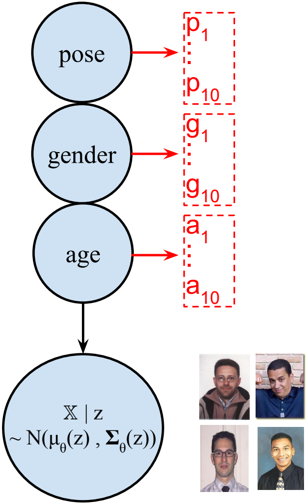


Figure: (images source: [1])

Scenario

As we can see $|\mathcal{Z}| = 10^3$. Assume we want to compute $p_\theta(\overline{\boldsymbol{x}})$, the true latent is $\overline{\boldsymbol{z}}$ and:

- $k = 10$
- $\mathcal{S}$ is the set of generated latent vectors

Then:

- $p\{\overline{\boldsymbol{z}} \in \mathcal{S}\} = 0.01 \Rightarrow p_\theta(\overline{\boldsymbol{x}}) > 0$
- $p\{\overline{\boldsymbol{z}} \notin \mathcal{S}\} = 0.99 \Rightarrow p_\theta(\overline{\boldsymbol{x}}) \simeq 0$

High Variance Estimate - Real World Example

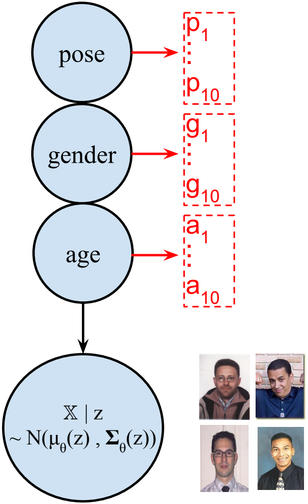


Figure: (images source: [1])

Scenario

As we can see $|\mathcal{Z}| = 10^3$. Assume we want to compute $p_\theta(\overline{\boldsymbol{x}})$, the true latent is $\overline{\boldsymbol{z}}$ and:

- $k = 10$
- $\mathcal{S}$ is the set of generated latent vectors

Then:

- $p\{\overline{\boldsymbol{z}} \in \mathcal{S}\} = 0.01 \Rightarrow p_\theta(\overline{\boldsymbol{x}}) > 0$
- $p\{\overline{\boldsymbol{z}} \notin \mathcal{S}\} = 0.99 \Rightarrow p_\theta(\overline{\boldsymbol{x}}) \simeq 0$

while the estimation has a high variance and cannot be used in practice.

☞ So, what can we do?

Update on our Challenges

Challenge 1

Calculating the likelihood and thus training the model seems to be intractable.

Update on our Challenges

Challenge 1

Calculating the likelihood and thus training the model seems to be intractable.

☞ We can cast the likelihood as expectation and estimate the expectation using a Mnote Carlo Estimation

Update on our Challenges

Challenge 1

Calculating the likelihood and thus training the model seems to be intractable.

- ☞ We can cast the likelihood as expectation and estimate the expectation using a Mnote Carlo Estimation

Challenge 2

Estimating the likelihood using prior distribution, while unbiased, represents high variance.

Section 4

Importance Sampling

Importance Sampling

To calculate the likelihood, we can use importance sampling:

Importance Sampling

To calculate the likelihood, we can use importance sampling:

$$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z}$$

Importance Sampling

To calculate the likelihood, we can use importance sampling:

$$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \int_{\boldsymbol{z}} q(\boldsymbol{z}) \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \mathrm{d}\boldsymbol{z}$$

Importance Sampling

To calculate the likelihood, we can use importance sampling:

$$
\begin{aligned}
p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} &= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \mathrm{d}\boldsymbol{z} \\
&= \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z})} \left[\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right]
\end{aligned}
$$

Importance Sampling

Importance Sampling

To calculate the likelihood, we can use importance sampling:

$$
\begin{aligned}
p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \int_{\boldsymbol{z}} q(\boldsymbol{z}) \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \mathrm{d}\boldsymbol{z} \\
= \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z})} \left[\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right]
\end{aligned}
$$

where $q(\boldsymbol{z})$ can be any arbitrary distribution function over $\mathbb{Z}$.

Importance Sampling

Importance Sampling

To calculate the likelihood, we can use importance sampling:

$$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \int_{\boldsymbol{z}} q(\boldsymbol{z}) \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \mathrm{d}\boldsymbol{z}$$
$$= \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z})} \left[\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right]$$

where $q(\boldsymbol{z})$ can be any arbitrary distribution function over $\mathbb{Z}$. We can estimate the distribution using the Monte Carlo estimate:

$$p_\theta(\boldsymbol{x}) \simeq \frac{1}{k} \sum_{i=1}^{k} \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z}_i)}{q(\boldsymbol{z}_i)}, \ \boldsymbol{z}_i \sim q(\mathbb{Z}) \text{ for } i = 1, \ldots, k$$

Importance Sampling

Importance Sampling

To calculate the likelihood, we can use importance sampling:

$$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \int_{\boldsymbol{z}} q(\boldsymbol{z}) \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \mathrm{d}\boldsymbol{z}$$
$$= \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z})} \left[\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right]$$

where $q(\boldsymbol{z})$ can be any arbitrary distribution function over $\mathbb{Z}$. We can estimate the distribution using the Monte Carlo estimate:

$$p_\theta(\boldsymbol{x}) \simeq \frac{1}{k} \sum_{i=1}^{k} \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z}_i)}{q(\boldsymbol{z}_i)}, \ \boldsymbol{z}_i \sim q(\mathbb{Z}) \text{ for } i = 1, \ldots, k$$

Challenge

What is a good option for $q(\mathbb{Z})$. Maybe we need optimization over q too!

Importance Sampling in Toy Example

Estimating Latent Posterior

Assume the toy example of GMM, we can calculate $p_\theta(\mathbb{Z}|\boldsymbol{x})$ as:

Importance Sampling in Toy Example

Estimating Latent Posterior

Assume the toy example of GMM, we can calculate $p_\theta(\mathbb{Z}|\boldsymbol{x})$ as:

$$p_\theta(\mathbb{Z} = i|\boldsymbol{x}) = \frac{p_\theta(\boldsymbol{x}, \mathbb{Z} = i)}{p_\theta(\boldsymbol{x})} \qquad \text{\#Bayes Rule}$$

Importance Sampling in Toy Example

Estimating Latent Posterior

Assume the toy example of GMM, we can calculate $p_\theta(\mathbb{Z}|\boldsymbol{x})$ as:

$$p_\theta(\mathbb{Z}=i|\boldsymbol{x}) = \frac{p_\theta(\boldsymbol{x},\mathbb{Z}=i)}{p_\theta(\boldsymbol{x})} \qquad \text{\#Bayes Rule}$$

$$= \frac{p_\theta(\boldsymbol{x},\mathbb{Z}=i)}{\sum_j p_\theta(\boldsymbol{x},\mathbb{Z}=j)} \qquad \text{\#Marginalization}$$

Importance Sampling in Toy Example

Estimating Latent Posterior

Assume the toy example of GMM, we can calculate $p_\theta(\mathbb{Z}|\boldsymbol{x})$ as:

$$p_\theta(\mathbb{Z} = i|\boldsymbol{x}) = \frac{p_\theta(\boldsymbol{x}, \mathbb{Z} = i)}{p_\theta(\boldsymbol{x})} \qquad \text{\#Bayes Rule}$$

$$= \frac{p_\theta(\boldsymbol{x}, \mathbb{Z} = i)}{\sum_j p_\theta(\boldsymbol{x}, \mathbb{Z} = j)} \qquad \text{\#Marginalization}$$

$$= \frac{p_\theta(\boldsymbol{x}|\mathbb{Z} = i) p_\theta(\mathbb{Z} = i)}{\sum_j p_\theta(\boldsymbol{x}|\mathbb{Z} = j) p_\theta(\mathbb{Z} = j)} \qquad \text{\#Chain Rule}$$

Importance Sampling in Toy Example

Estimating Latent Posterior

Assume the toy example of GMM, we can calculate $p_\theta(\mathbb{Z}|\boldsymbol{x})$ as:

$$p_\theta(\mathbb{Z}=i|\boldsymbol{x}) = \frac{p_\theta(\boldsymbol{x},\mathbb{Z}=i)}{p_\theta(\boldsymbol{x})} \qquad \text{\#Bayes Rule}$$

$$= \frac{p_\theta(\boldsymbol{x},\mathbb{Z}=i)}{\sum_j p_\theta(\boldsymbol{x},\mathbb{Z}=j)} \qquad \text{\#Marginalization}$$

$$= \frac{p_\theta(\boldsymbol{x}|\mathbb{Z}=i)p_\theta(\mathbb{Z}=i)}{\sum_j p_\theta(\boldsymbol{x}|\mathbb{Z}=j)p_\theta(\mathbb{Z}=j)} \qquad \text{\#Chain Rule}$$

$$= \frac{\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\pi_i}{\sum_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\pi_j} \qquad \text{Model Parameters}$$

Importance Sampling in Toy Example

Estimating Latent Posterior

Assume the toy example of GMM, we can calculate $p_\theta(\mathbb{Z}|\boldsymbol{x})$ as:

$$
\begin{aligned}
p_\theta(\mathbb{Z}=i|\boldsymbol{x}) =& \frac{p_\theta(\boldsymbol{x},\mathbb{Z}=i)}{p_\theta(\boldsymbol{x})} && \text{\#Bayes Rule} \\
=& \frac{p_\theta(\boldsymbol{x},\mathbb{Z}=i)}{\sum_j p_\theta(\boldsymbol{x},\mathbb{Z}=j)} && \text{\#Marginalization} \\
=& \frac{p_\theta(\boldsymbol{x}|\mathbb{Z}=i)p_\theta(\mathbb{Z}=i)}{\sum_j p_\theta(\boldsymbol{x}|\mathbb{Z}=j)p_\theta(\mathbb{Z}=j)} && \text{\#Chain Rule} \\
=& \frac{\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\pi_i}{\sum_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\pi_j} && \text{Model Parameters}
\end{aligned}
$$

Let's use importance sampling technique to calculate $p_\theta(\boldsymbol{x})$ with $q(Z) = p_\theta(Z|\boldsymbol{x})$

GMM - Posterior

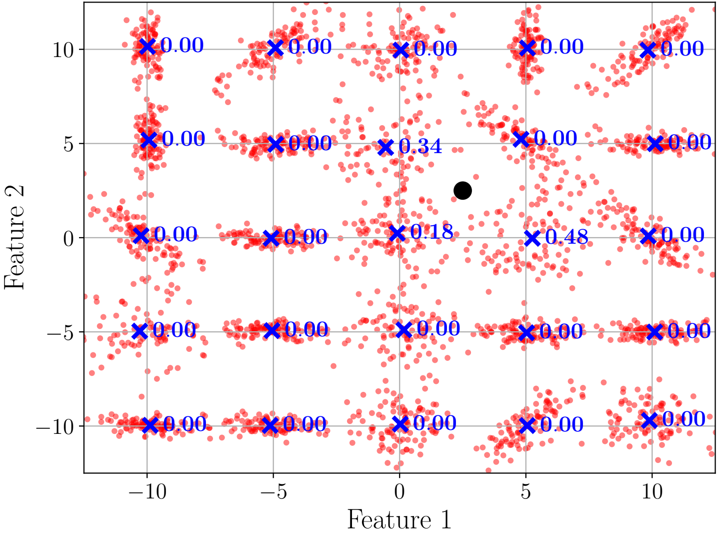


Figure: Posterior distribution $p(Z|\boldsymbol{x})$ (black circle represents $\boldsymbol{x}$)

High Variance Estimation - Toy Example

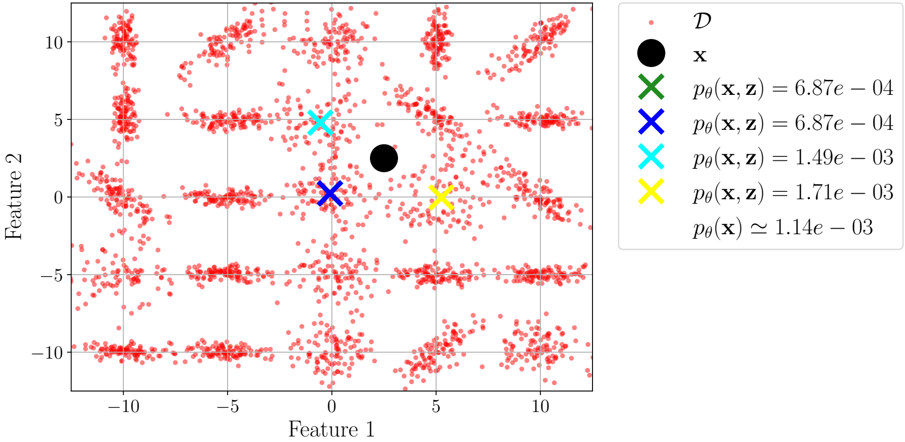


Figure: $[1.14 \times 10^{-3}]$

High Variance Estimation - Toy Example

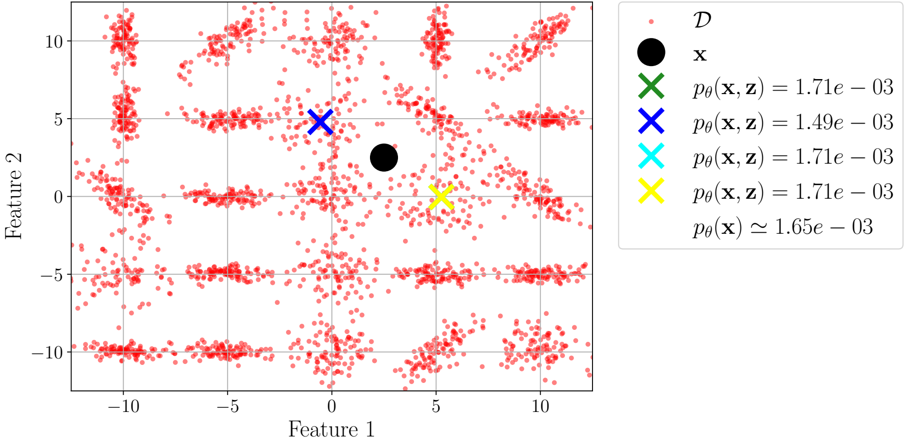


Figure: $[1.14 \times 10^{-3},\ 1.65 \times 10^{-3}]$

High Variance Estimation - Toy Example

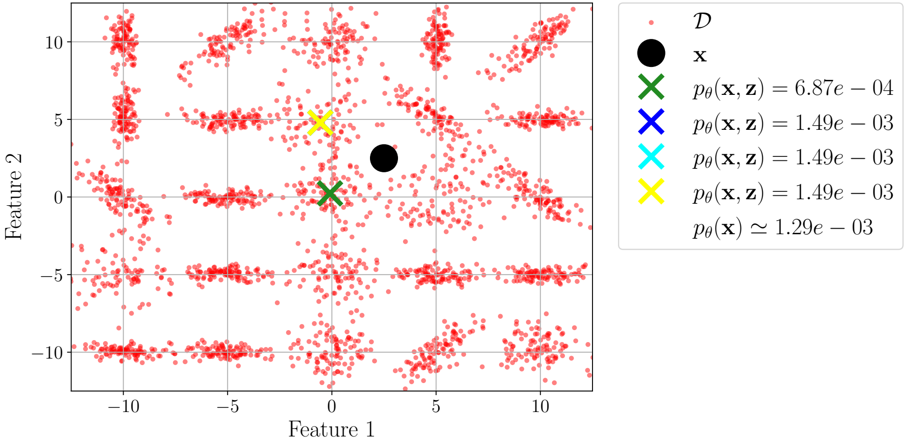


Figure: $[1.14 \times 10^{-3},\ 1.65 \times 10^{-3},\ 1.29 \times 10^{-3}]$

High Variance Estimation - Toy Example

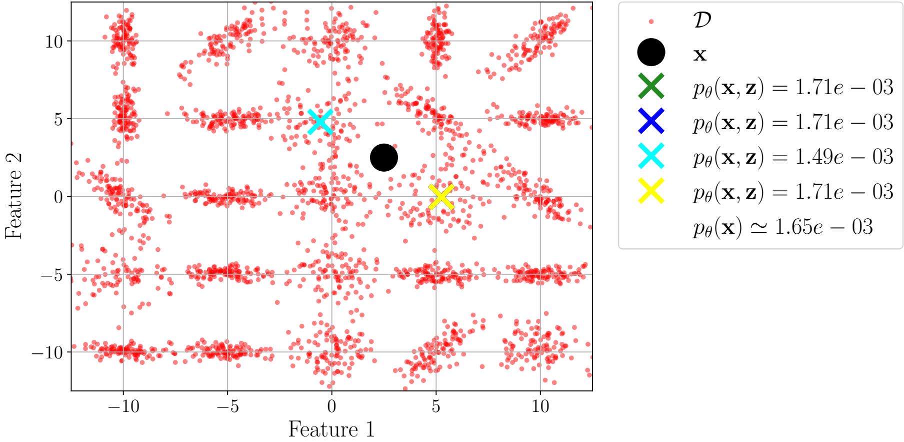


Figure: $[1.14 \times 10^{-3},\ 1.65 \times 10^{-3},\ 1.29 \times 10^{-3},\ 1.65 \times 10^{-3}]$

Lowering Estimation Variance

Real-World Case

Although for the case of GMM latent variable posterior is tractable, in real-world application we have (assuming the latent prior is standard Gaussian):

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

Lowering Estimation Variance

Real-World Case

Although for the case of GMM latent variable posterior is tractable, in real-world application we have (assuming the latent prior is standard Gaussian):

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

and we know that calculating $p_\theta(\boldsymbol{x})$ is challenging (indeed we are looking for it!).

Lowering Estimation Variance

Real-World Case

Although for the case of GMM latent variable posterior is tractable, in real-world application we have (assuming the latent prior is standard Gaussian):

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

and we know that calculating $p_\theta(\boldsymbol{x})$ is challenging (indeed we are looking for it!). So you can't use $p_\theta(\mathbb{Z}|\boldsymbol{x})$ as $q(\mathbb{Z})$ in the importance sampling technique.

Update on our Challenges

Challenge 2

Estimating the likelihood using prior distribution, while unbiased, represents high variance.

Update on our Challenges

Challenge 2

Estimating the likelihood using prior distribution, while unbiased, represents high variance.

- Importance sampling may help. We observe $q(\mathbb{Z}) = p_\theta(\mathbb{Z}|\boldsymbol{x})$ is a good option at least in the case of GMM using the importance sampling technique.

Update on our Challenges

Challenge 2

Estimating the likelihood using prior distribution, while unbiased, represents high variance.

- ☞ Importance sampling may help. We observe $q(\mathbb{Z}) = p_\theta(\mathbb{Z}|\boldsymbol{x})$ is a good option at least in the case of GMM using the importance sampling technique.

Challenge 3

We have no formal derivation for the suitability of $p_\theta(\mathbb{Z}|\boldsymbol{x})$ to be used for importance sampling in the general real-world case (other than GMM). So we don't know how to select $q(\mathbb{Z})$

Section 5

Evidence Lower BOund

Working on Log Likelihood

Log Likelihood

In the previous section, we have seen that:

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

But what we need in practice for training:

Working on Log Likelihood

Log Likelihood

In the previous section, we have seen that:

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

But what we need in practice for training:

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

Working on Log Likelihood

Log Likelihood

In the previous section, we have seen that:

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

But what we need in practice for training:

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

For two important reasons (among others), we need to change the order of expectation and log on RHS.

Working on Log Likelihood

Reason 1: Disentangling Parameters

If we can change the order then on RHS we have:

$$\mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] = \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z})\right] - \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log q(\boldsymbol{z})\right]$$

Working on Log Likelihood

Reason 1: Disentangling Parameters

If we can change the order then on RHS we have:

$$\mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] = \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z})\right] - \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log q(\boldsymbol{z})\right]$$

- The optimization over model parameters $\boldsymbol{\theta}$ is done using the first term.

Working on Log Likelihood

Reason 1: Disentangling Parameters

If we can change the order then on RHS we have:

$$\mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] = \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z})\right] - \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log q(\boldsymbol{z})\right]$$

- The optimization over model parameters $\boldsymbol{\theta}$ is done using the first term.
- The prospective optimization over q functions should still be done using both terms.

Working on Log Likelihood

Reason 2: Two Separate Learning Signals

For the log-likelihood, we have:

$$\log p_\theta(\boldsymbol{x}) = \log \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] \simeq \log\left[\frac{1}{k}\sum_k\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z}_i)}{q(\boldsymbol{z}_i)}\right]\right],\ \ \boldsymbol{z}_i \sim q(\mathbb{Z})$$

Working on Log Likelihood

Reason 2: Two Separate Learning Signals

For the log-likelihood, we have:

$$\log p_\theta(\boldsymbol{x}) = \log \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] \simeq \log\left[\frac{1}{k}\sum_k\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z}_i)}{q(\boldsymbol{z}_i)}\right]\right],\ \boldsymbol{z}_i \sim q(\mathbb{Z})$$

If we can change the order then on RHS we have:

Working on Log Likelihood

Reason 2: Two Separate Learning Signals

For the log-likelihood, we have:

$$\log p_\theta(\boldsymbol{x}) = \log \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] \simeq \log\left[\frac{1}{k}\sum_k \left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z}_i)}{q(\boldsymbol{z}_i)}\right]\right], \;\; \boldsymbol{z}_i \sim q(\mathbb{Z})$$

If we can change the order then on RHS we have:

$$\mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] = \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z})\right] - \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log q(\boldsymbol{z})\right]$$

Working on Log Likelihood

Reason 2: Two Separate Learning Signals

For the log-likelihood, we have:

$$\log p_\theta(\boldsymbol{x}) = \log \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] \simeq \log\left[\frac{1}{k}\sum_k\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z}_i)}{q(\boldsymbol{z}_i)}\right]\right],\ \boldsymbol{z}_i \sim q(\mathbb{Z})$$

If we can change the order then on RHS we have:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] &= \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z})\right] - \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log q(\boldsymbol{z})\right] \\ &\simeq \frac{1}{k}\sum_k \log p_\theta(\boldsymbol{x},\boldsymbol{z}_i) - \frac{1}{k}\sum_k \log q(\boldsymbol{z}_i),\ \boldsymbol{z}_i \sim q(\mathbb{Z})\end{aligned}$$

Working on Log Likelihood

Reason 2: Two Separate Learning Signals

For the log-likelihood, we have:

$$\log p_\theta(\boldsymbol{x}) = \log \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] \simeq \log\left[\frac{1}{k}\sum_k\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z}_i)}{q(\boldsymbol{z}_i)}\right]\right],\ \boldsymbol{z}_i \sim q(\mathbb{Z})$$

If we can change the order then on RHS we have:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] &= \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z})\right] - \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log q(\boldsymbol{z})\right] \\ &\simeq \frac{1}{k}\sum_k \log p_\theta(\boldsymbol{x},\boldsymbol{z}_i) - \frac{1}{k}\sum_k \log q(\boldsymbol{z}_i),\ \boldsymbol{z}_i \sim q(\mathbb{Z})\end{aligned}$$

Because we are estimating the expectations with MCE, then in the second case, the learning signal from each of the terms is separated leading to a better approximation.

Working on Log Likelihood

Challenge

The challenge is to relate the following terms:

$$\text{We have} \Rightarrow \log \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] \boxed{\begin{matrix}\leq \\ = \\ \geq\end{matrix}} \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] \Leftarrow \text{We want}$$

Working on Log Likelihood

Challenge

The challenge is to relate the following terms:

$$\text{We have} \Rightarrow \log \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] \boxed{\begin{matrix}\leq\\=\\\geq\end{matrix}} \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] \Leftarrow \text{We want}$$

Jensen's Inequality

Assume random vector $\mathbb{X}$ and concave function $\psi(\cdot)$. Then:

$$\psi\left(\mathbb{E}[\mathbb{X}]\right) \geq \mathbb{E}\left[\psi(X)\right]$$

Evidence Lower BOund (ELBO)

The Log function is concave. So using Jensen's inequality, we have:

Evidence Lower BOund

Evidence Lower BOund (ELBO)

The Log function is concave. So using Jensen's inequality, we have:

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

Evidence Lower BOund

Evidence Lower BOund (ELBO)

The Log function is concave. So using Jensen's inequality, we have:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}) &= \log \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] \\ &\geq \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log\left(\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right)\right]\end{aligned}$$

Evidence Lower BOund

Evidence Lower BOund (ELBO)

The Log function is concave. So using Jensen's inequality, we have:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}) =& \log \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] \\ \geq& \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log\left(\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right)\right]\end{aligned}$$

So by changing the order of log function and expectation, we get a lower bound for the log-likelihood known as *Evidence Lower BOund* or abbreviately ELBO.

Evidence Lower BOund

Evidence Lower BOund (ELBO)

The Log function is concave. So using Jensen's inequality, we have:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}) =& \log \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] \\ \geq& \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log\left(\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right)\right]\end{aligned}$$

So by changing the order of log function and expectation, we get a lower bound for the log-likelihood known as *Evidence Lower BOund* or abbreviately ELBO.

Notation

ELBO is defined for one sample $\boldsymbol{x}$, while the model parameters $\boldsymbol{\theta}$ and q distributions can be changed. Thus we use the following notation:

$$\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},q) = \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\log\left(\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right)\right]$$

ELBO

We can work on ELBO as:

$$\log p_\theta(\boldsymbol{x}) \geq \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$$

ELBO

ELBO

We can work on ELBO as:

$$\log p_\theta(\boldsymbol{x}) \geq \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$$

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

ELBO

We can work on ELBO as:

$$
\begin{aligned}
\log p_\theta(\boldsymbol{x}) \geq & \operatorname{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q) \\
= & \mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})}\left[\log\left(\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})}\right)\right] \\
= & \mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x}, \boldsymbol{z})\right] - \mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})}\left[\log q(\boldsymbol{z})\right]
\end{aligned}
$$

ELBO

We can work on ELBO as:

$$\begin{aligned}
\log p_\theta(\boldsymbol{x}) \geq & \operatorname{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q) \\
= & \mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})} \left[\log \left(\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right) \right] \\
= & \mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})} \left[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) \right] - \mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})} \left[\log q(\boldsymbol{z}) \right] \\
= & \mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})} \left[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) \right] + \mathbb{H}(q)
\end{aligned}$$

ELBO

We can work on ELBO as:

$$
\begin{aligned}
\log p_\theta(\boldsymbol{x}) \geq & \,\mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q) \\
= & \,\mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})}\left[\log\left(\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})}\right)\right] \\
= & \,\mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x}, \boldsymbol{z})\right] - \mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})}\left[\log q(\boldsymbol{z})\right] \\
= & \,\mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x}, \boldsymbol{z})\right] + \mathbb{H}(q)
\end{aligned}
$$

where $\mathbb{H}(q)$, called *distribution entropy*, is a well-known quantity in information theory.

ELBO Instead of Data Log-likelihood

Replacing Data Log-likelihood with ELBO

To train a model, we need to solve the following optimization:

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p_\theta(\boldsymbol{x})$$

ELBO Instead of Data Log-likelihood

Replacing Data Log-likelihood with ELBO

To train a model, we need to solve the following optimization:

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p_\theta(\boldsymbol{x})$$

We can replace the above above objective with its lower bound, $\mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$.

ELBO Instead of Data Log-likelihood

Replacing Data Log-likelihood with ELBO

To train a model, we need to solve the following optimization:

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p_\theta(\boldsymbol{x})$$

We can replace the above above objective with its lower bound, $\mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$. Hopefully maximizing the lower bound leads to maximizing the data log-likelihood. So:

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$$

ELBO Instead of Data Log-likelihood

Replacing Data Log-likelihood with ELBO

To train a model, we need to solve the following optimization:

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p_{\theta}(\boldsymbol{x})$$

We can replace the above above objective with its lower bound, $\mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$. Hopefully maximizing the lower bound leads to maximizing the data log-likelihood. So:

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$$

Tightness Matters

As we are replacing $\log p_{\theta}(\boldsymbol{x})$ with $\mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$, the tightness of the lower bound becomes important.

ELBO Tightness

Reframing ELBO

Based on KL divergence, we have:

$$\mathrm{KL}(q(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x})) = \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} \mathrm{d}\boldsymbol{z}$$

ELBO Tightness

Reframing ELBO

Based on KL divergence, we have:

$$\mathrm{KL}(q(\mathbb{Z})\|p_\theta(\mathbb{Z}|\boldsymbol{x})) = \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} \mathrm{d}\boldsymbol{z}$$

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

ELBO Tightness

Reframing ELBO

Based on KL divergence, we have:

$$
\begin{aligned}
\mathrm{KL}(q(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x})) &= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} \mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{p_\theta(\boldsymbol{x})}} \mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \left[\log p_\theta(\boldsymbol{x}) - \log p_\theta(\boldsymbol{x}, \boldsymbol{z}) + \log q(z)\right] \mathrm{d}\boldsymbol{z}
\end{aligned}
$$

ELBO Tightness

Reframing ELBO

Based on KL divergence, we have:

$$
\begin{aligned}
\mathrm{KL}(q(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x})) &= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} \mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{p_\theta(\boldsymbol{x})}} \mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \left[\log p_\theta(\boldsymbol{x}) - \log p_\theta(\boldsymbol{x}, \boldsymbol{z}) + \log q(z)\right] \mathrm{d}\boldsymbol{z} \\
&= \log p_\theta(\boldsymbol{x}) \int_{\boldsymbol{z}} q(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} - \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log p_\theta(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} + \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log q(z) \mathrm{d}\boldsymbol{z}
\end{aligned}
$$

ELBO Tightness

Reframing ELBO

Based on KL divergence, we have:

$$
\begin{aligned}
\mathrm{KL}(q(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x})) &= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} \mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{p_\theta(\boldsymbol{x})}} \mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \left[\log p_\theta(\boldsymbol{x}) - \log p_\theta(\boldsymbol{x}, \boldsymbol{z}) + \log q(z)\right] \mathrm{d}\boldsymbol{z} \\
&= \log p_\theta(\boldsymbol{x}) \int_{\boldsymbol{z}} q(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} - \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log p_\theta(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} + \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log q(z) \mathrm{d}\boldsymbol{z} \\
&= \underbrace{\log p_\theta(\boldsymbol{x})}_{\text{Loh-likelihood}} - \Bigg[\underbrace{\mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x}, \boldsymbol{z})\right] + \mathbb{H}(q)}_{\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},q)}\Bigg]
\end{aligned}
$$

ELBO Tightness

Tightest ELBO

So we have the following interesting equality:

$$\mathrm{KL}(q(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x})) = \log p_\theta(\boldsymbol{x}) - \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$$

Thus the gap is $\mathrm{KL}(q(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$.

ELBO Tightness

Tightest ELBO

So we have the following interesting equality:

$$\mathrm{KL}(q(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x})) = \log p_\theta(\boldsymbol{x}) - \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$$

Thus the gap is $\mathrm{KL}(q(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$.

- As the distance between $q(\mathbb{Z})$ and $p_\theta(\mathbb{Z}|\mathbf{x})$ increases, the ELBO becomes a less tight bound.

ELBO Tightness

Tightest ELBO

So we have the following interesting equality:

$$\mathrm{KL}(q(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x})) = \log p_\theta(\boldsymbol{x}) - \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$$

Thus the gap is $\mathrm{KL}(q(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$.

- As the distance between $q(\mathbb{Z})$ and $p_\theta(\mathbb{Z}|\mathbf{x})$ increases, the ELBO becomes a less tight bound.
- When you select $q(\mathbb{Z}) = p_\theta(\mathbb{Z}|\mathbf{x})$, ELBO touches model log Likelihood (as we see before).

ELBO Tightness

Tightest ELBO

So we have the following interesting equality:

$$\mathrm{KL}(q(\mathbb{Z})\|p_\theta(\mathbb{Z}|\boldsymbol{x})) = \log p_\theta(\boldsymbol{x}) - \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},q)$$

Thus the gap is $\mathrm{KL}(q(\mathbb{Z})\|p_\theta(\mathbb{Z}|\boldsymbol{x}))$.

- As the distance between $q(\mathbb{Z})$ and $p_\theta(\mathbb{Z}|\mathbf{x})$ increases, the ELBO becomes a less tight bound.
- When you select $q(\mathbb{Z}) = p_\theta(\mathbb{Z}|\mathbf{x})$, ELBO touches model log Likelihood (as we see before).

So for $q(\mathbb{Z}) = p_\theta(\mathbb{Z}|\boldsymbol{x})$, we have:

$$\mathrm{KL}(q(\mathbb{Z})\|p_\theta(\mathbb{Z}|\boldsymbol{x})) = 0 \Rightarrow \log p_\theta(\boldsymbol{x}) = \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},q)$$

ELBO Tightness

Tightest ELBO

So we have the following interesting equality:

$$\mathrm{KL}(q(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x})) = \log p_\theta(\boldsymbol{x}) - \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$$

Thus the gap is $\mathrm{KL}(q(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$.

- As the distance between $q(\mathbb{Z})$ and $p_\theta(\mathbb{Z}|\mathbf{x})$ increases, the ELBO becomes a less tight bound.
- When you select $q(\mathbb{Z}) = p_\theta(\mathbb{Z}|\mathbf{x})$, ELBO touches model log Likelihood (as we see before).

So for $q(\mathbb{Z}) = p_\theta(\mathbb{Z}|\boldsymbol{x})$, we have:

$$\mathrm{KL}(q(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x})) = 0 \Rightarrow \log p_\theta(\boldsymbol{x}) = \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$$

Consequently, we have found the $q(\mathbb{Z})$ distribution resulting in the tightest ELBO.

Update on our Challenges

Challenge 3

We have no formal derivation for the suitability of $p_\theta(\mathbb{Z}|\boldsymbol{x})$ to be used for importance sampling in the general real-world case (other than GMM). So we don't know how to select $q(\mathbb{Z})$

Update on our Challenges

Challenge 3

We have no formal derivation for the suitability of $p_\theta(\mathbb{Z}|\boldsymbol{x})$ to be used for importance sampling in the general real-world case (other than GMM). So we don't know how to select $q(\mathbb{Z})$

- ☞ $p_\theta(\mathbb{Z}|\boldsymbol{x})$ is the best option which leads ELBO touch data log-likelihood.

Update on our Challenges

Challenge 3

We have no formal derivation for the suitability of $p_\theta(\mathbb{Z}|\boldsymbol{x})$ to be used for importance sampling in the general real-world case (other than GMM). So we don't know how to select $q(\mathbb{Z})$

☞ $p_\theta(\mathbb{Z}|\boldsymbol{x})$ is the best option which leads ELBO touch data log-likelihood.

Challenge 4

Calculating $p_\theta(\mathbb{Z}|\boldsymbol{x})$ needs the data likelihood which is intractable as we observe in Slide 28. On the other hand, we need to decrease $\mathrm{KL}(q(\mathbb{Z})\|p_\theta(\mathbb{Z}|\boldsymbol{x}))$ to tighten the ELBO.

Section 6

ELBO Tightening

ELBO Tightening

Parameterizing q

To have control over distribution $q(\mathbb{Z})$, we should parameterize this distribution. An example can be:

- $q_\lambda(\mathbb{Z}) = \mathcal{N}(\mathbb{Z}|\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \sigma^2\}$

ELBO Tightening

Parameterizing q

To have control over distribution $q(\mathbb{Z})$, we should parameterize this distribution. An example can be:

- $q_\lambda(\mathbb{Z}) = \mathcal{N}(\mathbb{Z}|\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ and $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \sigma^2\}$

Note that we are currently focused on the likelihood of one sample point $\boldsymbol{x}$ and thus $\boldsymbol{\lambda}$ is the parameter of the distribution corresponding to $\boldsymbol{x}$.

ELBO Tightening

Parameterizing q

To have control over distribution $q(\mathbb{Z})$, we should parameterize this distribution. An example can be:

- $q_\lambda(\mathbb{Z}) = \mathcal{N}(\mathbb{Z}|\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ and $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \sigma^2\}$

Note that we are currently focused on the likelihood of one sample point $\boldsymbol{x}$ and thus $\boldsymbol{\lambda}$ is the parameter of the distribution corresponding to $\boldsymbol{x}$.

Notation Update

By parameterizing q, instead of $\text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$ we use:

$$\text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{z} \sim q_\lambda(\mathbb{Z})} \left[\log \left(\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\lambda(\boldsymbol{z})} \right) \right]$$

ELBO Tightening

Parameterizing q

To have control over distribution $q(\mathbb{Z})$, we should parameterize this distribution. An example can be:

- $q_\lambda(\mathbb{Z}) = \mathcal{N}(\mathbb{Z}|\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \sigma^2\}$

Note that we are currently focused on the likelihood of one sample point $\boldsymbol{x}$ and thus $\boldsymbol{\lambda}$ is the parameter of the distribution corresponding to $\boldsymbol{x}$.

Notation Update

By parameterizing q, instead of $\text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$ we use:

$$\text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{z} \sim q_\lambda(\mathbb{Z})}\left[\log\left(\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\lambda(\boldsymbol{z})}\right)\right]$$

And again pay attention that everything is just about one data point $\boldsymbol{x}$.

ELBO Tightening

Optimization Problem

For ELBO tightening in a model with parameter $\boldsymbol{\theta}$, we need to solve the following optimization problem:

$$\boldsymbol{\lambda}^\star(\boldsymbol{\theta}) = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \, \mathrm{KL}(q_\lambda(\mathbb{Z}) \,\|\, p_\theta(\mathbb{Z}|\boldsymbol{x}))$$

But pay attention to the following note:

ELBO Tightening

Optimization Problem

For ELBO tightening in a model with parameter $\boldsymbol{\theta}$, we need to solve the following optimization problem:

$$\boldsymbol{\lambda}^{\star}(\boldsymbol{\theta}) = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \, \mathrm{KL}(q_{\lambda}(\mathbb{Z}) \,\|\, p_{\theta}(\mathbb{Z}|\boldsymbol{x}))$$

But pay attention to the following note:

- If you can find $\boldsymbol{\lambda}^{\star}(\boldsymbol{\theta})$, then it is a function of $\boldsymbol{\theta}$. Thus by changing the model parameter $\boldsymbol{\theta}$, $\boldsymbol{\lambda}^{\star}$ is not optimum anymore and you should update it.

ELBO Tightening

Optimization Problem

For ELBO tightening in a model with parameter $\boldsymbol{\theta}$, we need to solve the following optimization problem:

$$\boldsymbol{\lambda}^{\star}(\boldsymbol{\theta}) = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \, \mathrm{KL}(q_{\lambda}(\mathbb{Z}) \, \| \, p_{\theta}(\mathbb{Z}|\boldsymbol{x}))$$

But pay attention to the following note:

- If you can find $\boldsymbol{\lambda}^{\star}(\boldsymbol{\theta})$, then it is a function of $\boldsymbol{\theta}$. Thus by changing the model parameter $\boldsymbol{\theta}$, $\boldsymbol{\lambda}^{\star}$ is not optimum anymore and you should update it.

Challenge 4

Solving the above optimization seems intractable as calculating $p_{\theta}(\mathbb{Z}|\boldsymbol{x})$ is intractable in general.

ELBO Tightening

Intuition

Pay attention to the following equality:

$$\log p_\theta(\boldsymbol{x}) = \text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) + \text{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$$

Two important points:

ELBO Tightening

Intuition

Pay attention to the following equality:

$$\log p_\theta(\boldsymbol{x}) = \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) + \mathrm{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$$

Two important points:

- LHS is independent of $\boldsymbol{\lambda}$.

ELBO Tightening

Intuition

Pay attention to the following equality:

$$\log p_\theta(\boldsymbol{x}) = \text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) + \text{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$$

Two important points:

- LHS is independent of $\boldsymbol{\lambda}$.
- Both terms on the RHS are functions of $\boldsymbol{\lambda}$.

ELBO Tightening

Intuition

Pay attention to the following equality:

$$\log p_\theta(\boldsymbol{x}) = \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) + \mathrm{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$$

Two important points:

- LHS is independent of $\boldsymbol{\lambda}$.
- Both terms on the RHS are functions of $\boldsymbol{\lambda}$.

☞ How can we decrese the KL divergence between $q_\lambda(\mathbb{Z})$ and $p_\theta(\mathbb{Z}|\boldsymbol{x})$?

ELBO Tightening

Tightening Solution

We know:

$$\log p_\theta(\boldsymbol{x}) = \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) + \mathrm{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$$

Thus:

ELBO Tightening

Tightening Solution

We know:

$$\log p_\theta(\boldsymbol{x}) = \text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) + \text{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$$

Thus:

$$\text{Because } \log p_\theta(\boldsymbol{x}) \text{ is independent of } \boldsymbol{\lambda} \Rightarrow \begin{cases} \text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) \ \uparrow \\ \\ \text{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x})) \ \downarrow \end{cases}$$

ELBO Tightening

Tightening Solution

We know:

$$\log p_\theta(\boldsymbol{x}) = \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) + \mathrm{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$$

Thus:

$$\text{Because } \log p_\theta(\boldsymbol{x}) \text{ is independent of } \boldsymbol{\lambda} \Rightarrow \begin{cases} \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) \ \uparrow \\ \mathrm{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x})) \ \downarrow \end{cases}$$

In other words for a fixed model $\boldsymbol{\theta}$:

$$\underset{\boldsymbol{\lambda}}{\operatorname{argmax}} \ \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) \equiv \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \ \mathrm{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$$

ELBO Tightening

Tightening Solution

We know:

$$\log p_\theta(\boldsymbol{x}) = \text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) + \text{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$$

Thus:

$$\text{Because } \log p_\theta(\boldsymbol{x}) \text{ is independent of } \boldsymbol{\lambda} \Rightarrow \begin{cases} \text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) \ \uparrow \\ \\ \text{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x})) \ \downarrow \end{cases}$$

In other words for a fixed model $\boldsymbol{\theta}$:

$$\underset{\boldsymbol{\lambda}}{\operatorname{argmax}} \ \text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) \equiv \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \ \text{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$$

Thus we can tighten the ELBO by maximizing it with respect to $\boldsymbol{\lambda}$.

ELBO Tightening

ELBO Tightening

Assume the case of simple GMM with two components as:

$$p_\theta(\boldsymbol{x}) = \sum_{z\in\{0,1\}} p_\theta(Z=z)p_\theta(\boldsymbol{x}|Z=z), \quad \begin{cases} p_\theta(Z=z) = \pi_Z \\ p_\theta(\boldsymbol{x}|Z=z) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) \end{cases}$$

So the model parameters are $\boldsymbol{\theta} = \{\pi_0, \pi_1 . \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \mu_1, \boldsymbol{\Sigma}_1\}$ and given.

ELBO Tightening

ELBO Tightening

Assume the case of simple GMM with two components as:

$$p_\theta(\boldsymbol{x}) = \sum_{z \in \{0,1\}} p_\theta(Z = z) p_\theta(\boldsymbol{x}|Z = z), \quad \begin{cases} p_\theta(Z = z) = \pi_Z \\ p_\theta(\boldsymbol{x}|Z = z) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) \end{cases}$$

So the model parameters are $\boldsymbol{\theta} = \{\pi_0, \pi_1.\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \mu_1, \boldsymbol{\Sigma}_1\}$ and given.
In this example, we want to show the following problems are equivalent:

$$\underset{\lambda}{\operatorname{argmax}} \;\; \text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \lambda) \equiv \underset{\lambda}{\operatorname{argmin}} \;\; \text{KL}(q_\lambda(Z) \| p_\theta(Z|\boldsymbol{x}))$$

We will follow by solving each optimization problem separately.

ELBO Tightening

RHS Optimization

We are interested in the following optimization:

$$\underset{\lambda}{\operatorname{argmin}} \;\; \mathrm{KL}(q_\lambda(Z) \| p_\theta(Z|\boldsymbol{x}))$$

ELBO Tightening

RHS Optimization

We are interested in the following optimization:

$$\underset{\lambda}{\text{argmin}} \;\; \mathrm{KL}(q_\lambda(Z) \| p_\theta(Z|\boldsymbol{x}))$$

- Because Z is binary thus:
    - $q_\lambda(Z) = \mathrm{Ber}(Z|\lambda)$

ELBO Tightening

RHS Optimization

We are interested in the following optimization:

$$\underset{\lambda}{\operatorname{argmin}} \ \ \mathrm{KL}(q_\lambda(Z) \| p_\theta(Z|\boldsymbol{x}))$$

- Because Z is binary thus:
  - $q_\lambda(Z) = \mathrm{Ber}(Z|\lambda)$
  - $p_\theta(Z|\boldsymbol{x}) = \mathrm{Ber}(Z|\beta(\boldsymbol{\theta}))$ (we write $\beta(\boldsymbol{\theta})$ to emphasize that $p_\theta(Z|\boldsymbol{x})$ has no new parameter and its parameter is a function of $\boldsymbol{\theta}$)

ELBO Tightening

RHS Optimization

We are interested in the following optimization:

$$\underset{\lambda}{\operatorname{argmin}} \; \mathrm{KL}(q_\lambda(Z) \| p_\theta(Z|\boldsymbol{x}))$$

- Because Z is binary thus:
  - $q_\lambda(Z) = \mathrm{Ber}(Z|\lambda)$
  - $p_\theta(Z|\boldsymbol{x}) = \mathrm{Ber}(Z|\beta(\boldsymbol{\theta}))$ (we write $\beta(\boldsymbol{\theta})$ to emphasize that $p_\theta(Z|\boldsymbol{x})$ has no new parameter and its parameter is a function of $\boldsymbol{\theta}$)
- The optimum λ based on RHS is:

$$\lambda^\star = \beta(\boldsymbol{\theta})$$

ELBO Tightening

RHS Optimization

We are interested in the following optimization:

$$\underset{\lambda}{\operatorname{argmin}} \ \mathrm{KL}(q_\lambda(Z) \| p_\theta(Z|\boldsymbol{x}))$$

- Because Z is binary thus:
  - $q_\lambda(Z) = \mathrm{Ber}(Z|\lambda)$
  - $p_\theta(Z|\boldsymbol{x}) = \mathrm{Ber}(Z|\beta(\boldsymbol{\theta}))$ (we write $\beta(\boldsymbol{\theta})$ to emphasize that $p_\theta(Z|\boldsymbol{x})$ has no new parameter and its parameter is a function of $\boldsymbol{\theta}$)
- The optimum λ based on RHS is:

$$\lambda^\star = \beta(\boldsymbol{\theta})$$

So we need to find $\beta(\boldsymbol{\theta}) = p_\theta(Z = 1|\boldsymbol{x})$

ELBO Tightening

RHS Optimization

$$\beta(\boldsymbol{\theta}) = p_\theta(Z = 1|\boldsymbol{x})$$

ELBO Tightening

RHS Optimization

$$\beta(\boldsymbol{\theta}) = p_\theta(Z = 1|\boldsymbol{x})$$

$$= \frac{p_\theta(\boldsymbol{x}, Z = 1)}{p_\theta(\boldsymbol{x})} \qquad \text{\#Bayes Rule}$$

ELBO Tightening

RHS Optimization

$$\beta(\boldsymbol{\theta}) = p_\theta(Z = 1|\boldsymbol{x})$$

$$= \frac{p_\theta(\boldsymbol{x}, Z = 1)}{p_\theta(\boldsymbol{x})} \quad \text{\#Bayes Rule}$$

$$= \frac{p_\theta(\boldsymbol{x}, Z = 1)}{\sum_{z \in \{0,1\}} p_\theta(\boldsymbol{x}, Z = z)} \quad \text{\#Marginalization}$$

RHS Optimization

$$
\begin{aligned}
\beta(\boldsymbol{\theta}) &= p_\theta(Z=1|\boldsymbol{x}) \\
&= \frac{p_\theta(\boldsymbol{x}, Z=1)}{p_\theta(\boldsymbol{x})} && \text{\#Bayes Rule} \\
&= \frac{p_\theta(\boldsymbol{x}, Z=1)}{\sum_{z\in\{0,1\}} p_\theta(\boldsymbol{x}, Z=z)} && \text{\#Marginalization} \\
&= \frac{p_\theta(\boldsymbol{x}|Z=1)p_\theta(Z=1)}{\sum_{z\in\{0,1\}} p_\theta(\boldsymbol{x}|Z=z)p_\theta(Z=z)} && \text{\#Conditioning}
\end{aligned}
$$

ELBO Tightening

RHS Optimization

$$\beta(\boldsymbol{\theta}) = p_\theta(Z = 1|\boldsymbol{x})$$

$$= \frac{p_\theta(\boldsymbol{x}, Z = 1)}{p_\theta(\boldsymbol{x})} \qquad \text{\#Bayes Rule}$$

$$= \frac{p_\theta(\boldsymbol{x}, Z = 1)}{\sum_{z\in\{0,1\}} p_\theta(\boldsymbol{x}, Z = z)} \qquad \text{\#Marginalization}$$

$$= \frac{p_\theta(\boldsymbol{x}|Z = 1)p_\theta(Z = 1)}{\sum_{z\in\{0,1\}} p_\theta(\boldsymbol{x}|Z = z)p_\theta(Z = z)} \qquad \text{\#Conditioning}$$

$$= \frac{\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\pi_1}{\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\pi_0 + \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\pi_1} = \lambda^\star$$

ELBO Tightening

RHS Optimization

Assume the model parameters are:

$$\boldsymbol{\theta} = \left\{ \begin{array}{lll} \pi_1 = 0.33 & \mu_1 = \begin{bmatrix} 4.04 \\ 3.83 \end{bmatrix} & \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.79 & -0.10 \\ -0.10 & 2.00 \end{bmatrix} \\ & & \\ \pi_0 = 0.67 & \mu_0 = \begin{bmatrix} 1.10 \\ 0.86 \end{bmatrix} & \boldsymbol{\Sigma}_0 = \begin{bmatrix} 1.20 & -0.97 \\ -0.97 & 1.15 \end{bmatrix} \end{array} \right\} \lambda^\star = \beta(\boldsymbol{\theta}) = 0.76$$

ELBO Tightening

RHS Optimization

Assume the model parameters are:

$$\boldsymbol{\theta} = \left\{ \begin{array}{ll} \pi_1 = 0.33 \quad \mu_1 = \begin{bmatrix} 4.04 \\ 3.83 \end{bmatrix} \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.79 & -0.10 \\ -0.10 & 2.00 \end{bmatrix} \\ \\ \pi_0 = 0.67 \quad \mu_0 = \begin{bmatrix} 1.10 \\ 0.86 \end{bmatrix} \boldsymbol{\Sigma}_0 = \begin{bmatrix} 1.20 & -0.97 \\ -0.97 & 1.15 \end{bmatrix} \end{array} \right\} \lambda^\star = \beta(\boldsymbol{\theta}) = 0.76$$

then using the equation in Slide 48 we have:

ELBO Tightening

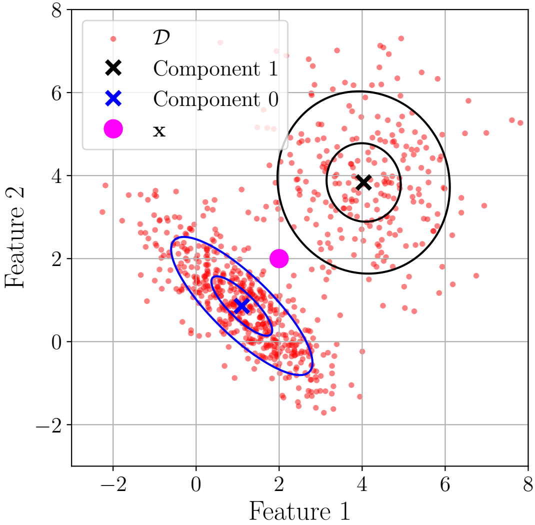


Figure: Dataset, GMM components and query point $\boldsymbol{x}$

ELBO Tightening

LHS Optimization

Assume distribution $q_\lambda(Z) = \mathrm{Ber}(Z|\lambda)$, then:

$$\mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \lambda) = \mathbb{E}_{z \sim q_\lambda(Z)} \log \left[\frac{p_\theta(\boldsymbol{x}, z)}{q_\lambda(z)} \right]$$

ELBO Tightening

LHS Optimization

Assume distribution $q_\lambda(Z) = \mathrm{Ber}(Z|\lambda)$, then:

$$\begin{aligned}
\mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \lambda) =& \mathbb{E}_{z \sim q_\lambda(Z)} \log \left[\frac{p_\theta(\boldsymbol{x}, z)}{q_\lambda(z)} \right] \\
=& q_\lambda(1) \log \left[\frac{p_\theta(\boldsymbol{x}, 1)}{q_\lambda(1)} \right] + q_\lambda(0) \log \left[\frac{p_\theta(\boldsymbol{x}, 0)}{q_\lambda(0)} \right]
\end{aligned}$$

ELBO Tightening

LHS Optimization

Assume distribution $q_\lambda(Z) = \text{Ber}(Z|\lambda)$, then:

$$
\begin{aligned}
\text{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\lambda) =& \mathbb{E}_{z\sim q_\lambda(Z)} \log\left[\frac{p_\theta(\boldsymbol{x},z)}{q_\lambda(z)}\right] \\
=& q_\lambda(1)\log\left[\frac{p_\theta(\boldsymbol{x},1)}{q_\lambda(1)}\right] + q_\lambda(0)\log\left[\frac{p_\theta(\boldsymbol{x},0)}{q_\lambda(0)}\right] \\
=& \lambda\log\left[\frac{\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\pi_1}{\lambda}\right] + (1-\lambda)\log\left[\frac{\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)\pi_0}{1-\lambda}\right]
\end{aligned}
$$

ELBO Tightening

LHS Optimization

Assume distribution $q_\lambda(Z) = \text{Ber}(Z|\lambda)$, then:

$$\begin{aligned}
\text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \lambda) =& \mathbb{E}_{z\sim q_\lambda(Z)} \log \left[\frac{p_\theta(\boldsymbol{x}, z)}{q_\lambda(z)}\right] \\
=& q_\lambda(1) \log \left[\frac{p_\theta(\boldsymbol{x}, 1)}{q_\lambda(1)}\right] + q_\lambda(0) \log \left[\frac{p_\theta(\boldsymbol{x}, 0)}{q_\lambda(0)}\right] \\
=& \lambda \log \left[\frac{\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\pi_1}{\lambda}\right] + (1-\lambda) \log \left[\frac{\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\pi_0}{1-\lambda}\right]
\end{aligned}$$

Assuming fixed model parameter $\boldsymbol{\theta}$, we can plot $\text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \lambda)$ as a function of λ.

ELBO Tightening

LHS Optimization

We know $p_\theta(Z|\boldsymbol{x}) = \mathrm{Ber}(Z|0.76)$ and $q_\lambda(Z) = \mathrm{Ber}(Z|\lambda)$, so:

ELBO Tightening

LHS Optimization

We know $p_\theta(Z|\boldsymbol{x}) = \mathrm{Ber}(Z|0.76)$ and $q_\lambda(Z) = \mathrm{Ber}(Z|\lambda)$, so:

$$\mathrm{KL}(q_\lambda(Z) \| p_\theta(Z|\boldsymbol{x})) = \mathbb{E}_{z\sim q_\lambda(Z)} \log \frac{q_\lambda(z)}{p_\theta(z|\boldsymbol{x})}$$

ELBO Tightening

LHS Optimization

We know $p_\theta(Z|\boldsymbol{x}) = \mathrm{Ber}(Z|0.76)$ and $q_\lambda(Z) = \mathrm{Ber}(Z|\lambda)$, so:

$$\begin{aligned}\mathrm{KL}(q_\lambda(Z)\|p_\theta(Z|\boldsymbol{x})) =& \mathbb{E}_{z\sim q_\lambda(Z)} \log \frac{q_\lambda(z)}{p_\theta(z|\boldsymbol{x})} \\ =& q_\lambda(1)\log\frac{q_\lambda(1)}{p_\theta(1|\boldsymbol{x})} + q_\lambda(0)\log\frac{q_\lambda(0)}{p_\theta(0|\boldsymbol{x})}\end{aligned}$$

ELBO Tightening

LHS Optimization

We know $p_\theta(Z|\boldsymbol{x}) = \text{Ber}(Z|0.76)$ and $q_\lambda(Z) = \text{Ber}(Z|\lambda)$, so:

$$\begin{aligned}
\text{KL}(q_\lambda(Z)\|p_\theta(Z|\boldsymbol{x})) =& \mathbb{E}_{z\sim q_\lambda(Z)} \log \frac{q_\lambda(z)}{p_\theta(z|\boldsymbol{x})} \\
=& q_\lambda(1) \log \frac{q_\lambda(1)}{p_\theta(1|\boldsymbol{x})} + q_\lambda(0) \log \frac{q_\lambda(0)}{p_\theta(0|\boldsymbol{x})} \\
=& \lambda \log \frac{\lambda}{0.76} + (1-\lambda) \log \frac{1-\lambda}{(1-0.76)}
\end{aligned}$$

ELBO Tightening

LHS Optimization

We know $p_\theta(Z|\boldsymbol{x}) = \text{Ber}(Z|0.76)$ and $q_\lambda(Z) = \text{Ber}(Z|\lambda)$, so:

$$\begin{aligned}
\text{KL}(q_\lambda(Z)\|p_\theta(Z|\boldsymbol{x})) =& \mathbb{E}_{z\sim q_\lambda(Z)} \log \frac{q_\lambda(z)}{p_\theta(z|\boldsymbol{x})} \\
=& q_\lambda(1) \log \frac{q_\lambda(1)}{p_\theta(1|\boldsymbol{x})} + q_\lambda(0) \log \frac{q_\lambda(0)}{p_\theta(0|\boldsymbol{x})} \\
=& \lambda \log \frac{\lambda}{0.76} + (1-\lambda) \log \frac{1-\lambda}{(1-0.76)}
\end{aligned}$$

In Slide 48, we have also calculate $p_\theta(\boldsymbol{x})$.

ELBO Tightening

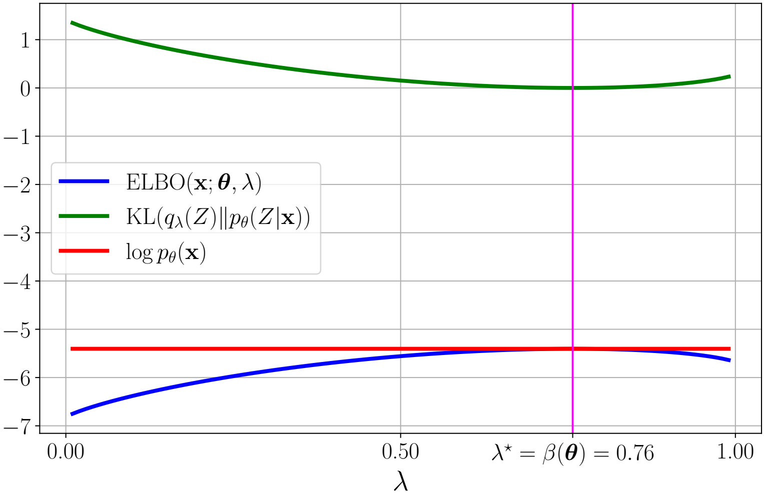


Figure: $\text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \lambda)$, $\text{KL}(q_\lambda(Z) \| p_\theta(Z|\boldsymbol{x}))$ and $\log p_\theta(\boldsymbol{x})$ as a function of λ

Altogether

Challenge 1 $\Rightarrow$ $$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} \mathcal{N}\big(\boldsymbol{x}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\big)\, p_\theta(\boldsymbol{z})\mathrm{d}\boldsymbol{z}$$

Altogether

Challenge 1

$$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\right) p_\theta(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}$$

Solution 1

$$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim p_\theta(\mathbb{Z})}\left[p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$$

Altogether

Challenge 1 $\rightarrow$ $p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} \mathcal{N}\big(\boldsymbol{x}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\big)\, p_\theta(\boldsymbol{z})\mathrm{d}\boldsymbol{z}$

$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim p_\theta(\mathbb{Z})}\left[p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$ $\leftarrow$ Solution 1

Challenge 2 $\rightarrow$ High Variance in Estimation with MCE

Altogether

Challenge 1

$$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} \mathcal{N}\left(\boldsymbol{x} | \boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\right) p_\theta(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}$$

Solution 1

$$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z} \sim p_\theta(\mathbb{Z})}\left[p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$$

Challenge 2

High Variance in Estimation with MCE

Solution 2

$$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})}\left[\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})}\right]$$

Altogether

Challenge 1 → $p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\right) p_\theta(\boldsymbol{z})\mathrm{d}\boldsymbol{z}$

$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim p_\theta(\mathbb{Z})}\left[p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$ ← Solution 1

Challenge 2 → High Variance in Estimation with MCE

$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right]$ ← Solution 2

Challenge 3 → Selecting $q(\mathbb{Z})$

Altogether

Challenge 1 → $p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\right) p_\theta(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}$

$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim p_\theta(\mathbb{Z})}\left[p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$ ← Solution 1

Challenge 2 → High Variance in Estimation with MCE

$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right]$ ← Solution 2

Challenge 3 → Selecting $q(\mathbb{Z})$

$\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},q)$ ← Solution 3

Altogether

Challenge 1 $\rightarrow$ $p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\right) p_\theta(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}$

$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z} \sim p_\theta(\mathbb{Z})}\left[p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$ $\leftarrow$ Solution 1

Challenge 2 $\rightarrow$ High Variance in Estimation with MCE

$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z} \sim q(\mathbb{Z})}\left[\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})}\right]$ $\leftarrow$ Solution 2

Challenge 3 $\rightarrow$ Selecting $q(\mathbb{Z})$

$\mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, q)$

$q(\mathbb{Z}) = p_\theta(\mathbb{Z}|\boldsymbol{x})$ $\leftarrow$ Solution 3

Altogether

Challenge 1 $p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\right) p_\theta(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}$

$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim p_\theta(\mathbb{Z})}\left[p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$ Solution 1

Challenge 2 High Variance in Estimation with MCE

$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right]$ Solution 2

Challenge 3 Selecting $q(\mathbb{Z})$

$\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},q)$

$q(\mathbb{Z}) = p_\theta(\mathbb{Z}|\boldsymbol{x})$ Solution 3

Challenge 4 Calulating $p_\theta(\mathbb{Z}|\boldsymbol{x})$

Altogether

Challenge 1 $\rightarrow$ $p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\right) p_\theta(\boldsymbol{z})\mathrm{d}\boldsymbol{z}$

$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim p_\theta(\mathbb{Z})}\left[p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$ $\leftarrow$ Solution 1

Challenge 2 $\rightarrow$ High Variance in Estimation with MCE

$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right]$ $\leftarrow$ Solution 2

Challenge 3 $\rightarrow$ Selecting $q(\mathbb{Z})$

$\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},q)$

$q(\mathbb{Z}) = p_\theta(\mathbb{Z}|\boldsymbol{x})$ $\leftarrow$ Solution 3

Challenge 4 $\rightarrow$ Calulating $p_\theta(\mathbb{Z}|\boldsymbol{x})$

$\min_\lambda \mathrm{KL}(q_\lambda(\mathbb{Z})\|p_\theta(\mathbb{Z}|\boldsymbol{x}))$ $\leftarrow$ Solution 4

Altogether

Challenge 1 → $p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\right) p_\theta(\boldsymbol{z})\mathrm{d}\boldsymbol{z}$

$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim p_\theta(\mathbb{Z})}\left[p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$ ← Solution 1

Challenge 2 → High Variance in Estimation with MCE

$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right]$ ← Solution 2

Challenge 3 → Selecting $q(\mathbb{Z})$

$\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},q)$

$q(\mathbb{Z}) = p_\theta(\mathbb{Z}|\boldsymbol{x})$ ← Solution 3

Challenge 4 → Calulating $p_\theta(\mathbb{Z}|\boldsymbol{x})$

$$\min_\lambda \mathrm{KL}(q_\lambda(\mathbb{Z})\|p_\theta(\mathbb{Z}|\boldsymbol{x})) \equiv \max_\lambda \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda})$$ ← Solution 4

Altogether

Challenge 1 → $p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z})\right) p_\theta(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}$

Solution 1 → $p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim p_\theta(\mathbb{Z})}\left[p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$

Challenge 2 → High Variance in Estimation with MCE

Solution 2 → $p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim q(\mathbb{Z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right]$

Challenge 3 → Selecting $q(\mathbb{Z})$

Solution 3 → $\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},q)$

$q(\mathbb{Z}) = p_\theta(\mathbb{Z}|\boldsymbol{x})$

Challenge 4 → Calulating $p_\theta(\mathbb{Z}|\boldsymbol{x})$

Solution 4 → $\min_\lambda \mathrm{KL}(q_\lambda(\mathbb{Z}) \| p_\theta(\mathbb{Z}|\boldsymbol{x}))$

$\equiv$

$\max_\lambda \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda})$

$$\boxed{\boxed{\max_{\boldsymbol{\lambda}}}\ \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda})}$$

Update on our Challenges

Challenge 4

Calculating $p_\theta(\mathbb{Z}|\boldsymbol{x})$ needs the data likelihood which is intractable as we observe in Slide 28. On the other hand, we need to decrease $\mathrm{KL}(q(\mathbb{Z})\|p_\theta(\mathbb{Z}|\boldsymbol{x}))$ to tighten the ELBO.

Update on our Challenges

Challenge 4

Calculating $p_\theta(\mathbb{Z}|\boldsymbol{x})$ needs the data likelihood which is intractable as we observe in Slide 28. On the other hand, we need to decrease $\mathrm{KL}(q(\mathbb{Z})\|p_\theta(\mathbb{Z}|\boldsymbol{x}))$ to tighten the ELBO.

- ☞ We can tighten the ELBO using the following optimization problem:

$$\max_{\boldsymbol{\lambda}} \ \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda})$$

Update on our Challenges

Challenge 4

Calculating $p_\theta(\mathbb{Z}|\boldsymbol{x})$ needs the data likelihood which is intractable as we observe in Slide 28. On the other hand, we need to decrease $\mathrm{KL}(q(\mathbb{Z})\|p_\theta(\mathbb{Z}|\boldsymbol{x}))$ to tighten the ELBO.

☞ We can tighten the ELBO using the following optimization problem:

$$\max_{\boldsymbol{\lambda}} \ \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda})$$

Challenge 5

We need to maximize ELBO with respect to both $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. Thus we need the following gradient vectors:

- $\nabla_\theta \,\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda})$
- $\nabla_\lambda \,\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda})$

Section 7

ELBO Optimization

ELBO Gradients

With Respect to $\boldsymbol{\theta}$

For the gradient w.r.t to $\boldsymbol{\theta}$, we have:

$$\nabla_\theta \operatorname{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) = \nabla_\theta \left(\mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\right]\right)$$

ELBO Gradients

With Respect to $\boldsymbol{\theta}$

For the gradient w.r.t to $\boldsymbol{\theta}$, we have:

$$\begin{aligned}
\nabla_\theta \operatorname{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) =& \nabla_\theta \left(\mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\right]\right) \\
\overset{a}{=}& \mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\nabla_\theta \left(\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\right)\right]
\end{aligned}$$

ELBO Gradients

With Respect to $\boldsymbol{\theta}$

For the gradient w.r.t to $\boldsymbol{\theta}$, we have:

$$\begin{aligned}
\nabla_\theta \operatorname{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) =& \nabla_\theta \left(\mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\right]\right) \\
\overset{a}{=}& \mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\nabla_\theta \left(\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\right)\right] \\
\overset{b}{=}& \mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\nabla_\theta \log p_\theta(\boldsymbol{x},\boldsymbol{z})\right]
\end{aligned}$$

ELBO Gradients

With Respect to $\boldsymbol{\theta}$

For the gradient w.r.t to $\boldsymbol{\theta}$, we have:

$$
\begin{aligned}
\nabla_\theta \operatorname{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) =& \nabla_\theta \left(\mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\right]\right) \\
\overset{a}{=}& \mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\nabla_\theta \left(\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\right)\right] \\
\overset{b}{=}& \mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\nabla_\theta \log p_\theta(\boldsymbol{x},\boldsymbol{z})\right] \\
\overset{c}{\simeq}& \frac{1}{K}\sum_{k=1}^{K} \nabla_\theta \log p_\theta(\boldsymbol{x},\boldsymbol{z}_k)
\end{aligned}
$$

ELBO Gradients

With Respect to $\boldsymbol{\theta}$

For the gradient w.r.t to $\boldsymbol{\theta}$, we have:

$$\begin{aligned}
\nabla_\theta \, \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) =& \nabla_\theta \left(\mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\right]\right) \\
\stackrel{a}{=}& \mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\nabla_\theta \left(\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\right)\right] \\
\stackrel{b}{=}& \mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\nabla_\theta \log p_\theta(\boldsymbol{x},\boldsymbol{z})\right] \\
\stackrel{c}{\simeq}& \frac{1}{K}\sum_{k=1}^{K} \nabla_\theta \log p_\theta(\boldsymbol{x},\boldsymbol{z}_k)
\end{aligned}$$

where:

- $a \Rightarrow$ Expectation is with respect to a distribution independent of $\boldsymbol{\theta}$
- $b \Rightarrow q_\lambda(\boldsymbol{z})$ is independent of $\boldsymbol{\theta}$
- $c \Rightarrow$ Monte-Carlo estimation

ELBO Gradients

With Respect to $\boldsymbol{\lambda}$

For the gradient w.r.t to $\boldsymbol{\lambda}$, we have:

$$\nabla_\lambda \, \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) = \nabla_\lambda \left(\mathbb{E}_{q_\lambda(Z)} \left[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q_\lambda(\boldsymbol{z}) \right] \right)$$

ELBO Gradients

With Respect to $\boldsymbol{\lambda}$

For the gradient w.r.t to $\boldsymbol{\lambda}$, we have:

$$\nabla_\lambda \operatorname{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) = \nabla_\lambda \left(\mathbb{E}_{q_\lambda(Z)}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_\lambda(\boldsymbol{z})\right]\right)$$

Note that here we can not change the order of gradient and expectation as the distribution of expectation is a function of $\boldsymbol{\lambda}$.

Update on our Challenges

Challenge 5

Calculating $p_\theta(\mathbb{Z}|\boldsymbol{x})$ needs the data likelihood which is intractable as we observe in Slide 28. On the other hand, we need to decrease $\mathrm{KL}(q(\mathbb{Z})\|p_\theta(\mathbb{Z}|\boldsymbol{x}))$ to tighten the ELBO.

Update on our Challenges

Challenge 5

Calculating $p_\theta(\mathbb{Z}|\boldsymbol{x})$ needs the data likelihood which is intractable as we observe in Slide 28. On the other hand, we need to decrease $\mathrm{KL}(q(\mathbb{Z})\|p_\theta(\mathbb{Z}|\boldsymbol{x}))$ to tighten the ELBO.

- ☞ We can tighten the ELBO using the following optimization problem:

$$\max_{\boldsymbol{\lambda}} \ \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda})$$

Update on our Challenges

Challenge 5

Calculating $p_\theta(\mathbb{Z}|\boldsymbol{x})$ needs the data likelihood which is intractable as we observe in Slide 28. On the other hand, we need to decrease $\mathrm{KL}(q(\mathbb{Z})\|p_\theta(\mathbb{Z}|\boldsymbol{x}))$ to tighten the ELBO.

☞ We can tighten the ELBO using the following optimization problem:

$$\max_{\boldsymbol{\lambda}} \ \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda})$$

Challenge 6

We need to maximize ELBO with respect to $\boldsymbol{\lambda}$. Thus we need the following gradient vector:

- $\nabla_\lambda \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda})$

Section 8

Reparameterization Trick

Reparameterization Trick

Reparameterization Trick

Again pay attention to the following gradient.

$$\nabla_\lambda \,\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) = \nabla_\lambda \left(\mathbb{E}_{q_\lambda(Z)}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_\lambda(\boldsymbol{z})\right]\right)$$

Reparameterization Trick

Again pay attention to the following gradient.

$$\nabla_\lambda \operatorname{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) = \nabla_\lambda \left(\mathbb{E}_{q_\lambda(Z)}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_\lambda(\boldsymbol{z})\right]\right)$$

To handle the above challenge we use the following steps:

Reparameterization Trick

Again pay attention to the following gradient.

$$\nabla_\lambda \, \text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) = \nabla_\lambda \left(\mathbb{E}_{q_\lambda(Z)} \left[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q_\lambda(\boldsymbol{z}) \right] \right)$$

To handle the above challenge we use the following steps:

- Limit the $q_\lambda(\mathbb{Z})$ to Gaussian distribtuion as:

$$q_\lambda(\mathbb{Z}) = \mathcal{N}(\mathbb{Z} | \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \Rightarrow \boldsymbol{\lambda} = \{\boldsymbol{\mu}, \sigma^2\}$$

Reparameterization Trick

Reparameterization Trick

Again pay attention to the following gradient.

$$\nabla_\lambda \, \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) = \nabla_\lambda \left(\mathbb{E}_{q_\lambda(Z)}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_\lambda(\boldsymbol{z})\right]\right)$$

To handle the above challenge we use the following steps:

- Limit the $q_\lambda(\mathbb{Z})$ to Gaussian distribtuion as:

$$q_\lambda(\mathbb{Z}) = \mathcal{N}(\mathbb{Z}|\boldsymbol{\mu},\sigma^2\boldsymbol{I}) \Rightarrow \boldsymbol{\lambda} = \{\boldsymbol{\mu},\sigma^2\}$$

- Reparameterize the random vector as:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I}) \Rightarrow \boldsymbol{z} = \boldsymbol{\mu} + \sigma\boldsymbol{\epsilon} = g(\boldsymbol{\epsilon};\boldsymbol{\lambda}) \sim \mathcal{N}(\mathbb{Z}\boldsymbol{\mu},\sigma^2\boldsymbol{I}) = q_\lambda(\mathbb{Z})$$

ELBO Gradients

With Respect to $\boldsymbol{\lambda}$

For the gradient w.r.t to $\boldsymbol{\lambda}$, we have:

$$\nabla_\lambda \, \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) = \nabla_\lambda \left(\mathbb{E}_{\boldsymbol{z} \sim q_\lambda(\mathbb{Z})} \left[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z}) \right] \right)$$

ELBO Gradients

With Respect to $\boldsymbol{\lambda}$

For the gradient w.r.t to $\boldsymbol{\lambda}$, we have:

$$\begin{aligned}\nabla_\lambda \, \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) =& \nabla_\lambda \left(\mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\right]\right)\\ \overset{a}{=}& \nabla_\lambda \left(\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\log p_\theta(\boldsymbol{x},g(\boldsymbol{\epsilon};\boldsymbol{\lambda})) - \log q_{\boldsymbol{\lambda}}(g(\boldsymbol{\epsilon};\boldsymbol{\lambda}))\right]\right)\end{aligned}$$

ELBO Gradients

With Respect to $\boldsymbol{\lambda}$

For the gradient w.r.t to $\boldsymbol{\lambda}$, we have:

$$\begin{aligned}
\nabla_\lambda \operatorname{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) =& \nabla_\lambda \left(\mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\right]\right) \\
\stackrel{a}{=}& \nabla_\lambda \left(\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\log p_\theta(\boldsymbol{x}, g(\boldsymbol{\epsilon};\boldsymbol{\lambda})) - \log q_{\boldsymbol{\lambda}}(g(\boldsymbol{\epsilon};\boldsymbol{\lambda}))\right]\right) \\
\stackrel{b}{=}& \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\nabla_\lambda \left(\log p_\theta(\boldsymbol{x}, g(\boldsymbol{\epsilon};\boldsymbol{\lambda})) - \log q_{\boldsymbol{\lambda}}(g(\boldsymbol{\epsilon};\boldsymbol{\lambda}))\right)\right]
\end{aligned}$$

ELBO Gradients

With Respect to $\boldsymbol{\lambda}$

For the gradient w.r.t to $\boldsymbol{\lambda}$, we have:

$$\begin{aligned}
\nabla_\lambda \,\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) =& \nabla_\lambda \left(\mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\right]\right) \\
\overset{a}{=}& \nabla_\lambda \left(\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\log p_\theta(\boldsymbol{x}, g(\boldsymbol{\epsilon};\boldsymbol{\lambda})) - \log q_{\boldsymbol{\lambda}}(g(\boldsymbol{\epsilon};\boldsymbol{\lambda}))\right]\right) \\
\overset{b}{=}& \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\nabla_\lambda \left(\log p_\theta(\boldsymbol{x}, g(\boldsymbol{\epsilon};\boldsymbol{\lambda})) - \log q_{\boldsymbol{\lambda}}(g(\boldsymbol{\epsilon};\boldsymbol{\lambda}))\right)\right] \\
\overset{c}{=}& \frac{1}{K}\sum_{k=1}^{K} \nabla_\lambda \left(\log p_\theta(\boldsymbol{x}, g(\boldsymbol{\epsilon}_k;\boldsymbol{\lambda})) - \log q_{\boldsymbol{\lambda}}(g(\boldsymbol{\epsilon}_k;\boldsymbol{\lambda}))\right)
\end{aligned}$$

ELBO Gradients

With Respect to $\boldsymbol{\lambda}$

For the gradient w.r.t to $\boldsymbol{\lambda}$, we have:

$$\begin{aligned}
\nabla_\lambda \operatorname{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda}) =& \nabla_\lambda \left(\mathbb{E}_{\boldsymbol{z}\sim q_\lambda(\mathbb{Z})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\right]\right)\\
\overset{a}{=}& \nabla_\lambda \left(\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\log p_\theta(\boldsymbol{x},g(\boldsymbol{\epsilon};\boldsymbol{\lambda})) - \log q_{\boldsymbol{\lambda}}(g(\boldsymbol{\epsilon};\boldsymbol{\lambda}))\right]\right)\\
\overset{b}{=}& \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\nabla_\lambda\left(\log p_\theta(\boldsymbol{x},g(\boldsymbol{\epsilon};\boldsymbol{\lambda})) - \log q_{\boldsymbol{\lambda}}(g(\boldsymbol{\epsilon};\boldsymbol{\lambda}))\right)\right]\\
\overset{c}{=}& \frac{1}{K}\sum_{k=1}^{K}\nabla_\lambda\left(\log p_\theta(\boldsymbol{x},g(\boldsymbol{\epsilon}_k;\boldsymbol{\lambda})) - \log q_{\boldsymbol{\lambda}}(g(\boldsymbol{\epsilon}_k;\boldsymbol{\lambda}))\right)
\end{aligned}$$

where:

- $a \Rightarrow$ Reparameterization trick
- $b \Rightarrow q_\lambda(\boldsymbol{z})$ Expectation is with respect to a distribution independent of $\boldsymbol{\lambda}$
- $c \Rightarrow$ Monte-Carlo estimation

ELBO Optimization

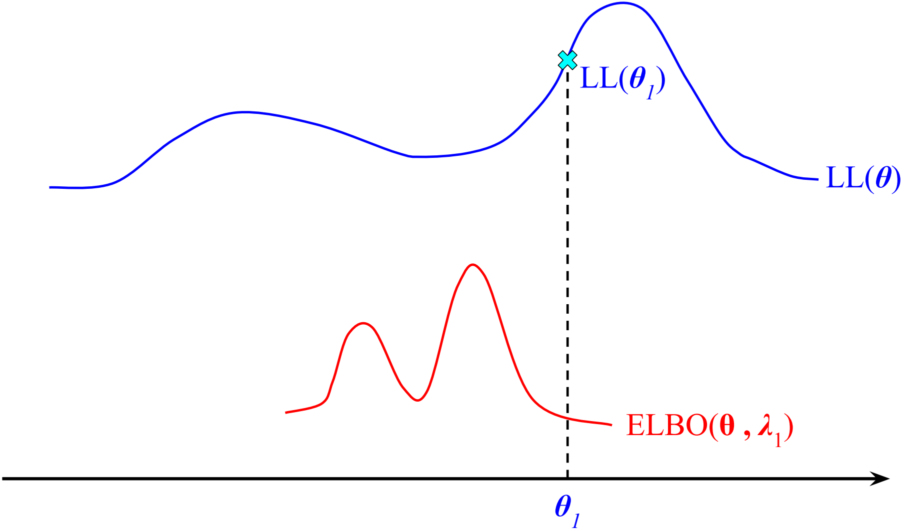


Figure: Starting point $\boldsymbol{\theta}_1$ and $\boldsymbol{\lambda}_1$

ELBO Optimization

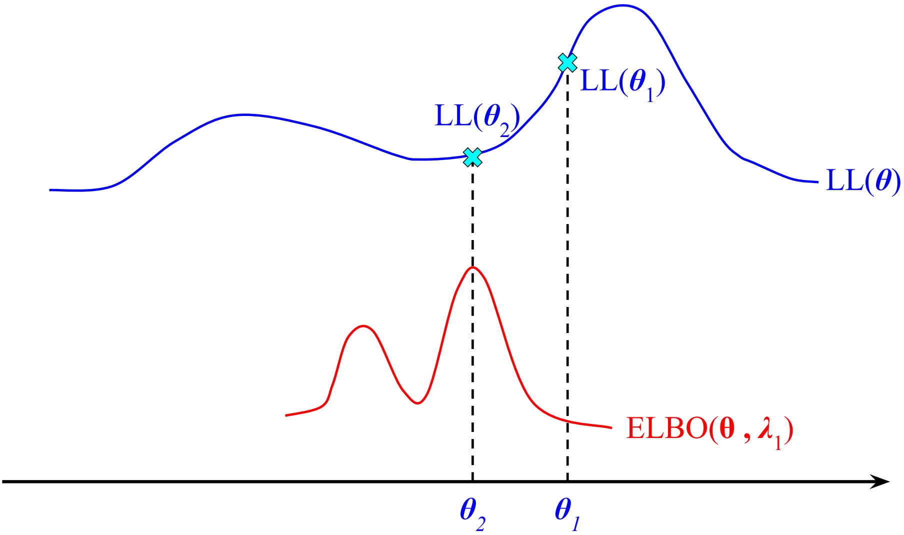


Figure: $\boldsymbol{\theta}_2 = \text{argmax}_{\boldsymbol{\theta}}\, \text{ELBO}(\boldsymbol{\theta}, \boldsymbol{\lambda}_1)$

ELBO Optimization

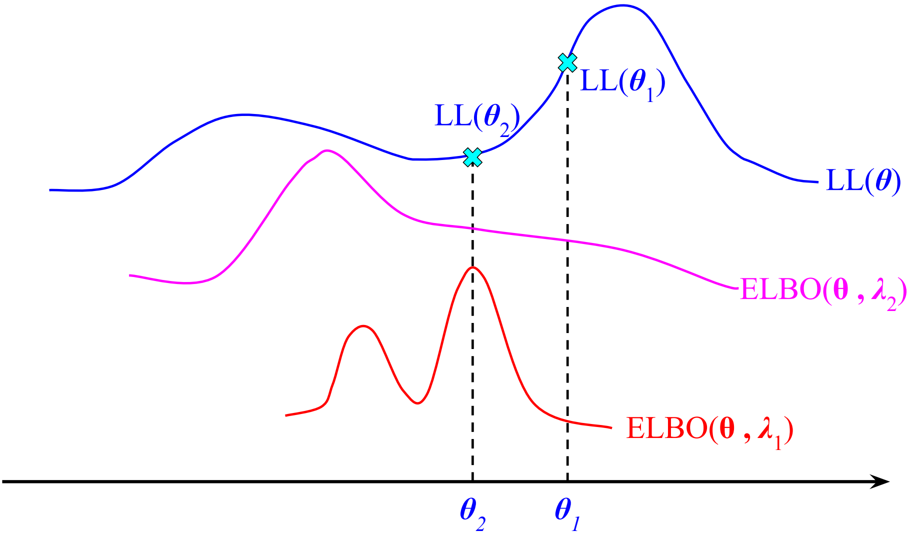


Figure: $\boldsymbol{\lambda}_2 = \mathrm{argmax}_{\boldsymbol{\lambda}}\, \mathrm{ELBO}(\boldsymbol{\theta}_2, \boldsymbol{\lambda})$

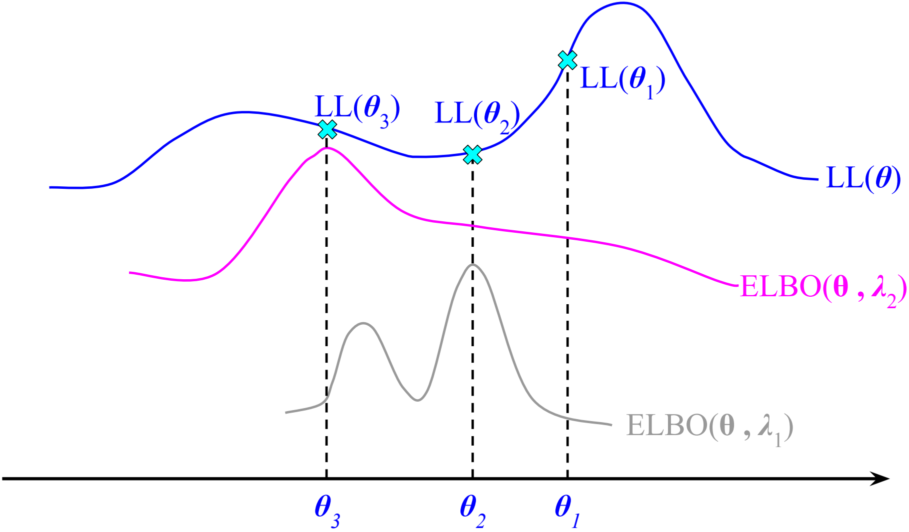


Figure: $\boldsymbol{\theta}_3 = \operatorname{argmax}_{\boldsymbol{\theta}} \mathrm{ELBO}(\boldsymbol{\theta}, \boldsymbol{\lambda}_2)$

Update on our Challenges

Challenge 6

We need to maximize ELBO with respect to $\boldsymbol{\lambda}$. Thus we need the following gradient vector:

- $\nabla_\lambda \, \text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda})$

Update on our Challenges

Challenge 6

We need to maximize ELBO with respect to $\boldsymbol{\lambda}$. Thus we need the following gradient vector:

- $\nabla_\lambda \operatorname{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda})$

☞ We can use the reparameterization trick to calculate the above gradient.

Update on our Challenges

Challenge 6

We need to maximize ELBO with respect to $\boldsymbol{\lambda}$. Thus we need the following gradient vector:

- $\nabla_\lambda \, \text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda})$

☞ We can use the reparameterization trick to calculate the above gradient.

Challenge 7

To train a model, we need to maximize the log-likelihood on the dataset $\mathcal{D}$ not an isolated sample $\boldsymbol{x}$.

Section 9

Learning

From Data Point to Dataset

Till now, we found a practical approach to increase the ELBO for a typical sample $\boldsymbol{x}$ as:

$$\log p_\theta(\boldsymbol{x}) \geq \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda})$$

From Data Point to Dataset

Till now, we found a practical approach to increase the ELBO for a typical sample $\boldsymbol{x}$ as:

$$\log p_\theta(\boldsymbol{x}) \geq \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda})$$

Now assume we have a complete dataset $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^N$. We have the following optimization problem:

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}_i)$$

Lower Bounding the Objective

We have:

$$\log p_\theta(\boldsymbol{x}_i) \geq \underset{\boldsymbol{\lambda}_i}{\operatorname{argmax}}\ \mathrm{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

Learning

Lower Bounding the Objective

We have:

$$\log p_\theta(\boldsymbol{x}_i) \geq \underset{\boldsymbol{\lambda}_i}{\operatorname{argmax}}\ \mathrm{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

$$\Rightarrow \sum_i \log p_\theta(\boldsymbol{x}_i) \geq \sum_i \underset{\boldsymbol{\lambda}_i}{\operatorname{argmax}}\ \mathrm{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

Lower Bounding the Objective

We have:

$$\log p_\theta(\boldsymbol{x}_i) \geq \underset{\boldsymbol{\lambda}_i}{\operatorname{argmax}} \ \text{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

$$\Rightarrow \sum_i \log p_\theta(\boldsymbol{x}_i) \geq \sum_i \underset{\boldsymbol{\lambda}_i}{\operatorname{argmax}} \ \text{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

$$= \underset{\boldsymbol{\lambda}_1, \ldots \boldsymbol{\lambda}_N}{\operatorname{argmax}} \sum_i \ \text{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

Learning

Lower Bounding the Objective

We have:

$$\log p_\theta(\boldsymbol{x}_i) \geq \underset{\boldsymbol{\lambda}_i}{\operatorname{argmax}}\ \mathrm{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

$$\Rightarrow \sum_i \log p_\theta(\boldsymbol{x}_i) \geq \sum_i \underset{\boldsymbol{\lambda}_i}{\operatorname{argmax}}\ \mathrm{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

$$= \underset{\boldsymbol{\lambda}_1, \ldots \boldsymbol{\lambda}_N}{\operatorname{argmax}} \sum_i\ \mathrm{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

So:

$$\frac{1}{N} \sum_i \log p_\theta(\boldsymbol{x}_i) \geq \underset{\boldsymbol{\lambda}_1, \ldots \boldsymbol{\lambda}_N}{\operatorname{argmax}} \frac{1}{N} \sum_i\ \mathrm{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

Learning Problem

We see:

$$\frac{1}{N}\sum_i \log p_\theta(\boldsymbol{x}_i) \geq \underset{\boldsymbol{\lambda}_1,\dots\boldsymbol{\lambda}_N}{\operatorname{argmax}} \frac{1}{N}\sum_i \ \mathrm{ELBO}(\boldsymbol{x}_i;\boldsymbol{\theta},\boldsymbol{\lambda}_i)$$

Learning

Learning Problem

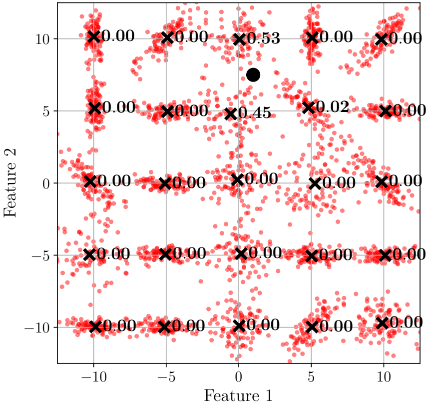
We see:

$$\frac{1}{N}\sum_i \log p_\theta(\boldsymbol{x}_i) \geq \underset{\boldsymbol{\lambda}_1,\ldots\boldsymbol{\lambda}_N}{\operatorname{argmax}} \frac{1}{N}\sum_i \ \mathrm{ELBO}(\boldsymbol{x}_i;\boldsymbol{\theta},\boldsymbol{\lambda}_i)$$

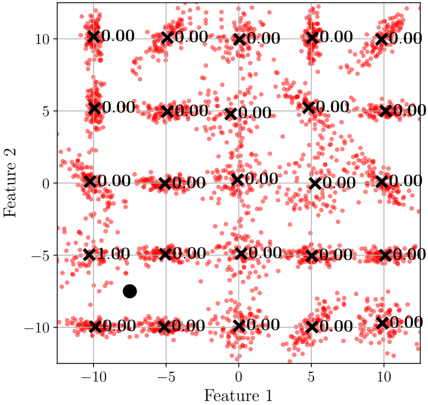
As a result, the learning problem is:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underset{\boldsymbol{\lambda}_1,\ldots\boldsymbol{\lambda}_N}{\operatorname{argmax}} \sum_i \ \mathrm{ELBO}(\boldsymbol{x}_i;\boldsymbol{\theta},\boldsymbol{\lambda}_i)$$

Different Sampling Distributions for Different Data Samples



(a) $q(Z|\boldsymbol{x}_1)$



(b) $q(Z|\boldsymbol{x}_2)$

Figure: The model posterior over latent variables for two different samples $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (so we need one distribution for each sample)

Stochastic Variational Inference (SVI) Learning

Algorithm 1: Stochastic Variational Learning

Input : Dataset $\mathcal{D} = \{\boldsymbol{x}_i\}$

Initialization: $\boldsymbol{\theta}, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N$

begin

    for $j = 1 : D$ **do**

        Select a Sample $\boldsymbol{x}_i$ from $\mathcal{D}$ randomly

        for $k = 1 : K$ **do**

            $\boldsymbol{\lambda}_i \leftarrow \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}_i} \text{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$

        end

        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \mu \nabla_{\theta} \text{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$

    end

end

Output : $\boldsymbol{\theta}$

Update on our Challenges

Challenge 7

To train a model, we need to maximize the log-likelihood on the dataset $\mathcal{D}$, not an isolated sample $\boldsymbol{x}$.

Update on our Challenges

Challenge 7

To train a model, we need to maximize the log-likelihood on the dataset $\mathcal{D}$, not an isolated sample $\boldsymbol{x}$.

☞ We derive the objective for the complete dataset as:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underset{\boldsymbol{\lambda}_1,\ldots\boldsymbol{\lambda}_N}{\operatorname{argmax}} \sum_i \operatorname{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

Update on our Challenges

Challenge 7

To train a model, we need to maximize the log-likelihood on the dataset $\mathcal{D}$, not an isolated sample $\boldsymbol{x}$.

☞ We derive the objective for the complete dataset as:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underset{\boldsymbol{\lambda}_1, \ldots \boldsymbol{\lambda}_N}{\operatorname{argmax}} \sum_i \operatorname{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

Challenge 8

As we can see in the above optimization problem, the number of parameters is not scalable with the dataset size. Each training sample adds one $\boldsymbol{\lambda}$ vector to the optimization problem.

Section 10

Amortization

Amortization

Scalability of SVI

Using the SVI, we can solve the following maximization problem:

$$\boldsymbol{\theta}^\star, \boldsymbol{\lambda}_1^\star, \boldsymbol{\lambda}_2^\star, \ldots, \boldsymbol{\lambda}_N^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underset{\boldsymbol{\lambda}_1, \ldots \boldsymbol{\lambda}_N}{\operatorname{argmax}} \sum_i \operatorname{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

Amortization

Scalability of SVI

Using the SVI, we can solve the following maximization problem:

$$\boldsymbol{\theta}^\star, \boldsymbol{\lambda}_1^\star, \boldsymbol{\lambda}_2^\star, \ldots, \boldsymbol{\lambda}_N^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underset{\boldsymbol{\lambda}_1, \ldots \boldsymbol{\lambda}_N}{\operatorname{argmax}} \sum_i \operatorname{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\lambda}_i)$$

Amortization is a type of parameter sharing to achieve scalability with respect to dataset size.

Amortization

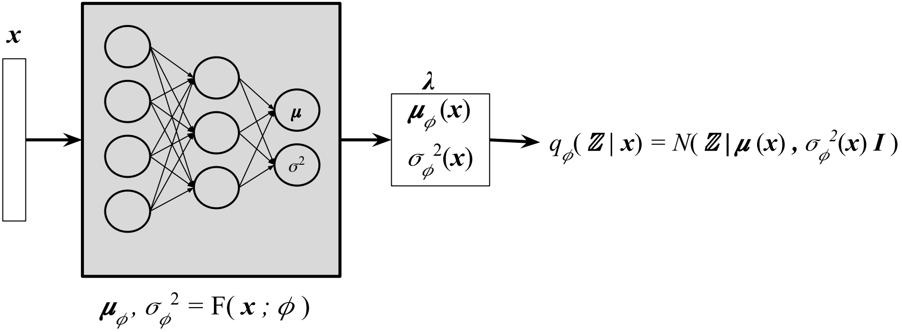


Figure: Amortization: The parameters of neural network $F(\cdot;\cdot)$ is shared between all the training samples.

Amortization

Notation Update

By using the Amortization technique, $\boldsymbol{\phi}$ the parameters of F are shared between all training data samples. So instead of $\text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda})$ we use:

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

Amortization

Notation Update

By using the Amortization technique, $\boldsymbol{\phi}$ the parameters of F are shared between all training data samples. So instead of $\text{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\lambda})$ we use:

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

Amortized Inference

Using Amortization, the ELBO for sample $\boldsymbol{x}_i$ can be written as:

$$\text{ELBO}(\boldsymbol{x}_i;\boldsymbol{\theta},\boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{z}\sim q_{\phi}(\mathbb{Z}|\boldsymbol{x}_i)}\left[\log p_{\theta}(\boldsymbol{x},\boldsymbol{z}) - \log q_{\phi}(\boldsymbol{z}|\boldsymbol{x}_i)\right]$$

Amortization

Algorithm 2: Amortized Stochastic Variational Inference

Input : Dataset $\mathcal{D} = \{\boldsymbol{x}_i\}$
Initialization: $\boldsymbol{\theta}, \boldsymbol{\phi}$
begin
  for $j = 1 : D$ **do**
    Select a Sample $\boldsymbol{x}_i$ from $\mathcal{D}$ randomly
    $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \eta \nabla_{\boldsymbol{\phi}} \mathrm{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\phi})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \mu \nabla_{\theta} \mathrm{ELBO}(\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\phi})$
  end
end
Output : $\boldsymbol{\theta}, \boldsymbol{\phi}$

Section 11

Revisiting Objective Function

Variational Autoencoder

Revisiting ELBO

We can reformulate ELBO as:

$$\text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{z} \sim q_\phi(\mathbb{Z}|\boldsymbol{x})} \left[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x}) \right]$$

Revisiting ELBO

We can reformulate ELBO as:

$$\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\mathbb{Z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x})\right]$$

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

Variational Autoencoder

Revisiting ELBO

We can reformulate ELBO as:

$$
\begin{aligned}
\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi}) &= \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\mathbb{Z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x})\right] \\
&= \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\mathbb{Z}|\boldsymbol{x})}\left[\underbrace{\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log p_\theta(\boldsymbol{z})}_{\log p_\theta(\boldsymbol{x}|\boldsymbol{z})} + \underbrace{\log p_\theta(\boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x})}_{-\log \frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z})}}\right] \\
&= \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\mathbb{Z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - \mathrm{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p_\theta(\boldsymbol{z})\right)
\end{aligned}
$$

Variational Autoencoder Overview

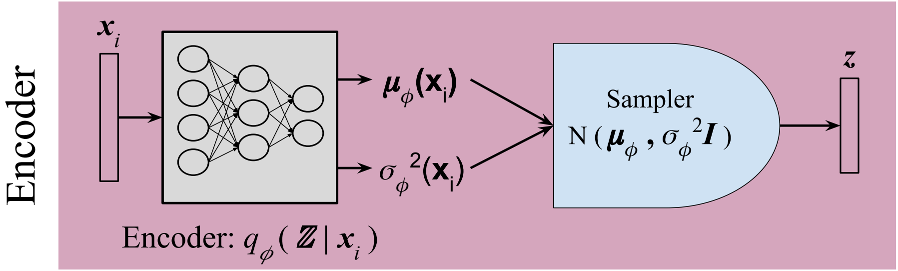


Figure: Encoding

Variational Autoencoder Overview

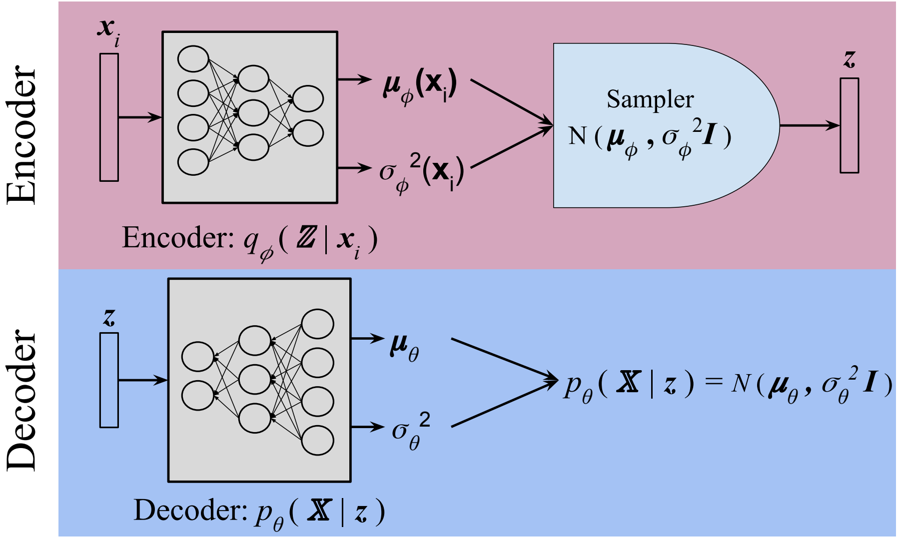


Figure: Encoding and Decoding

Revisiting ELBO

Let's focus on the first term:

$$\mathbb{E}_{\boldsymbol{z} \sim q_\phi(\mathbb{Z}|\boldsymbol{x}_i)} \left[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z})\right]$$

Variational Autoencoder

Revisiting ELBO

Let's focus on the first term:

$$\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\mathbb{Z}|\boldsymbol{x}_i)}\left[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z})\right]$$

Conceptually, we have:

- Select a training sample $\boldsymbol{x}_i$.

Variational Autoencoder

Revisiting ELBO

Let's focus on the first term:

$$\mathbb{E}_{\boldsymbol{z} \sim q_\phi(\mathbb{Z}|\boldsymbol{x}_i)} \left[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z})\right]$$

Conceptually, we have:

- Select a training sample $\boldsymbol{x}_i$.
- Generate the approximate posterior over the latent vector given $\boldsymbol{x}_i$

Revisiting ELBO

Let's focus on the first term:

$$\mathbb{E}_{\boldsymbol{z} \sim q_\phi(\mathbb{Z}|\boldsymbol{x}_i)} \left[\log p_\theta(\boldsymbol{x}_i | \boldsymbol{z}) \right]$$

Conceptually, we have:

- Select a training sample $\boldsymbol{x}_i$.
- Generate the approximate posterior over the latent vector given $\boldsymbol{x}_i$
- Sample $\boldsymbol{z}$ from the approximate posterior

Variational Autoencoder

Revisiting ELBO

Let's focus on the first term:

$$\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\mathbb{Z}|\boldsymbol{x}_i)}\left[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z})\right]$$

Conceptually, we have:

- Select a training sample $\boldsymbol{x}_i$.
- Generate the approximate posterior over the latent vector given $\boldsymbol{x}_i$
- Sample $\boldsymbol{z}$ from the approximate posterior
- Generate the model posterior over $\mathbb{X}$ given $\boldsymbol{z}$ or $p_\theta(\mathbb{X}|\boldsymbol{z})$

Variational Autoencoder

Revisiting ELBO

Let's focus on the first term:

$$\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\mathbb{Z}|\boldsymbol{x}_i)}\left[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z})\right]$$

Conceptually, we have:

- Select a training sample $\boldsymbol{x}_i$.
- Generate the approximate posterior over the latent vector given $\boldsymbol{x}_i$
- Sample $\boldsymbol{z}$ from the approximate posterior
- Generate the model posterior over $\mathbb{X}$ given $\boldsymbol{z}$ or $p_\theta(\mathbb{X}|\boldsymbol{z})$

A good model is one where $\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z})$ is maximized.

Revisiting ELBO

Let's focus on the first term:

$$\mathbb{E}_{\boldsymbol{z} \sim q_\phi(\mathbb{Z}|\boldsymbol{x}_i)} \left[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z}) \right]$$

Variational Autoencoder

Revisiting ELBO

Let's focus on the first term:

$$\mathbb{E}_{\boldsymbol{z} \sim q_\phi(\mathbb{Z}|\boldsymbol{x}_i)} \left[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z})\right]$$

In terms of formulations, we have:

$$\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z}) = \log \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_\theta, \sigma_\theta^2 \boldsymbol{I})$$

Variational Autoencoder

Revisiting ELBO

Let's focus on the first term:

$$\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\mathbb{Z}|\boldsymbol{x}_i)}\left[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z})\right]$$

In terms of formulations, we have:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z}) &= \log \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_\theta, \sigma_\theta^2\boldsymbol{I})\\ &= -\frac{D}{2}\log(2\pi\sigma_\theta^2) - \frac{1}{2\sigma_\theta^2}\|\boldsymbol{x}_i - \boldsymbol{\mu}_\theta\|_2^2\end{aligned}$$

Revisiting ELBO

Let's focus on the first term:

$$\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\mathbb{Z}|\boldsymbol{x}_i)}\left[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z})\right]$$

In terms of formulations, we have:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z}) &= \log \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_\theta, \sigma_\theta^2\boldsymbol{I}) \\ &= -\frac{D}{2}\log(2\pi\sigma_\theta^2) - \frac{1}{2\sigma_\theta^2}\|\boldsymbol{x}_i - \boldsymbol{\mu}_\theta\|_2^2\end{aligned}$$

Note that:

- The optimization w.r.t. $\boldsymbol{\theta}$ is straightforward.

Variational Autoencoder

Revisiting ELBO

Let's focus on the first term:

$$\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\mathbb{Z}|\boldsymbol{x}_i)}\left[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z})\right]$$

In terms of formulations, we have:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z}) &= \log \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_\theta, \sigma_\theta^2 \boldsymbol{I}) \\ &= -\frac{D}{2}\log(2\pi\sigma_\theta^2) - \frac{1}{2\sigma_\theta^2}\|\boldsymbol{x}_i - \boldsymbol{\mu}_\theta\|_2^2\end{aligned}$$

Note that:

- The optimization w.r.t. $\boldsymbol{\theta}$ is straightforward.
- The optimization w.r.t. $\boldsymbol{\phi}$ needs reparameterization trick.

Variational Autoencoder with Reparameterization Trick

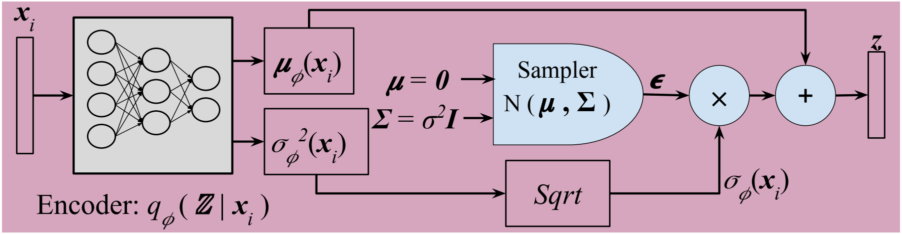


Figure: Encoding with Reparameterization Trick

Variational Autoencoder with Reparameterization Trick

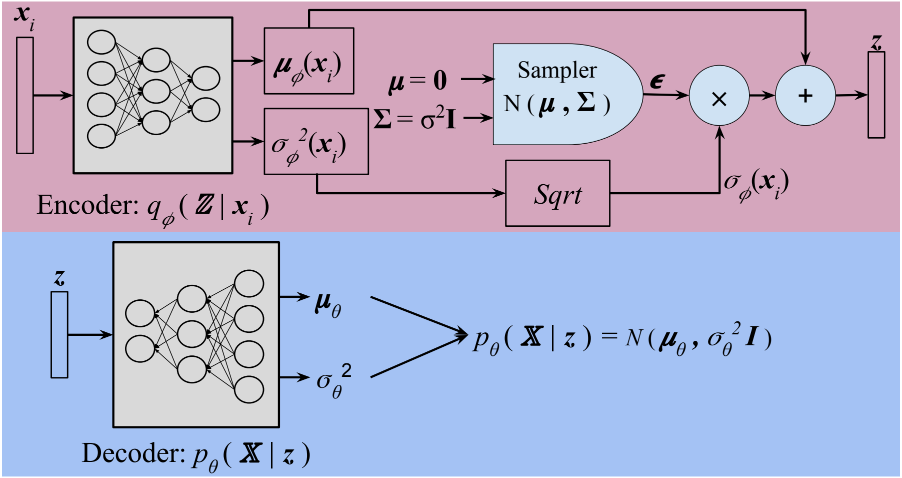


Figure: Encoding and Decoding

Revisiting ELBO

Now let's move on to the second term:

$$-\operatorname{KL}\left(q_\phi(\mathbb{Z}|\boldsymbol{x}_i)\|p_\theta(\mathbb{Z})\right)$$

Variational Autoencoder

Revisiting ELBO

Now let's move on to the second term:

$$-\,\mathrm{KL}\left(q_\phi(\mathbb{Z}|\boldsymbol{x}_i)\|p_\theta(\mathbb{Z})\right)$$

After training and at the end of the day:

- $q_\phi(\mathbb{Z}|\boldsymbol{x}_i), i = 1, \ldots, N$ be near the prior distribution $p_\theta(\mathbb{Z})$ in KLD sense.

Variational Autoencoder

Revisiting ELBO

Now let's move on to the second term:

$$-\operatorname{KL}\left(q_\phi(\mathbb{Z}|\boldsymbol{x}_i)\|p_\theta(\mathbb{Z})\right)$$

After training and at the end of the day:

- $q_\phi(\mathbb{Z}|\boldsymbol{x}_i), i = 1, \ldots, N$ be near the prior distribution $p_\theta(\mathbb{Z})$ in KLD sense.

What does it imply?

Variational Autoencoder

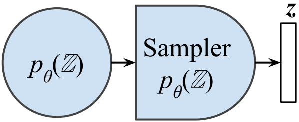


Figure: Encoding with Reparameterization Trick

Variational Autoencoder

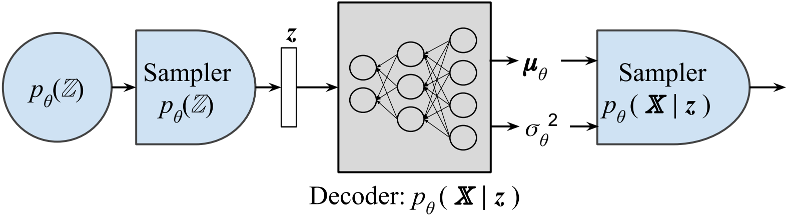


Figure: Encoding with Reparameterization Trick

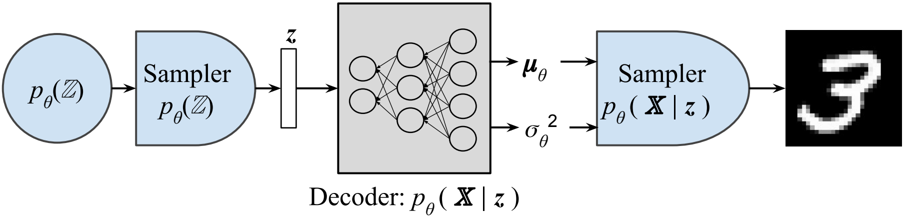


Figure: Encoding with Reparameterization Trick

Section 12

Attribute Vectors in Code Space

Concept

SW: Smiling Woman
NW: Neutral Woman
SM: Smiling Man
NM: Neutral Man

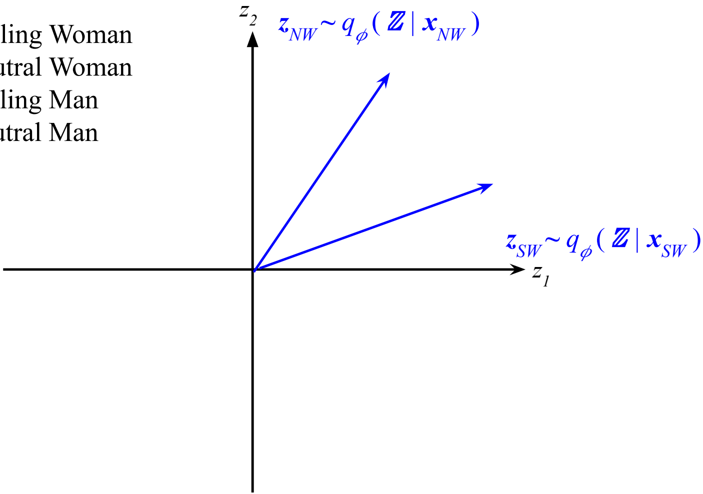


Figure: Encoding of *Neutral Woman* and *Smiling Woman* images

Concept

SW: Smiling Woman
NW: Neutral Woman
SM: Smiling Man
NM: Neutral Man

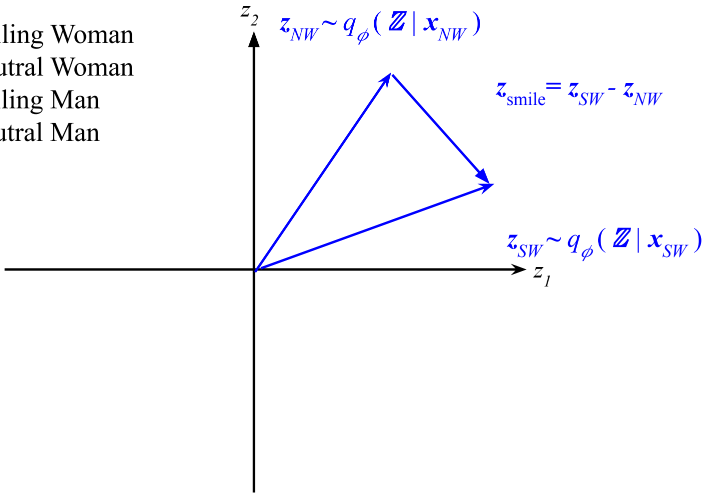


Figure: Calculating the attribute vector of *Smile* in latent space

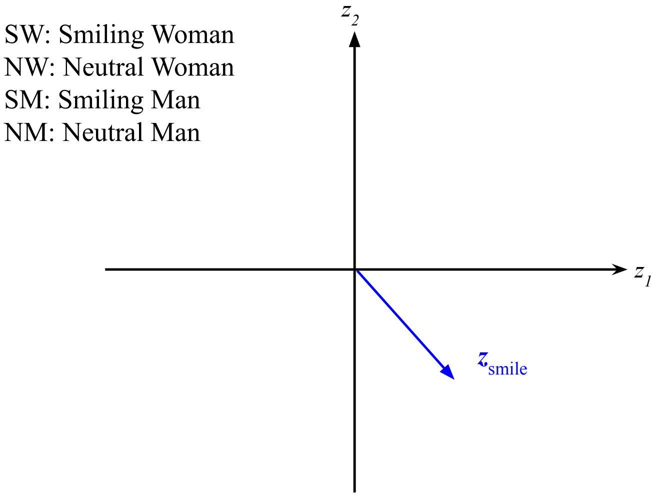


Figure: *Smile* attribute vector

Concept

SW: Smiling Woman
NW: Neutral Woman
SM: Smiling Man
NM: Neutral Man

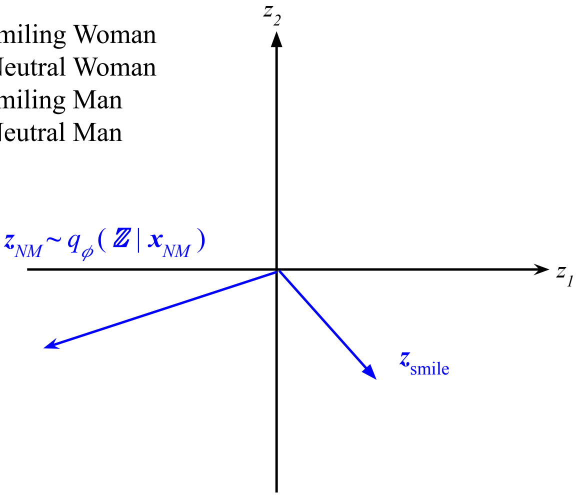


Figure: Encoding of *Neutral Man* image

Concept

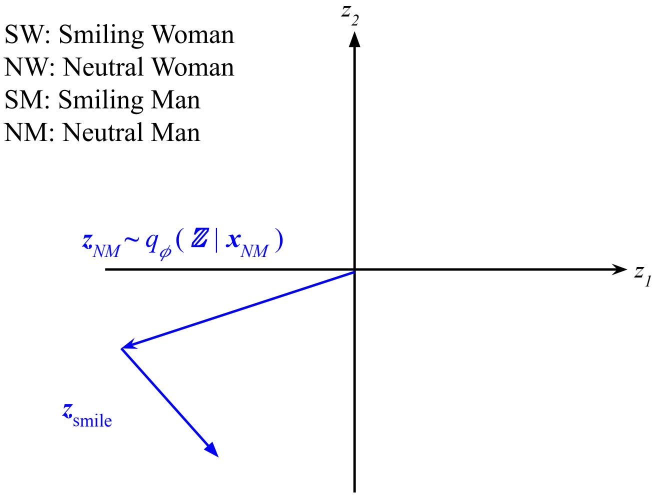


Figure: Adding *Smile* attribute vector to *Neutral Man* latent vector

Concept

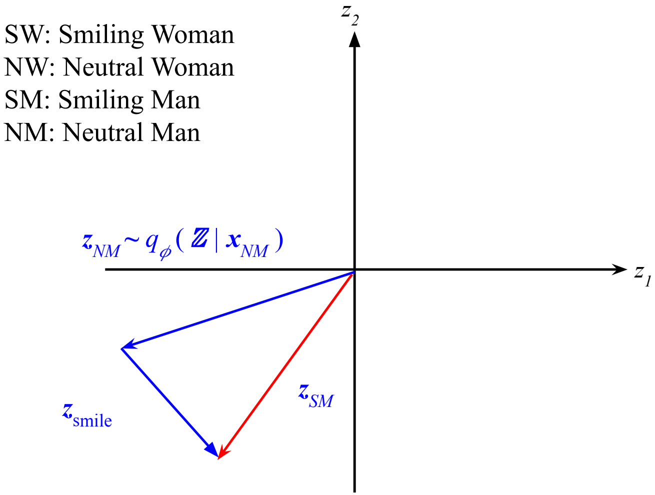


Figure: The latent vector of *Smiling Man*

Sample of Using Attribute Vectors

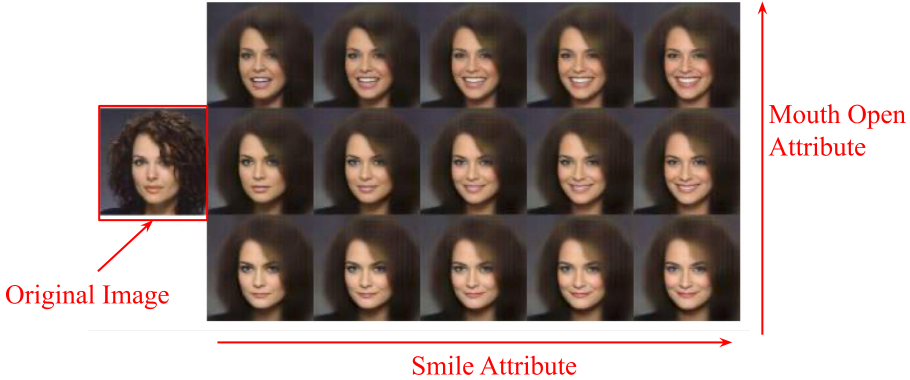


Figure: Application of attribute vectors in real worl application of face images (source: [2])

List of Abbreviations

Complete	Abbreviation
Evidence Lower-BOund	ELBO
Gaussian Mixture Model	GMM
Kullback–Leibler	KL
Left-Hand Side	LHS
Log-Likelihood	LL
Right-Hand Side	RHS
Stochastic Variational Inference	SVI
Variational AutoEncoder	VAE

References I

M. Fink and P. Perona,
"Caltech 10k web faces (1.0) [data set].," https://doi.org/10.22002/D1.20132.

T White,
"Sampling generative networks: notes on a few effective techniques corr (2016),"
arXiv preprint arXiv:1609.04468.