香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

COMP 5212
Machine Learning
Lecture 11

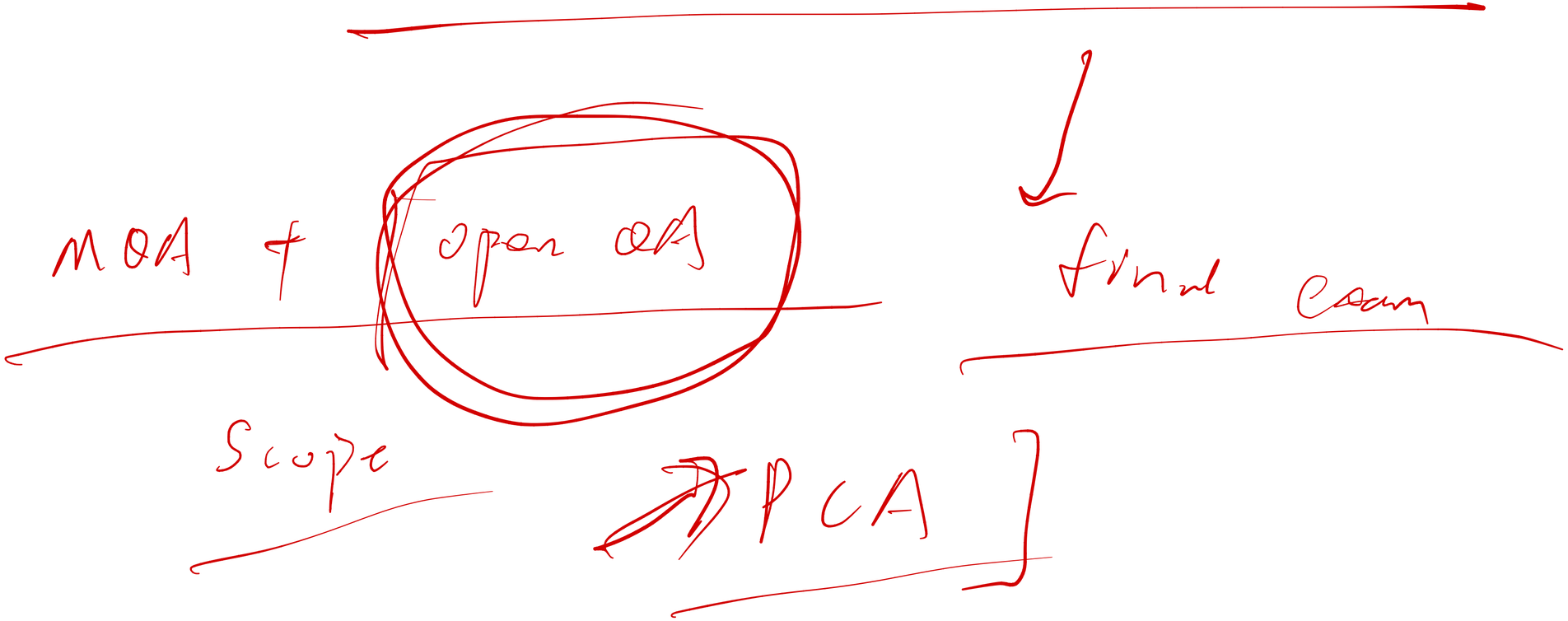
Expectation Maximization

Junxian He
Mar 19, 2026

Midterm Exam

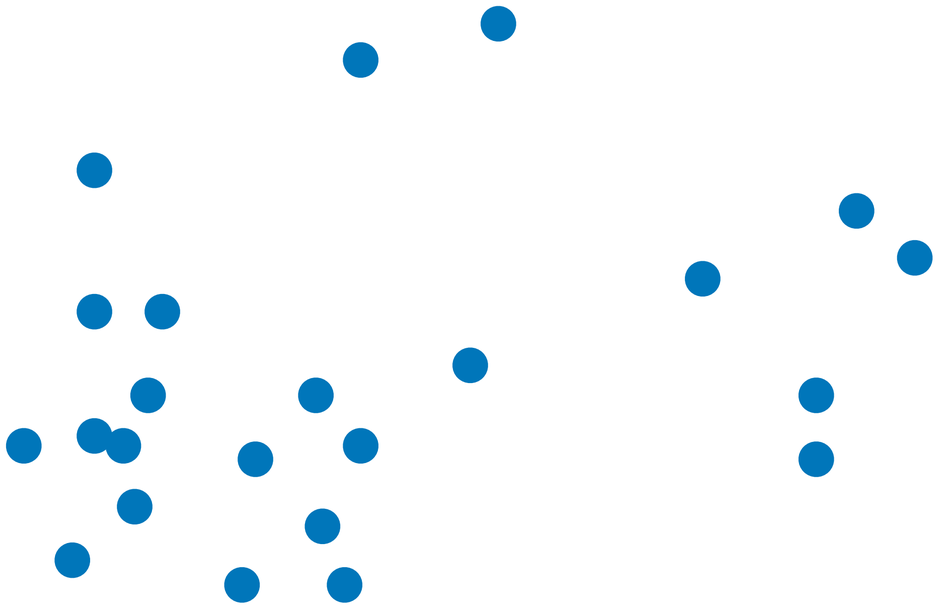
80 min

31/03, in-class, same classroom, 430pm-550pm, one A4-size double-sided cheetsheet is allowed (either printing or handwriting is fine).



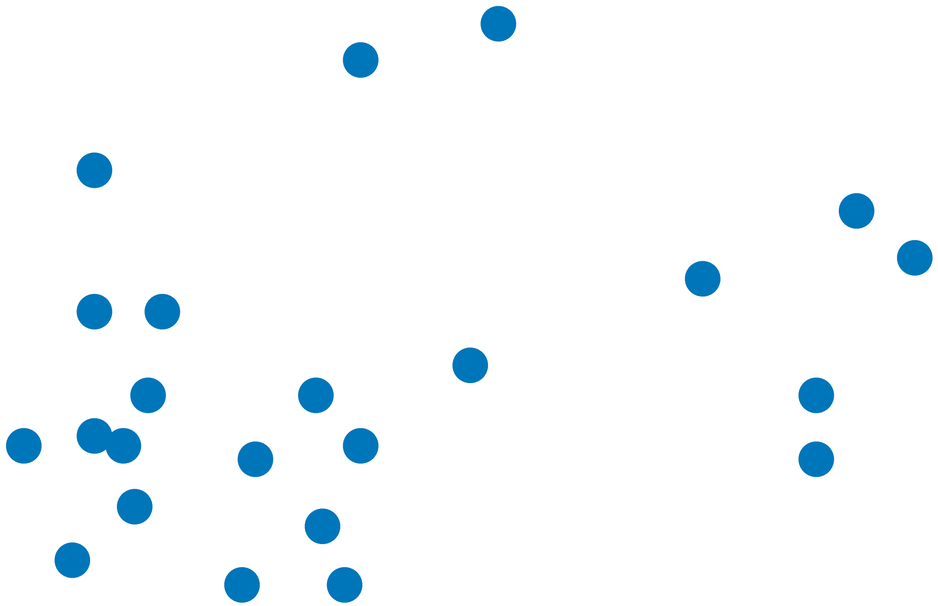
Recap: Generative Models

Recap: Generative Models



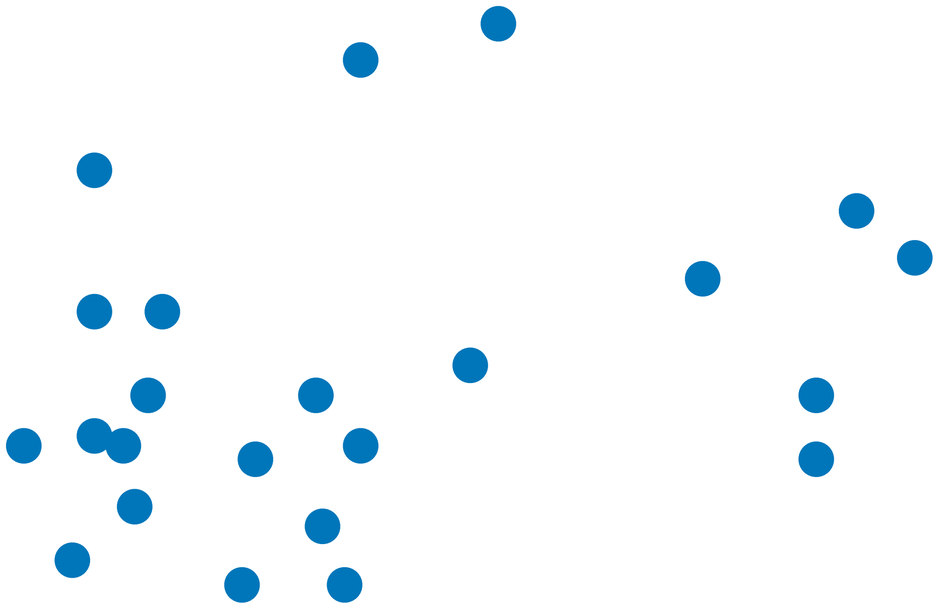
Recap: Generative Models

We want to model $p(x)$



Recap: Generative Models

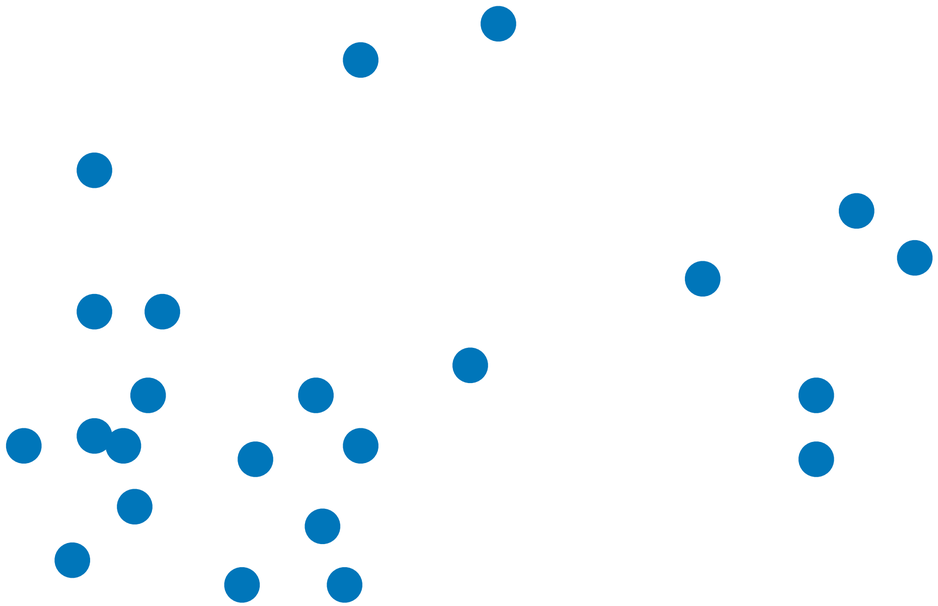
We want to model $p(x)$



In discriminative models, we need to "design" model to make assumption about the function: linear regression, logistic regression, kernel methods ….

Recap: Generative Models

We want to model $p(x)$



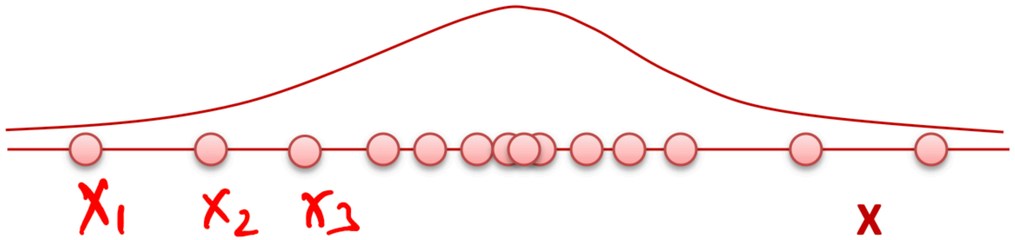
In discriminative models, we need to "design" model to make assumption about the function: linear regression, logistic regression, kernel methods ….

In generative models, we "design" the model and make assumptions about the data, through defining a distribution family

Recap: Generative Models

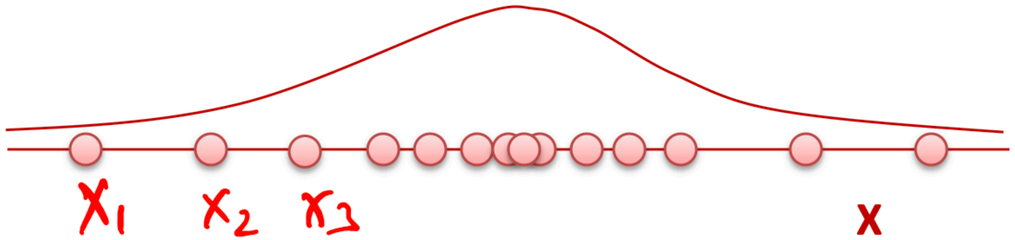
Recap: Generative Models

Data, D =



Recap: Generative Models

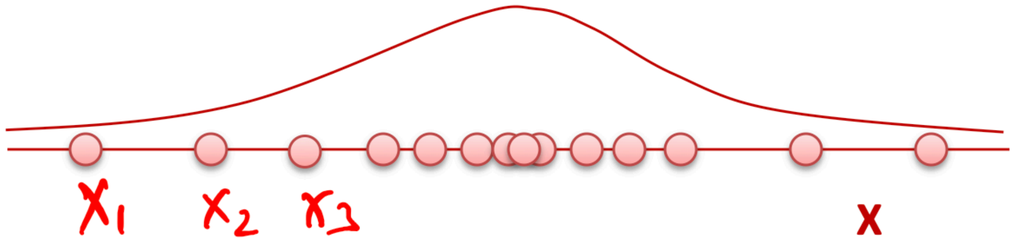
Data, D =



As a simplest case, we directly assume $x \sim N(\mu, \Sigma)$

Recap: Generative Models

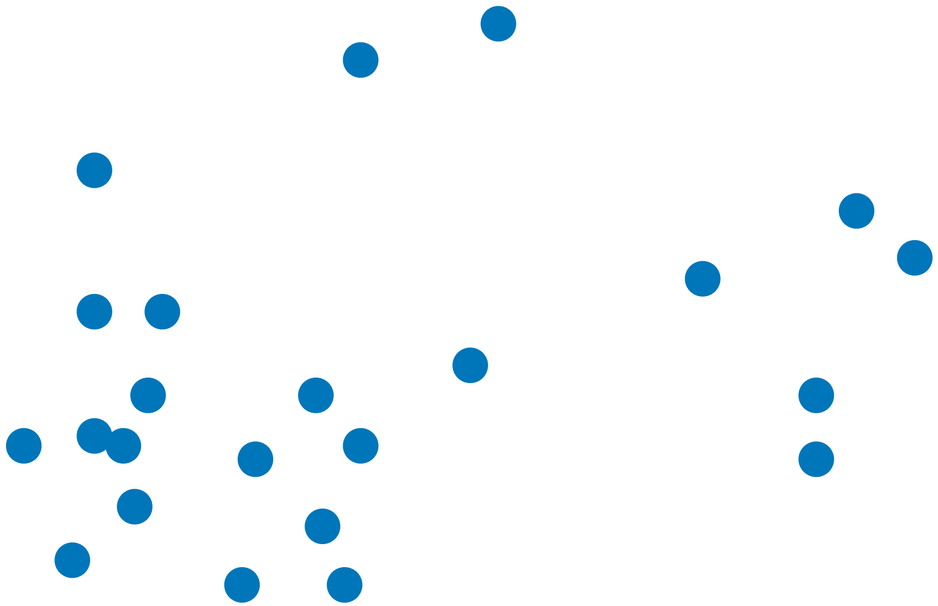
Data, D =



As a simplest case, we directly assume $x \sim N(\mu, \Sigma)$

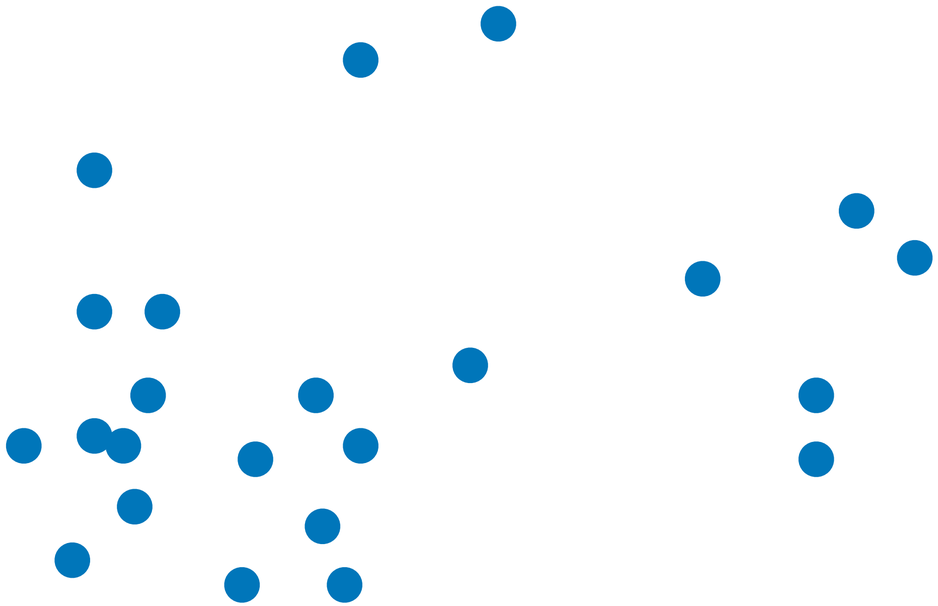
By varying the parameters (μ, Σ), the model represents different distributions that belong to the Gaussian family

Recap: Generative Models



How to construct more complex distribution family?

Recap: Generative Models

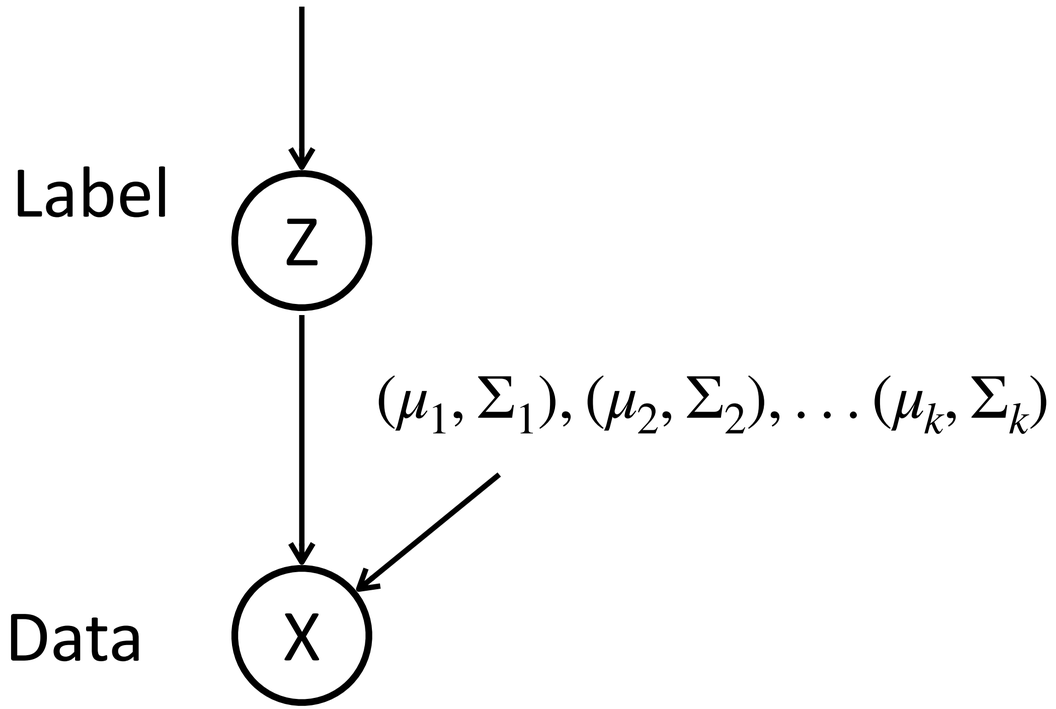


How to construct more complex distribution family?

Introducing more latent variables

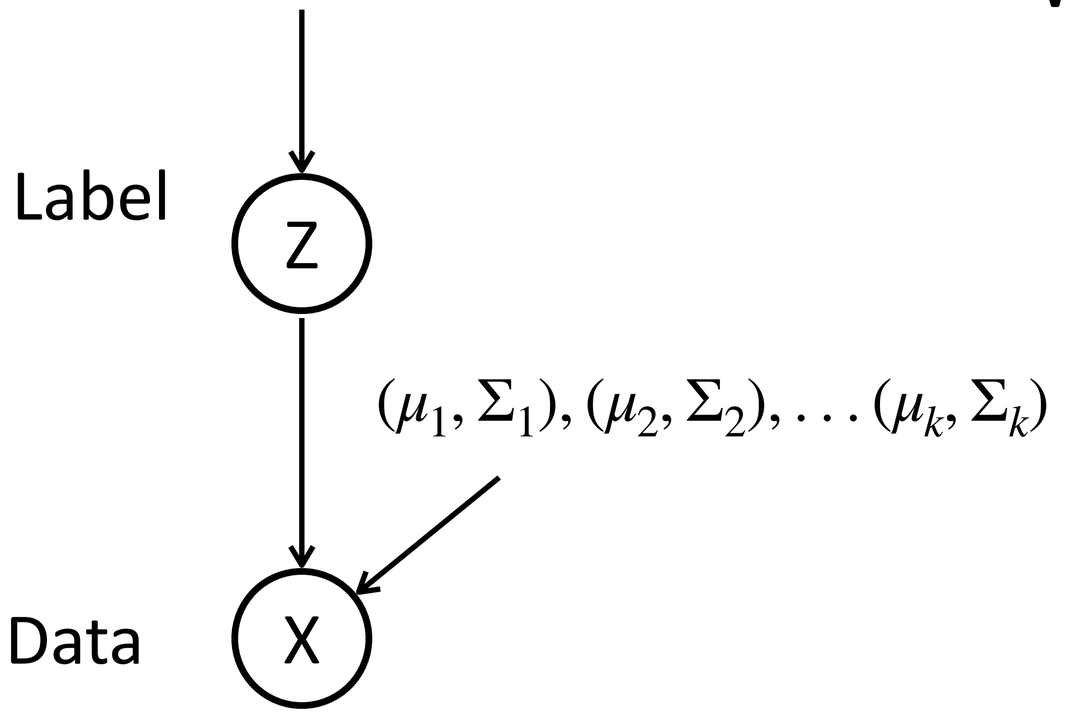
Recap: Gaussian Mixture Model

p(z): multinomial , k classes(e.g. uniform)



Recap: Gaussian Mixture Model

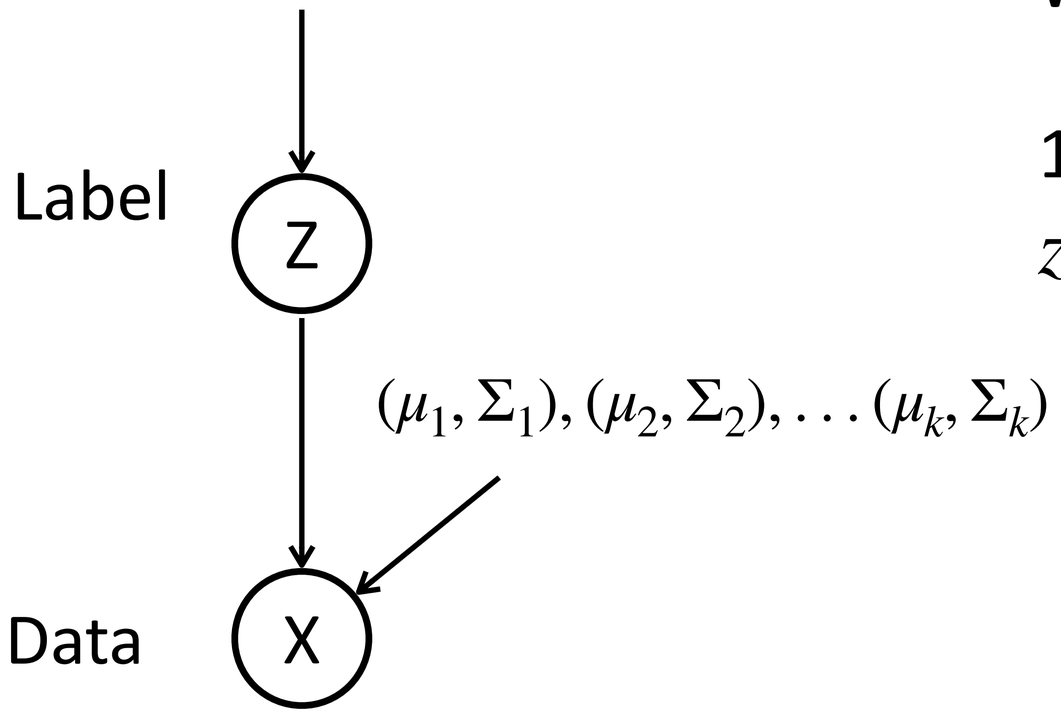
p(z): multinomial , k classes(e.g. uniform)



We assume the generative process as:

Recap: Gaussian Mixture Model

p(z): multinomial , k classes(e.g. uniform)

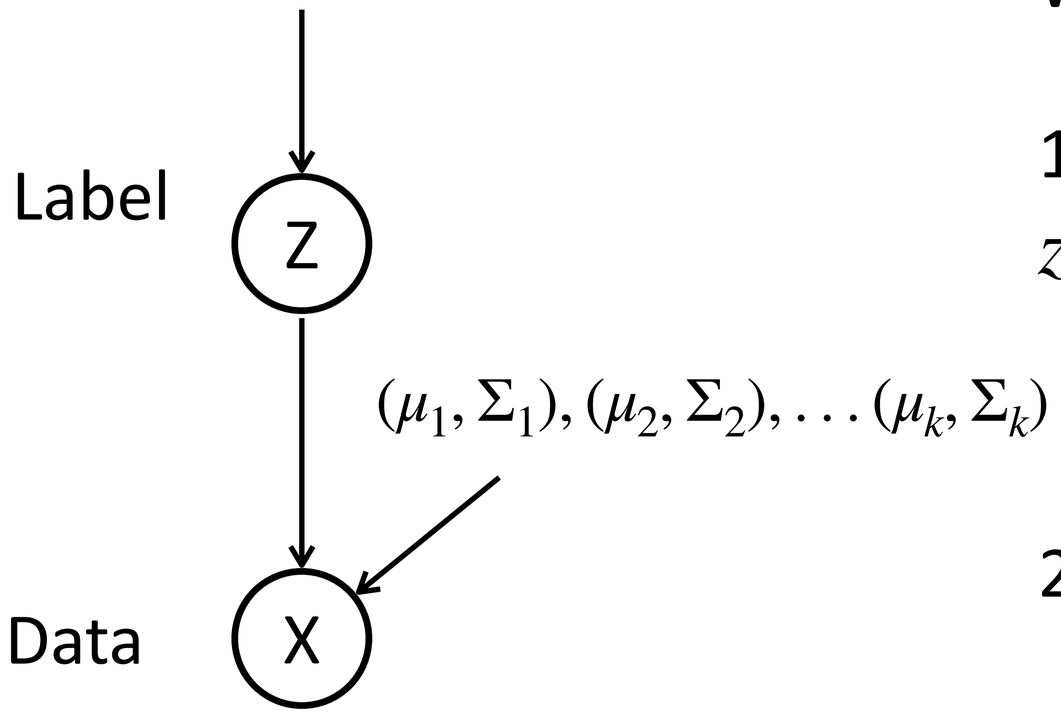


We assume the generative process as:

1. For each data point, sample its label z_i from p(z)

Recap: Gaussian Mixture Model

p(z): multinomial , k classes(e.g. uniform)



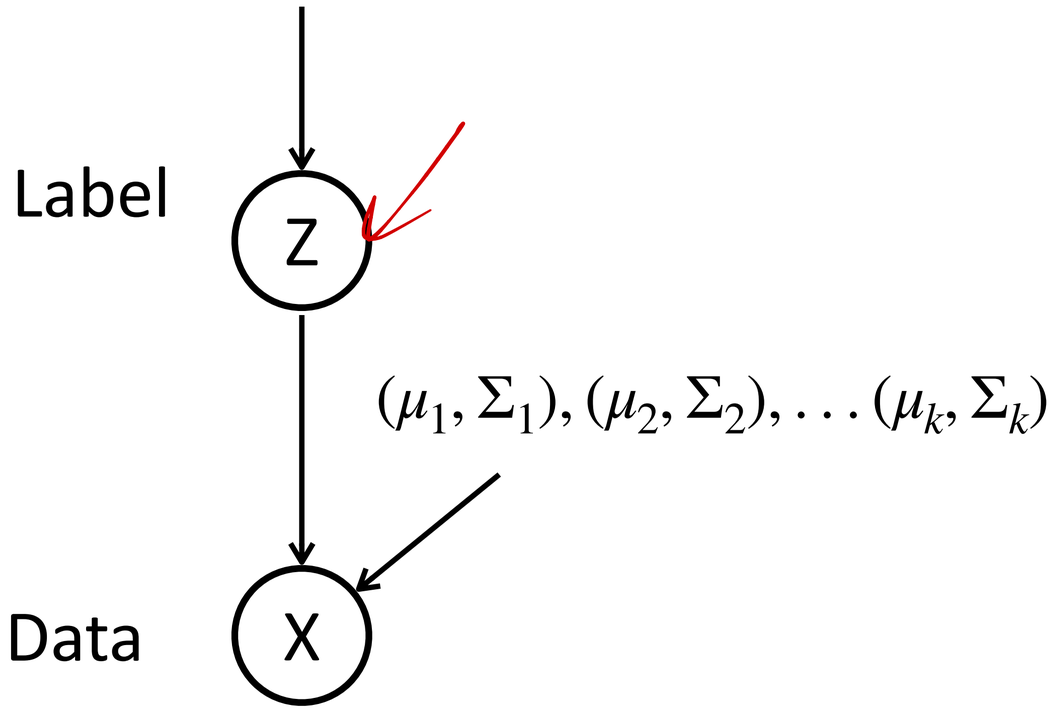
We assume the generative process as:

1. For each data point, sample its label z_i from p(z)

2. Sample $x_i \sim N(\mu_{z_i}, \Sigma_{z_i})$

Recap: MLE for GMM

p(z): multinomial , k classes(e.g. uniform)



Unsupervised:

$$\text{argmax}_{\phi,\mu,\Sigma} \log p(x)$$

How to compute this?

$\log P(x, z)$ supervised learning

$\log P(x)$

Recap: MLE for GMM

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{n} \log p(x^{(i)}; \phi, \mu, \Sigma)$$

$$= \sum_{i=1}^{n} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)}|z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi).$$

Recap: MLE for GMM

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{n} \log p(x^{(i)}; \phi, \mu, \Sigma)$$

$$= \sum_{i=1}^{n} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)}|z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi).$$

if z is continuous, $\int$

1. Intractable (no closed-form for the solution)

Recap: MLE for GMM

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{n} \log p(x^{(i)}; \phi, \mu, \Sigma)$$

$$= \sum_{i=1}^{n} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)} | z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi).$$

1. Intractable (no closed-form for the solution)
2. Large variance in gradient descent

Recap: MLE for GMM

$$\begin{aligned}\ell(\phi, \mu, \Sigma) &= \sum_{i=1}^{n} \log p(x^{(i)}; \phi, \mu, \Sigma) \\ &= \sum_{i=1}^{n} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)}|z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi).\end{aligned}$$

1. Intractable (no closed-form for the solution)
2. Large variance in gradient descent

Expectation Maximization is to address the MLE optimization problem

Things are easy when we know z..

In case we know z

Things are easy when we know z..

In case we know z

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{n} \log p(x^{(i)}|z^{(i)}; \mu, \Sigma) + \log p(z^{(i)}; \phi).$$

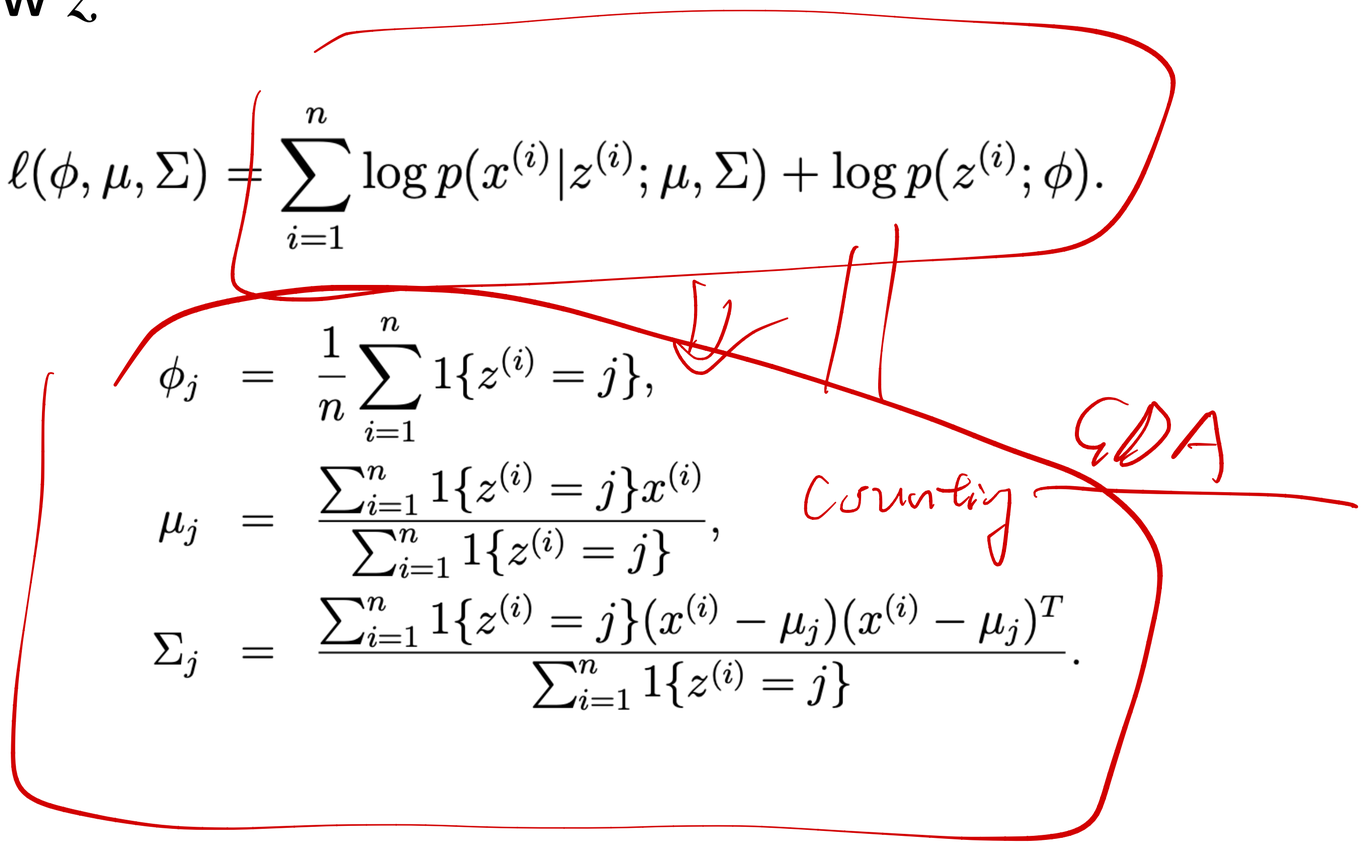
$\log p(x, t)$

Things are easy when we know z..

In case we know z

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{n} \log p(x^{(i)}|z^{(i)}; \mu, \Sigma) + \log p(z^{(i)}; \phi).$$

$$\begin{aligned}
\phi_j &= \frac{1}{n}\sum_{i=1}^{n} 1\{z^{(i)} = j\}, \\
\mu_j &= \frac{\sum_{i=1}^{n} 1\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{n} 1\{z^{(i)} = j\}}, \\
\Sigma_j &= \frac{\sum_{i=1}^{n} 1\{z^{(i)} = j\}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{n} 1\{z^{(i)} = j\}}.
\end{aligned}$$



Things are easy when we know z..

In case we know z

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{n} \log p(x^{(i)}|z^{(i)}; \mu, \Sigma) + \log p(z^{(i)}; \phi).$$

$$\begin{aligned}
\phi_j &= \frac{1}{n}\sum_{i=1}^{n} 1\{z^{(i)} = j\}, \\
\mu_j &= \frac{\sum_{i=1}^{n} 1\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{n} 1\{z^{(i)} = j\}}, \\
\Sigma_j &= \frac{\sum_{i=1}^{n} 1\{z^{(i)} = j\} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{n} 1\{z^{(i)} = j\}}.
\end{aligned}$$

Expectation maximization is to infer the latent variables first (z here), and maximize the likelihood given the inferred z

Expectation Maximization for GMM

Repeat until convergence:

{

}

Expectation Maximization for GMM

Repeat until convergence:

{

(E-step) For each i, j, set

$$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

}

Expectation Maximization for GMM

Repeat until convergence:

{

(E-step) For each i, j, set

$$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

Compute the posterior distribution, given current parameters

}

Expectation Maximization for GMM

Repeat until convergence:

{

No parameter change in E-step

(E-step) For each i, j, set

$$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

Compute the posterior distribution, given current parameters

}

Expectation Maximization for GMM

Repeat until convergence:

{

No parameter change in E-step

(E-step) For each i, j, set

$$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

Compute the posterior distribution, given current parameters

(M-step) Update the parameters:

$$\phi_j := \frac{1}{n} \sum_{i=1}^{n} w_j^{(i)},$$

$$\mu_j := \frac{\sum_{i=1}^{n} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{n} w_j^{(i)}},$$

$$\Sigma_j := \frac{\sum_{i=1}^{n} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{n} w_j^{(i)}}$$

}

Expectation Maximization

- Why does it work?
- What is its relation to MLE estimation?
- How is convergence guaranteed?
- When we perform EM, what is the real objective that we are optimizing?

General EM Algorithm

General EM Algorithm

$$p(x;\theta) = \sum_{z} p(x,z;\theta)$$

General EM Algorithm

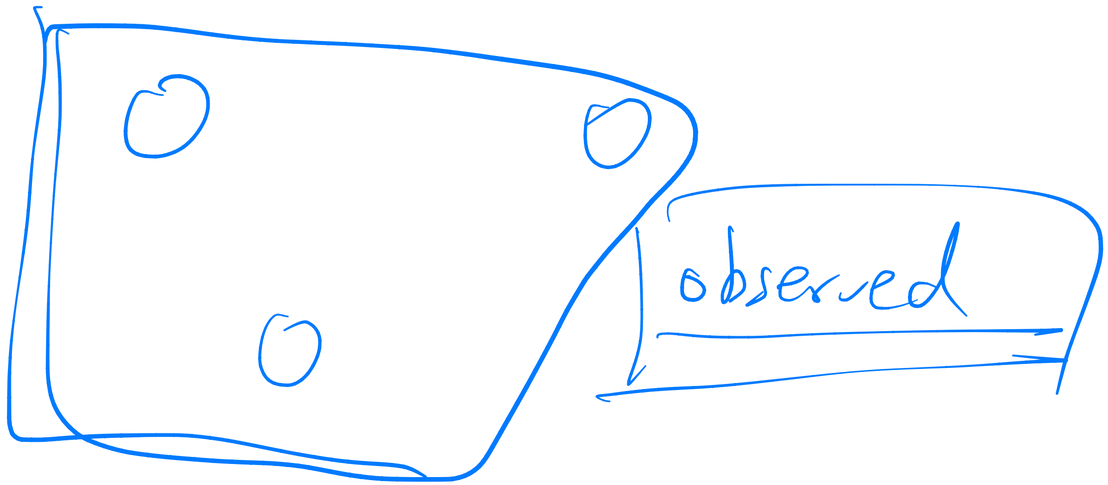
$$p(x;\theta) = \sum_z p(x,z;\theta)$$

z

$\vec{z} = \{z_1, z_2 \cdots z_3\}$

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x^{(i)};\theta)$$

$$= \sum_{i=1}^{n} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta).$$



General EM Algorithm

$$p(x;\theta) = \sum_z p(x,z;\theta)$$

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{n} \log p(x^{(i)};\theta) \\ &= \sum_{i=1}^{n} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta).\end{aligned}$$

Let Q to be a distribution over z

General EM Algorithm

$$p(x;\theta) = \sum_z p(x,z;\theta)$$

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{n} \log p(x^{(i)};\theta) \\ &= \sum_{i=1}^{n} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta).\end{aligned}$$

Let Q to be a distribution over z

$$\begin{aligned}\log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\ &= \log \sum_z Q(z) \frac{p(x,z;\theta)}{Q(z)} \\ &\geq \sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)}\end{aligned}$$

General EM Algorithm

$$p(x;\theta) = \sum_z p(x,z;\theta)$$

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{n} \log p(x^{(i)};\theta) \\ &= \sum_{i=1}^{n} \log \sum_{z^{(i)}} p(x^{(i)},z^{(i)};\theta).\end{aligned}$$

Let Q to be a distribution over z

This lower bound holds for any Q(z)

$$\begin{aligned}\log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\ &= \log \sum_z Q(z)\frac{p(x,z;\theta)}{Q(z)} \\ &\geq \sum_z Q(z)\log\frac{p(x,z;\theta)}{Q(z)}\end{aligned}$$

General EM Algorithm

$$p(x;\theta) = \sum_z p(x,z;\theta)$$

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{n} \log p(x^{(i)};\theta) \\ &= \sum_{i=1}^{n} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta).\end{aligned}$$

Let Q to be a distribution over z

This lower bound holds for any Q(z)

$$\begin{aligned}\log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\ &= \log \sum_z Q(z) \frac{p(x,z;\theta)}{Q(z)} \\ &\geq \sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)}\end{aligned}$$

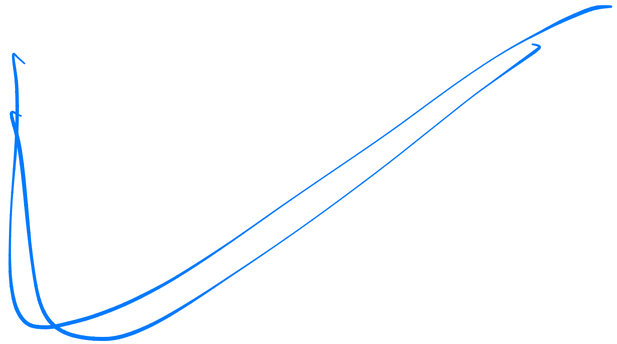
Jensen inequality

log concave

Jensen Inequality

For a convex function f, and $t \in [0,1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$



Jensen Inequality

For a convex function f, and $t \in [0,1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

In probability:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Jensen Inequality

For a convex function f, and $t \in [0,1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

In probability:

X is constant $f(E[x_1]) = E[f(x_1)]$

$$f(\mathbb{E}[X]) \leq [f(X)]$$

If f is strictly convex, then equality holds only when X is a constant

Evidence Lower Bound (ELBO)

Evidence Lower Bound (ELBO)

$$
\begin{aligned}
\log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\
&= \log \sum_z Q(z) \frac{p(x,z;\theta)}{Q(z)} \\
&\geq \sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)}
\end{aligned}
$$

ELBO

Evidence Lower Bound (ELBO)

$$
\begin{aligned}
\log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\
&= \log \sum_z Q(z) \frac{p(x,z;\theta)}{Q(z)} \\
&\geq \sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)}
\end{aligned}
$$

ELBO

Evidence Lower Bound (ELBO)

$$
\begin{aligned}
\log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\
&= \log \sum_z Q(z) \frac{p(x,z;\theta)}{Q(z)} \\
&\geq \sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)}
\end{aligned}
$$

ELBO

Q(z) uniform

Q(z) N(0,1)

Because the log likelihood is intractable, people often optimize its lower bound instead

Evidence Lower Bound (ELBO)

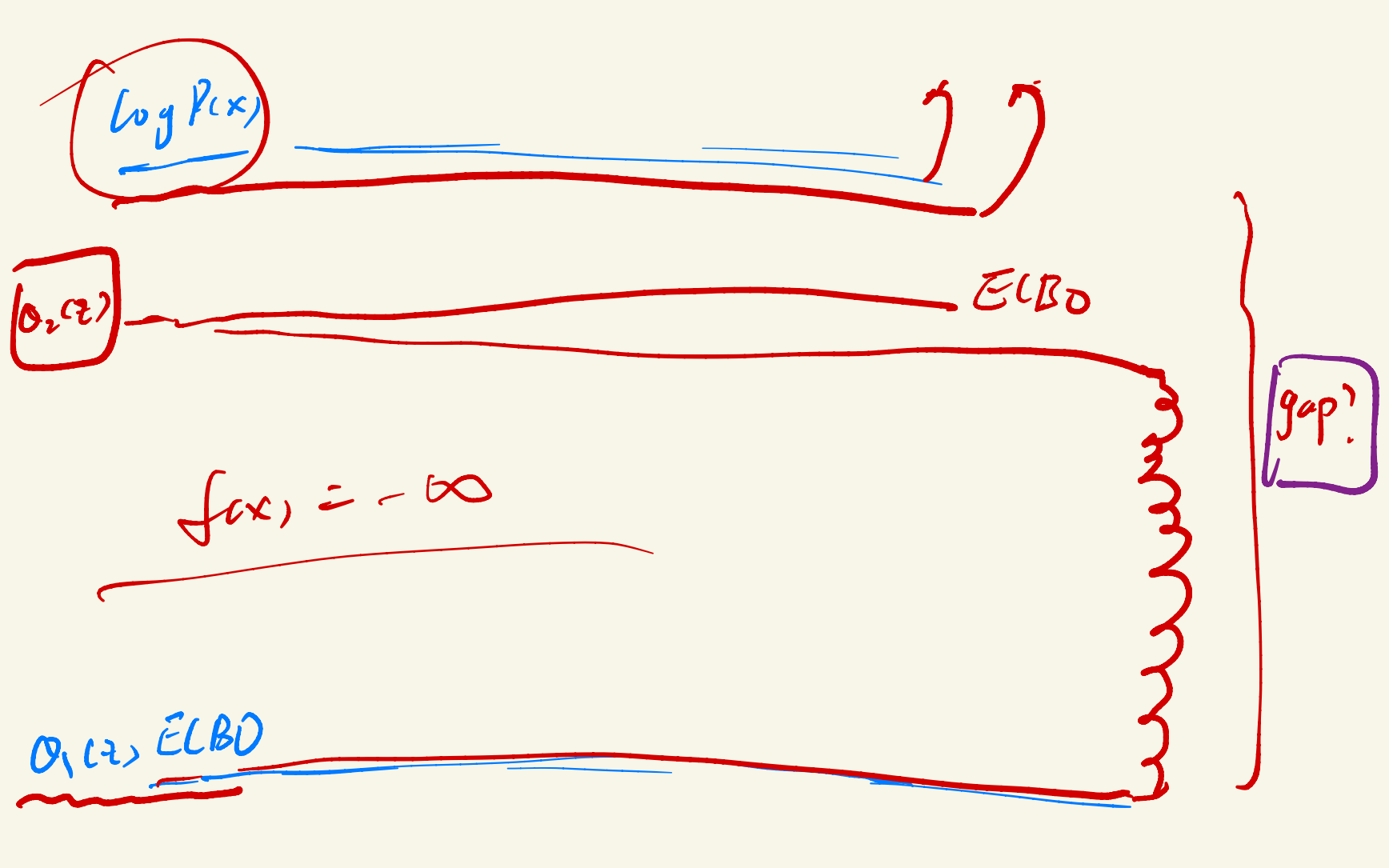
$$\begin{aligned}
\log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\
&= \log \sum_z Q(z) \frac{p(x,z;\theta)}{Q(z)} \\
&\geq \sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)}
\end{aligned}$$

ELBO

tight lower bound

Because the log likelihood is intractable, people often optimize its lower bound instead

Why optimizing lower bound works? How to choose Q(z), why we computed posterior in the E step, what is the benefit?


log P(x)
θ2(z)
ELBO
gap
f(x) = -∞
θ1(z) ELBO

Evidence Lower Bound (ELBO)

$$
\begin{aligned}
\log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\
&= \log \sum_z Q(z) \frac{p(x,z;\theta)}{Q(z)} \\
&\geq \sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)}
\end{aligned}
$$

Evidence Lower Bound (ELBO)

$$\begin{aligned}
\log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\
&= \log \sum_z Q(z) \frac{p(x,z;\theta)}{Q(z)} \\
&\geq \sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)}
\end{aligned}$$

Jensen Inequality

When is the lower bound tight?

$\log P(x) = ELBO$

Evidence Lower Bound (ELBO)

$$\begin{aligned}
\log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\
&= \log \sum_z Q(z) \frac{p(x,z;\theta)}{Q(z)} \\
&\geq \sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)}
\end{aligned}$$

$E[\ ]$

When is the lower bound tight?

$$\frac{p(x,z;\theta)}{Q(z)} = c$$

Evidence Lower Bound (ELBO)

$$\begin{aligned}
\log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\
&= \log \sum_z Q(z) \frac{p(x,z;\theta)}{Q(z)} \\
&\geq \sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)}
\end{aligned}$$

$\sum_z Q(z) = 1$

$Q(z)$

When is the lower bound tight?

$$\frac{p(x,z;\theta)}{Q(z)} = c$$

$$\begin{aligned}
Q(z) &= \frac{p(x,z;\theta)}{\sum_z p(x,z;\theta)} \\
&= \frac{p(x,z;\theta)}{p(x;\theta)} \\
&= p(z|x;\theta)
\end{aligned}$$

$Q(z) = \frac{p(x,z)}{c}$

Evidence Lower Bound (ELBO)

Evidence Lower Bound (ELBO)

Verify $\sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$ when Q(z) = p(z|x) ?

$= \log P(x)$

$$\sum_z P(z|x) \log \frac{P(x, z; \theta)}{P(z|x)}$$

$$P(x, z) = P(z|x) P(x)$$

$$= \sum_z P(z|x) \log P(x)$$

$$= \log P(x) \cdot \sum_z P(z|x) = 1$$

$$= \log P(x) = \text{ELBO}$$

Evidence Lower Bound (ELBO)

Verify $\sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$ when Q(z) = p(z|x) ?

$$\text{ELBO}(x; Q, \theta) = \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$$

Evidence Lower Bound (ELBO)

Verify $\sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$ when Q(z) = p(z|x) ?

$$\text{ELBO}(x; Q, \theta) = \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$$

$$\forall Q, \theta, x, \quad \log p(x; \theta) \geq \text{ELBO}(x; Q, \theta)$$

Evidence Lower Bound (ELBO)

Verify $\sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$ when Q(z) = p(z|x) ?

$$\text{ELBO}(x; Q, \theta) = \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$$

$$\forall Q, \theta, x, \quad \log p(x; \theta) \geq \text{ELBO}(x; Q, \theta)$$

For a dataset of many data samples

$$\ell(\theta) \geq \sum_i \text{ELBO}(x^{(i)}; Q_i, \theta)$$

$$= \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

Evidence Lower Bound (ELBO)

$$\mathrm{ELBO}(x; Q, \theta) = \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$$

Evidence Lower Bound (ELBO)

$$\text{ELBO}(x; Q, \theta) = \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$$

What is $\text{argmax}_{Q(z)} \text{ELBO}(x; Q, \theta)$?

$Q^*(z) = P(z|x)$

$P(z|x)$

$\text{ELBO} \leq \log P(x)$

The General EM Algorithm

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_{i=1}^{n} \text{ELBO}(x^{(i)}; Q_i, \theta)$$

$$= \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}.$$

}

The General EM Algorithm

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

Based on current θ, model parameters does not change in E-step

(M-step) Set

$$\theta := \arg\max_\theta \sum_{i=1}^{n} \text{ELBO}(x^{(i)}; Q_i, \theta)$$

$$= \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}.$$

}

The General EM Algorithm

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

Based on current θ, model parameters does not change in E-step

(M-step) Set

$$\theta := \arg\max_\theta \sum_{i=1}^{n} \text{ELBO}(x^{(i)}; Q_i, \theta)$$

$$= \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}.$$

$Q(z)$ is not relevant to θ, and $Q(z)$does not change in the M-step

}

The General EM Algorithm

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

Based on current θ, model parameters does not change in E-step

(M-step) Set

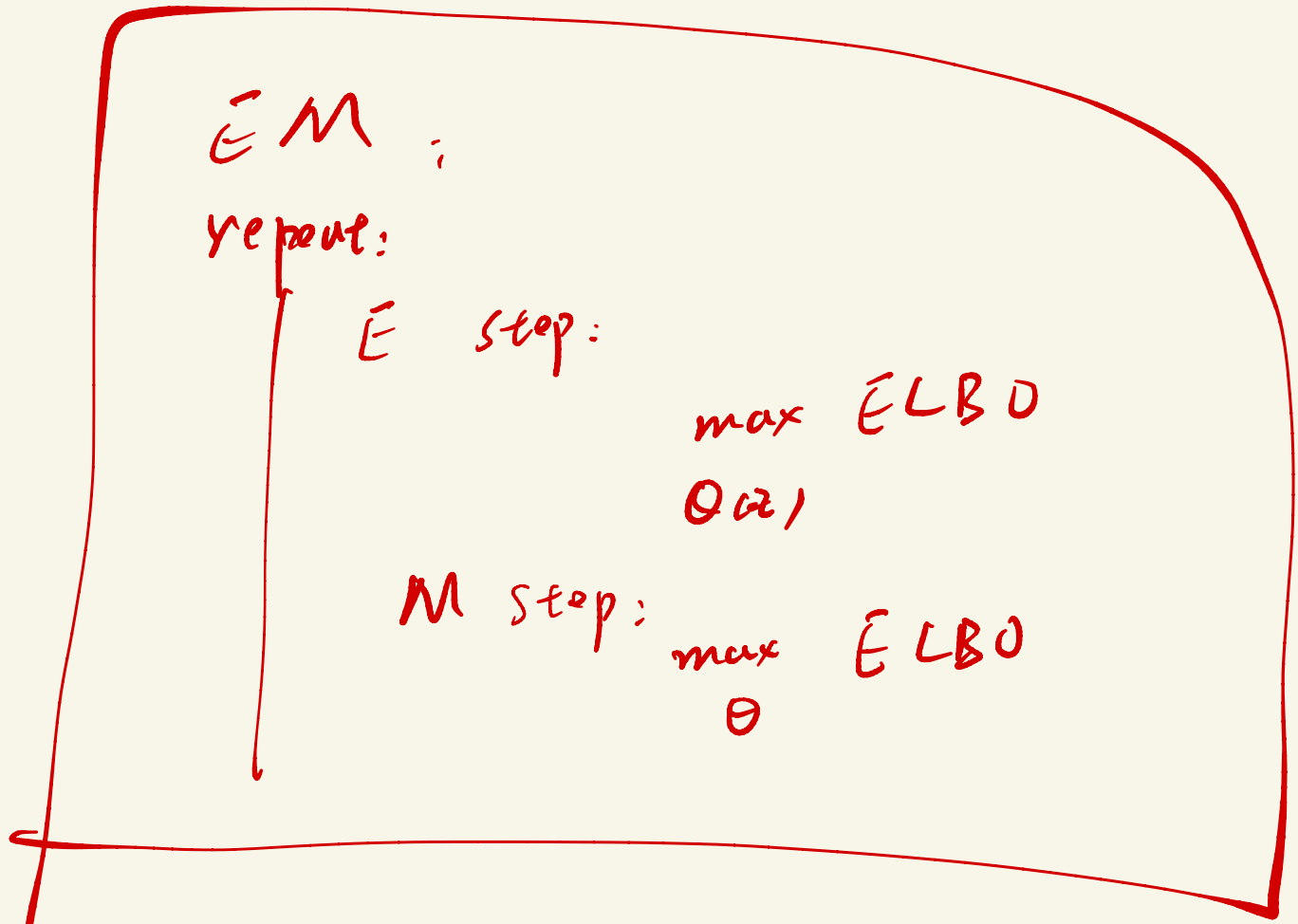
$$\theta := \arg\max_\theta \sum_{i=1}^{n} \text{ELBO}(x^{(i)}; Q_i, \theta)$$

$$= \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}.$$

$Q(z)$ is not relevant to θ, and $Q(z)$does not change in the M-step

}

E-step is maximizing ELBO over Q(z), M-step is maximizing ELBO overθ


EM:
repeat:
E step:
max ELBO
Q α
M step:
max ELBO
θ

The General EM Algorithm

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

Based on current θ, model parameters does not change in E-step

(M-step) Set

$$\theta := \arg\max_\theta \sum_{i=1}^{n} \text{ELBO}(x^{(i)}; Q_i, \theta)$$

$$= \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}.$$

$Q(z)$ is not relevant to θ, and $Q(z)$does not change in the M-step

}

E-step is maximizing ELBO over Q(z), M-step is maximizing ELBO overθ

Why is maximizing lower-bound sufficient?

EM is Hill Climbing

$\log p(x;\theta)$

ELBO

Larger

true objective

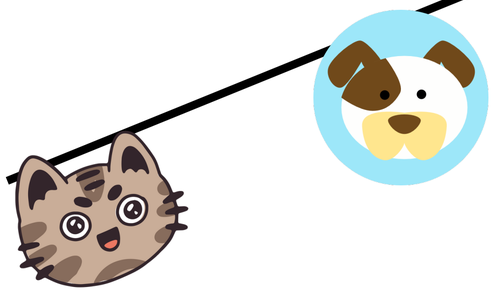
EM is Hill Climbing

$\log p(x;\theta)$

Only related to θ, no z

ELBO

Larger



EM is Hill Climbing

$\log p(x;\theta)$

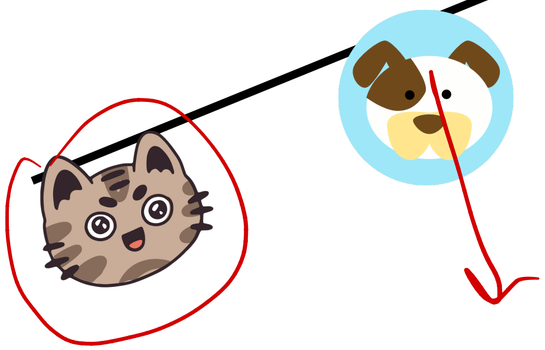
ELBO

EM is Hill Climbing

$\log p(x;\theta)$

ELBO

Larger



E-step: $Q(z) = p(z \mid x;\theta)$, making ELBO tight

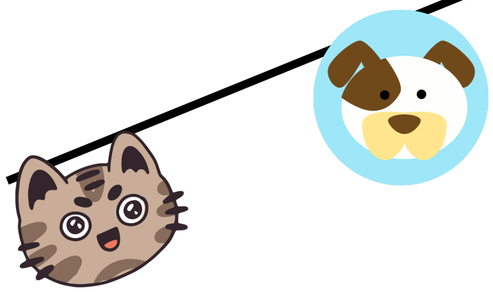
not change

EM is Hill Climbing

$\log p(x;\theta)$

ELBO

Larger



E-step: $Q(z) = p(z \mid x;\theta)$, making ELBO tight

"dog" doesn't change, because θ does not change

EM is Hill Climbing

$\log p(x;\theta)$

ELBO

Larger

E-step: $Q(z) = p(z \mid x;\theta)$, making ELBO tight

"dog" doesn't change, because θ does not change

EM is Hill Climbing

$\log p(x;\theta)$

ELBO

Larger

M-step: $\max_{\theta} ELBO$

$Q(z)$ not change

$P(z|x;\theta)$

ELBO becomes larger, and it is not tight anymore because posterior changes

EM is Hill Climbing

$\log p(x;\theta)$

ELBO

ELBO becomes larger, and it is not tight anymore because posterior changes

EM is Hill Climbing

$\log p(x;\theta)$

ELBO

ELBO becomes larger, and it is not tight anymore because posterior changes

EM is Hill Climbing

$\log p(x;\theta)$

ELBO

EM is Hill Climbing

$\log p(x;\theta)$

ELBO

E-step: $Q(z) = p(z \mid x;\theta)$, making ELBO tight

"dog" doesn't change, because θ does not change

EM is Hill Climbing

$\log p(x;\theta)$

ELBO

M-step: $\max_{\theta} ELBO$

ELBO becomes larger, and it is not tight anymore because posterior changes

EM is Hill Climbing

$\log p(x;\theta)$

ELBO

M-step: $\max_{\theta} ELBO$

ELBO becomes larger, and it is not tight anymore because posterior changes

$\log p(x;\theta)$ is monotonically increasing!

EM is Hill Climbing

$\log p(x;\theta)$

ELBO

M-step: $\max_{\theta} ELBO$

ELBO becomes larger, and it is not tight anymore because posterior changes

$\log p(x;\theta)$ is monotonically increasing!

We are doing MLE implicitly!

EM is Hill Climbing

$\log p(x;\theta)$

ELBO

$\log p(x;\theta) \geq ?$

Larger

M-step: $\max_{\theta} ELBO$

ELBO becomes larger, and it is not tight anymore because posterior changes

$\log p(x;\theta)$ is monotonically increasing!

We are doing MLE implicitly!

Convergence is guaranteed

Revisit the E-Step

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_{i=1}^{n} \text{ELBO}(x^{(i)}; Q_i, \theta)$$

$$= \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}.$$

}

Revisit the E-Step

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_{i=1}^{n} \text{ELBO}(x^{(i)}; Q_i, \theta)$$

$$= \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}.$$

}

Computable posterior is important. If Q(z) is not the posterior, then there is no guarantee that $\log p(x)$ is improved at every iteration

Revisit the E-Step

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_{i=1}^{n} \text{ELBO}(x^{(i)}; Q_i, \theta)$$

$$= \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}.$$

}

Computable posterior is important. If Q(z) is not the posterior, then there is no guarantee that $\log p(x)$ is improved at every iteration

Still remember conjugate prior? Which is for easy-to-compute posterior

Revisit the M-Step

Revisit the M-Step

$$\text{argmax}_\theta \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)} = \text{argmax}_\theta \sum_z Q(z) \log p(x, z; \theta)$$

$$\text{argmax}_\theta \sum_z Q(z) \log p(x, z; \theta) - \sum_z Q(z) \log Q(z)$$

constant

Revisit the M-Step

$$\text{argmax}_\theta \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)} = \text{argmax}_\theta \sum_z Q(z) \log p(x, z; \theta)$$

Sometimes the sum is computable, but sometimes not

Revisit the M-Step

$$\text{argmax}_\theta \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)} = \text{argmax}_\theta \sum_z Q(z) \log p(x, z; \theta)$$

Sometimes the sum is computable, but sometimes not

$$\text{argmax}_\theta \sum_z Q(z) \log p(x, z; \theta) = \text{argmax}_\theta \mathbb{E}_{z \sim Q(z)} \log p(x, z; \theta)$$

Revisit the M-Step

$$\text{argmax}_\theta \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)} = \text{argmax}_\theta \sum_z Q(z) \log p(x, z; \theta)$$

Sometimes the sum is computable, but sometimes not

$$\text{argmax}_\theta \sum_z Q(z) \log p(x, z; \theta) = \text{argmax}_\theta \mathbb{E}_{z \sim Q(z)} \log p(x, z; \theta)$$

$\frac{\sum_{i=1}^{m} \log p(x, z^{(i)}; \theta)}{m}$ $z^{(i)} \sim Q(z)$

We can use Monto-Carlo sampling to approximate the expectation

Comparing Direct Maximization and EM

Direct maximization:

$$\text{argmax}_\theta \log \sum_z p(x \mid z; \theta) p(z) = \text{argmax}_\theta \log \mathbb{E}_{z \sim p(z)} p(x \mid z; \theta)$$

Comparing Direct Maximization and EM

Direct maximization:

$$\text{argmax}_\theta \log \sum_z p(x \mid z; \theta) p(z) = \text{argmax}_\theta \log \mathbb{E}_{z \sim p(z)} p(x \mid z; \theta)$$

M-Step in EM:

$$\text{argmax}_\theta \sum_z Q(z) \log p(x, z; \theta) = \text{argmax}_\theta \mathbb{E}_{z \sim Q(z)} \log p(x, z; \theta)$$

Comparing Direct Maximization and EM

Direct maximization:

$$\text{argmax}_\theta \log \sum_z p(x \mid z; \theta) p(z) = \text{argmax}_\theta \log \mathbb{E}_{z \sim p(z)} p(x \mid z; \theta)$$

M-Step in EM:

$$\text{argmax}_\theta \sum_z Q(z) \log p(x, z; \theta) = \text{argmax}_\theta \mathbb{E}_{z \sim Q(z)} \log p(x, z; \theta)$$

Why don't we use MC sampling to approximate expectation in direct maximization?

Comparing Direct Maximization and EM

Direct maximization:

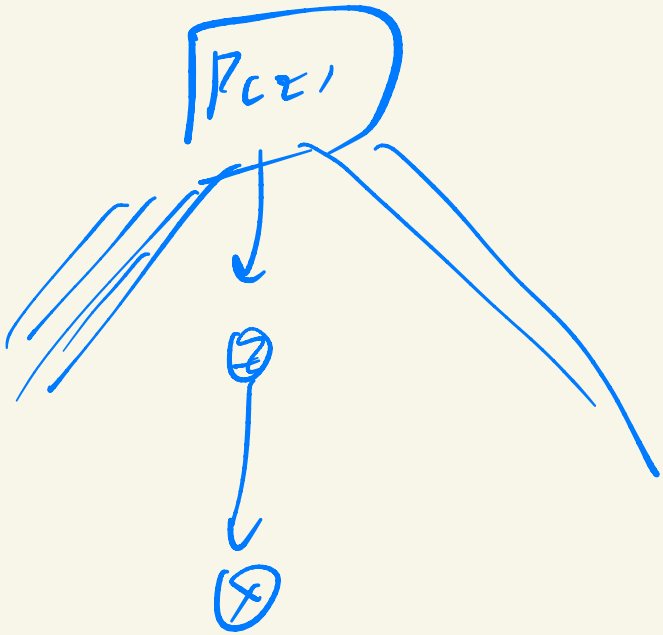
$$\text{argmax}_\theta \log \sum_z p(x \mid z; \theta) p(z) = \text{argmax}_\theta \log \mathbb{E}_{z \sim p(z)} p(x \mid z; \theta)$$

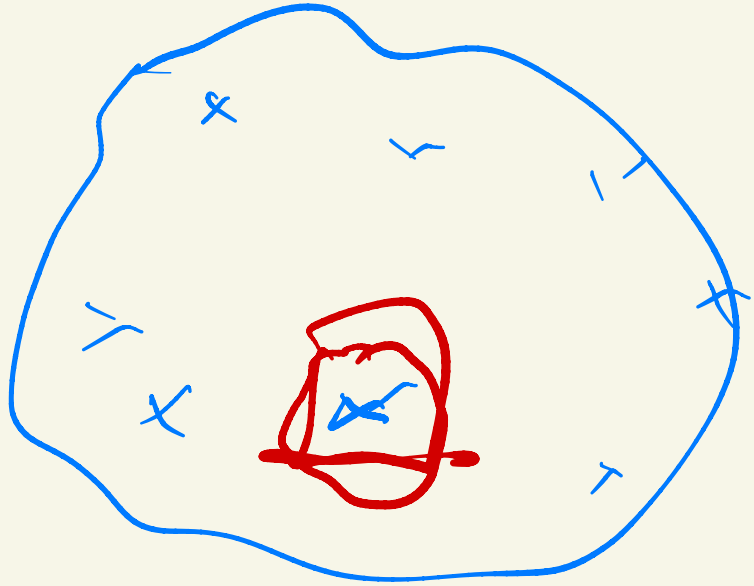
M-Step in EM:

$$\text{argmax}_\theta \sum_z Q(z) \log p(x, z; \theta) = \text{argmax}_\theta \mathbb{E}_{z \sim Q(z)} \log p(x, z; \theta)$$

Why don't we use MC sampling to approximate expectation in direct maximization?

It may need a large number of samples to have a good approximation





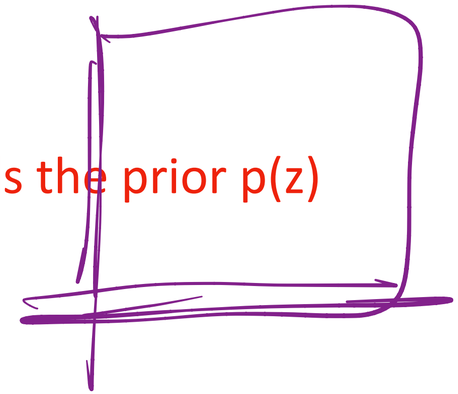
$E_{z \sim P(z)}[x]$ —

$q(z) = P(z|x)$

Other Interpretations of ELBO

$$\begin{aligned}\mathrm{ELBO}(x; Q, \theta) &= \mathrm{E}_{z\sim Q}[\log p(x, z; \theta)] - \mathrm{E}_{z\sim Q}[\log Q(z)] \\ &= \mathrm{E}_{z\sim Q}[\log p(x|z; \theta)] - D_{KL}(Q\|p_z)\end{aligned}$$

Regularize Q(z) towards the prior p(z)



$$\mathrm{ELBO} = E_{z\sim Q}\left[\frac{\log p(x, z)}{Q(z)}\right]$$

$$\mathrm{ELBO}(x; Q, \theta) = \log p(x) - D_{KL}(Q\|p_{z|x})$$

Maximizing ELBO over Q(z) is essentially solving the posterior distribution p(z|x)

$$\text{gap} = KL(Q(z) \,||\, P(z|x))$$

$$\max \text{ELBO} = E_{Q(z)}\left[\log \frac{P(x,z)}{Q(z)}\right]$$

$$= E_{Q(z)}[\log P(x,z)] - E_{Q(z)}[\log Q(z)]$$

$$= E_{Q(z)}[\log P(x|z) + \log P(z)] - E_{Q(z)}[\log Q(z)]$$

$$= E_{Q(z)}[\log P(x|z)] - E_{Q(z)}\log \frac{Q(z)}{P(z)}$$

$E_{Q(z)}[\log P(x|z)]$: reconstruction

$E_{Q(z)}\log \frac{Q(z)}{P(z)}$: $KL(Q(z) \| P(z))$

$P(z|x)$ $P(x|z)$

Further Questions

Further Questions

- What if we do not have closed-form model posterior?

Further Questions

- What if we do not have closed-form model posterior? —> Variational EM

Further Questions

- What if we do not have closed-form model posterior? —> Variational EM

The process of approximating the model posterior is called variational inference

Further Questions

- What if we do not have closed-form model posterior? —> Variational EM

The process of approximating the model posterior is called variational inference

We will learn variational autoencoder later

Thank You!
Q & A