



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

COMP 5212
Machine Learning
Lecture 11

Expectation Maximization

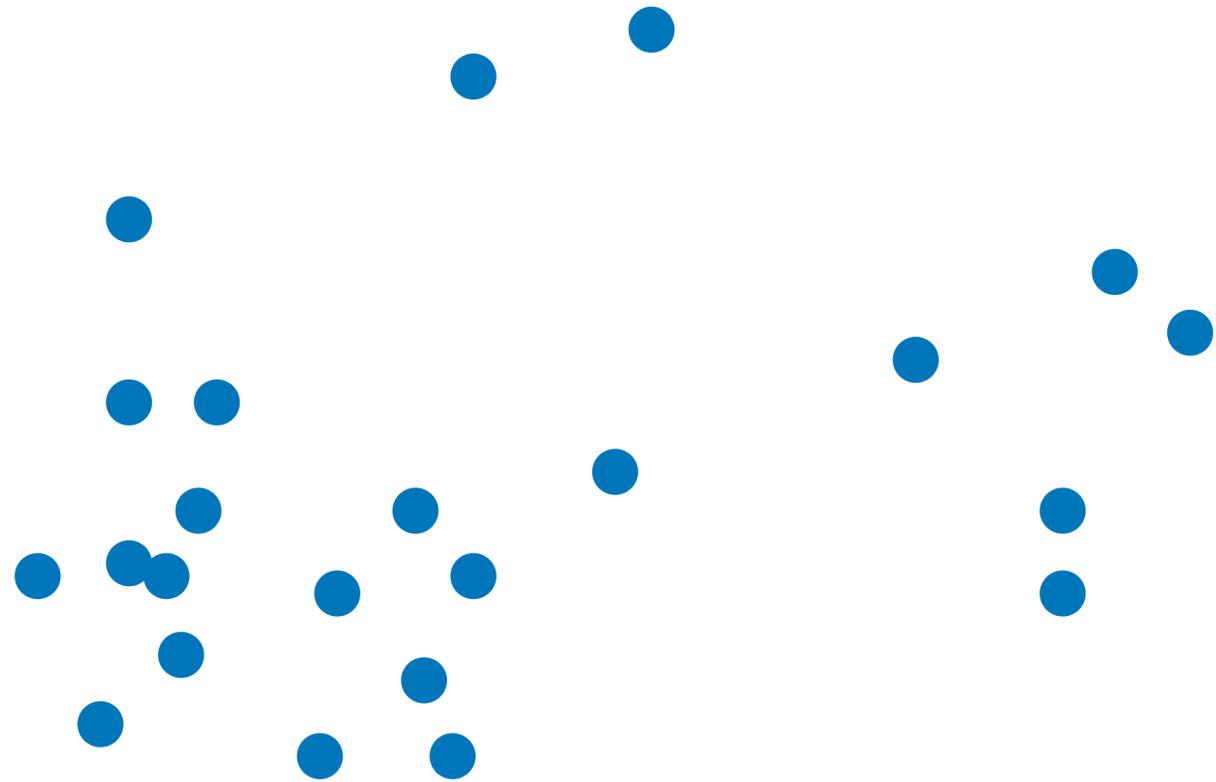
Junxian He
Mar 19, 2026

Midterm Exam

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

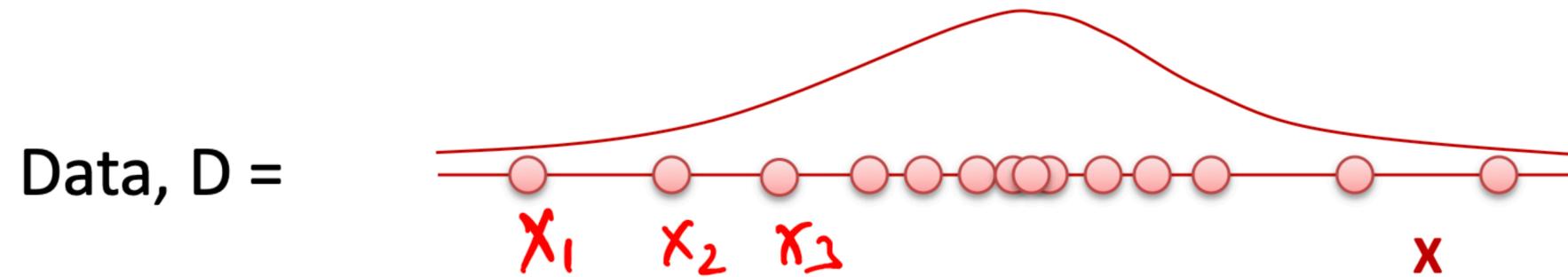
We want to model $p(x)$



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

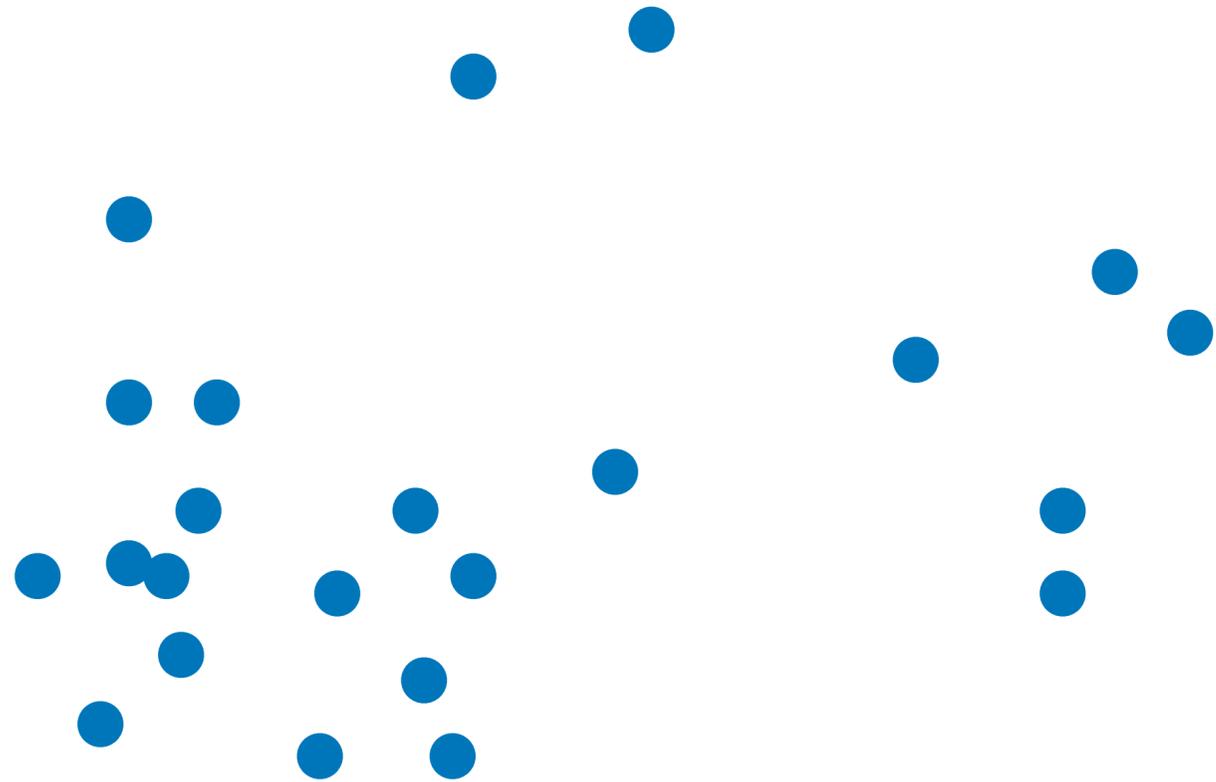
Recap: Generative Models



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

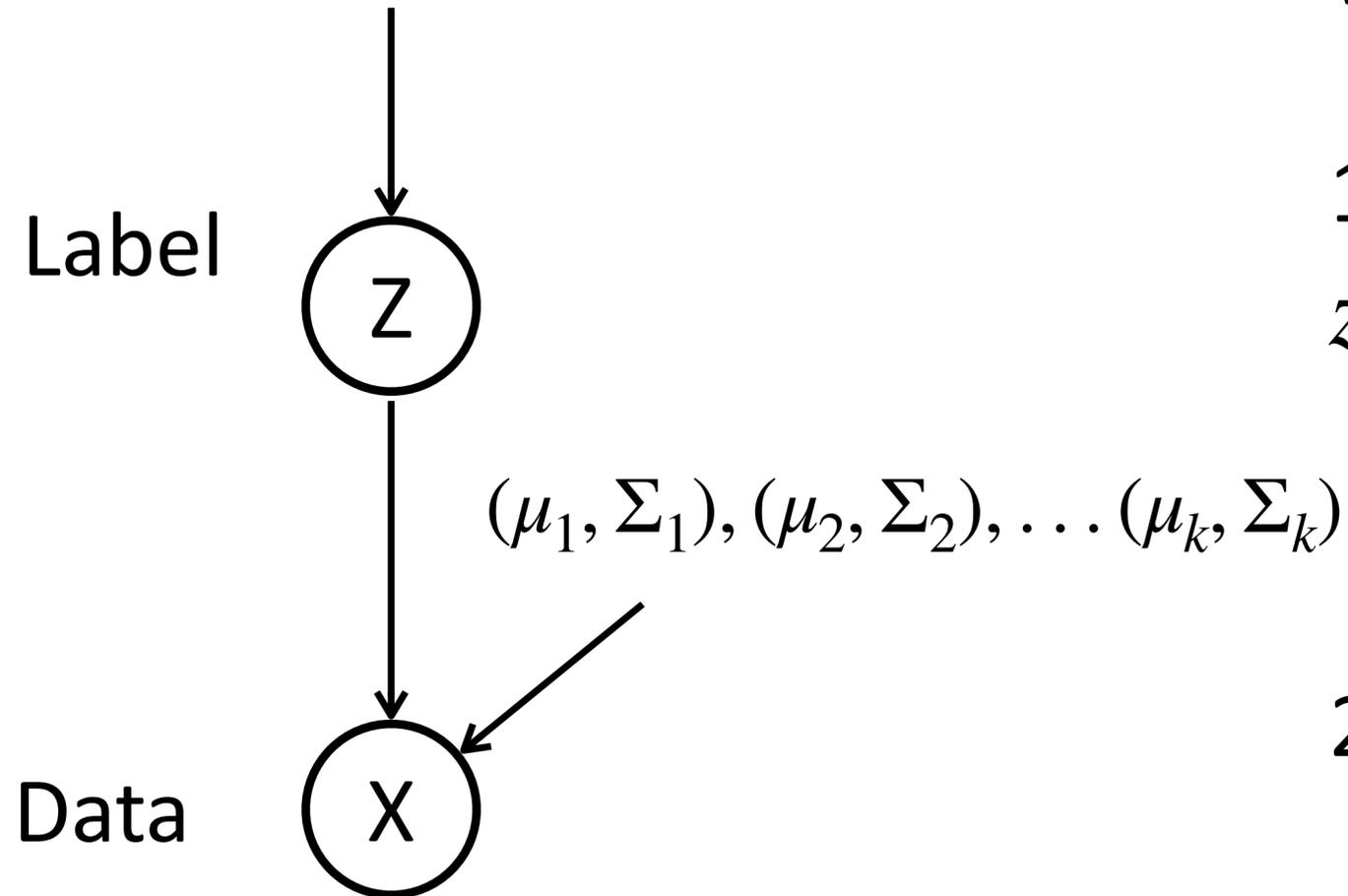


How to construct more complex distribution family?

Introducing more latent variables

Recap: Gaussian Mixture Model

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



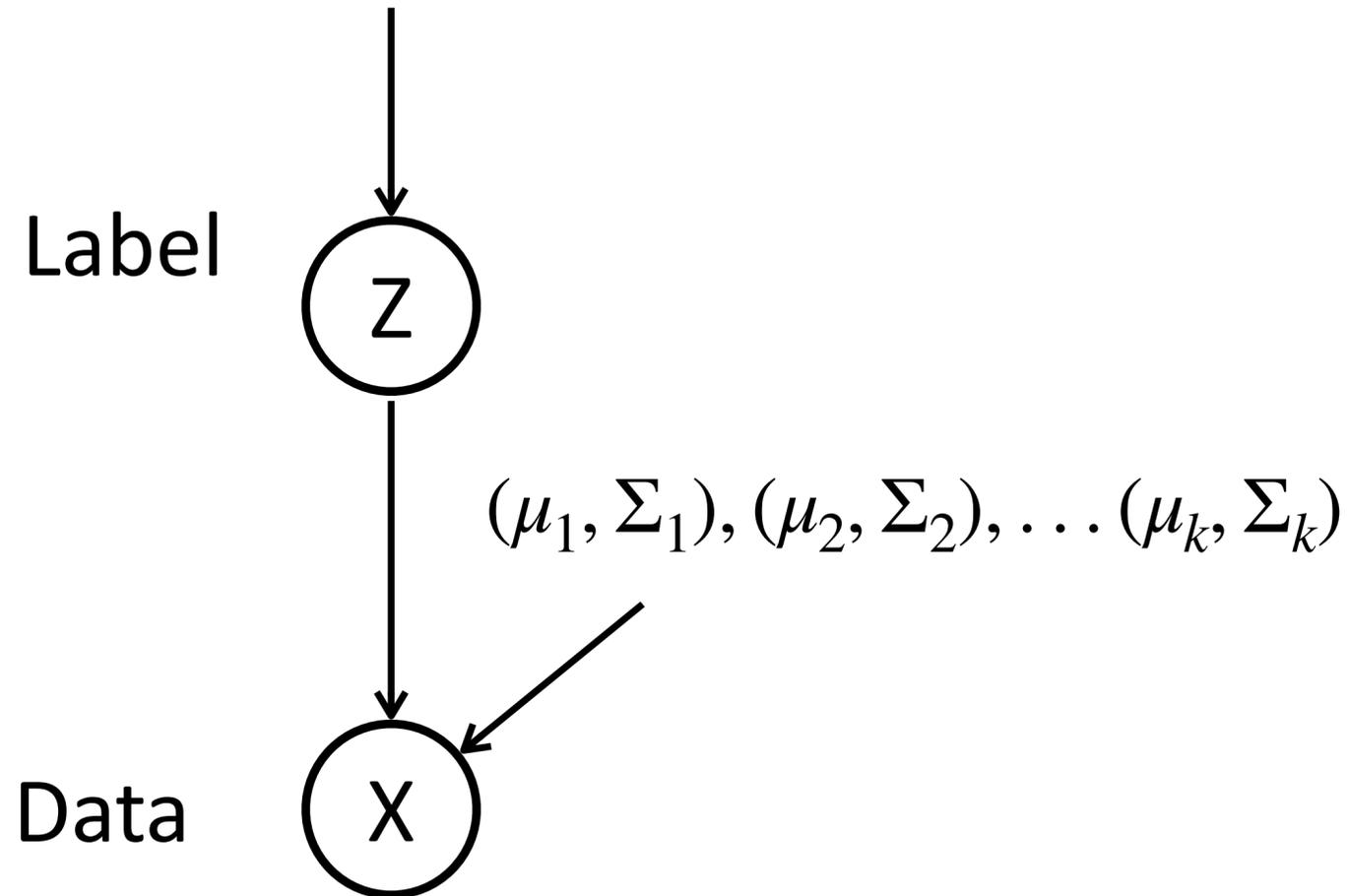
We assume the generative process as:

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:

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

How to compute this?

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

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

$$\phi_j = \frac{1}{n} \sum_{i=1}^n 1\{z^{(i)} = j\},$$

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

$$\Sigma_j = \frac{\sum_{i=1}^n 1\{z^{(i)} = j\} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^n 1\{z^{(i)} = j\}}.$$

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:

{

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

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

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

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

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:

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

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

Evidence Lower Bound (ELBO)

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

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

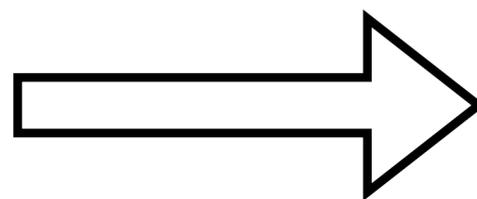
Why optimizing lower bound works? How to choose $Q(z)$, why we computed posterior in the E step, what is the benefit?

Evidence Lower Bound (ELBO)

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

When is the lower bound tight?

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



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

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

$$\begin{aligned} \ell(\theta) &\geq \sum_i \text{ELBO}(x^{(i)}; Q_i, \theta) \\ &= \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \end{aligned}$$

Evidence Lower Bound (ELBO)

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

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

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

$$\begin{aligned} \theta &:= \arg \max_{\theta} \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i, \theta) \\ &= \arg \max_{\theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}. \end{aligned}$$

$Q(z)$ 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 over θ

Why is maximizing lower-bound sufficient?

EM is Hill Climbing



$\log p(x; \theta)$

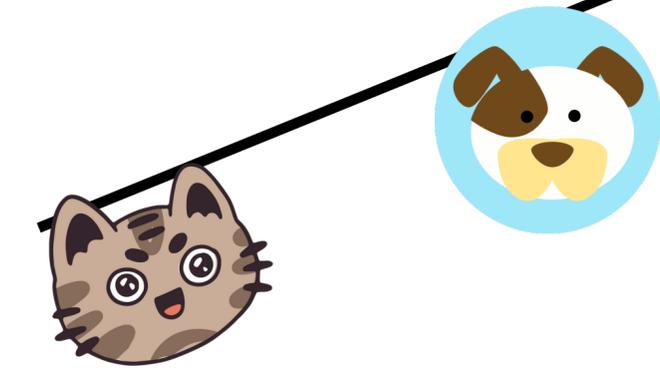
Only related to θ , no z



Larger



ELBO



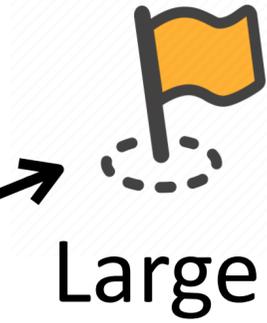
EM is Hill Climbing



$\log p(x; \theta)$



ELBO



E-step: $Q(z) = p(z | x; \theta)$, 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



Larger

EM is Hill Climbing



$\log p(x; \theta)$



ELBO



Larger

E-step: $Q(z) = p(z | x; \theta)$, making ELBO tight
“dog” doesn’t change, because θ does not change

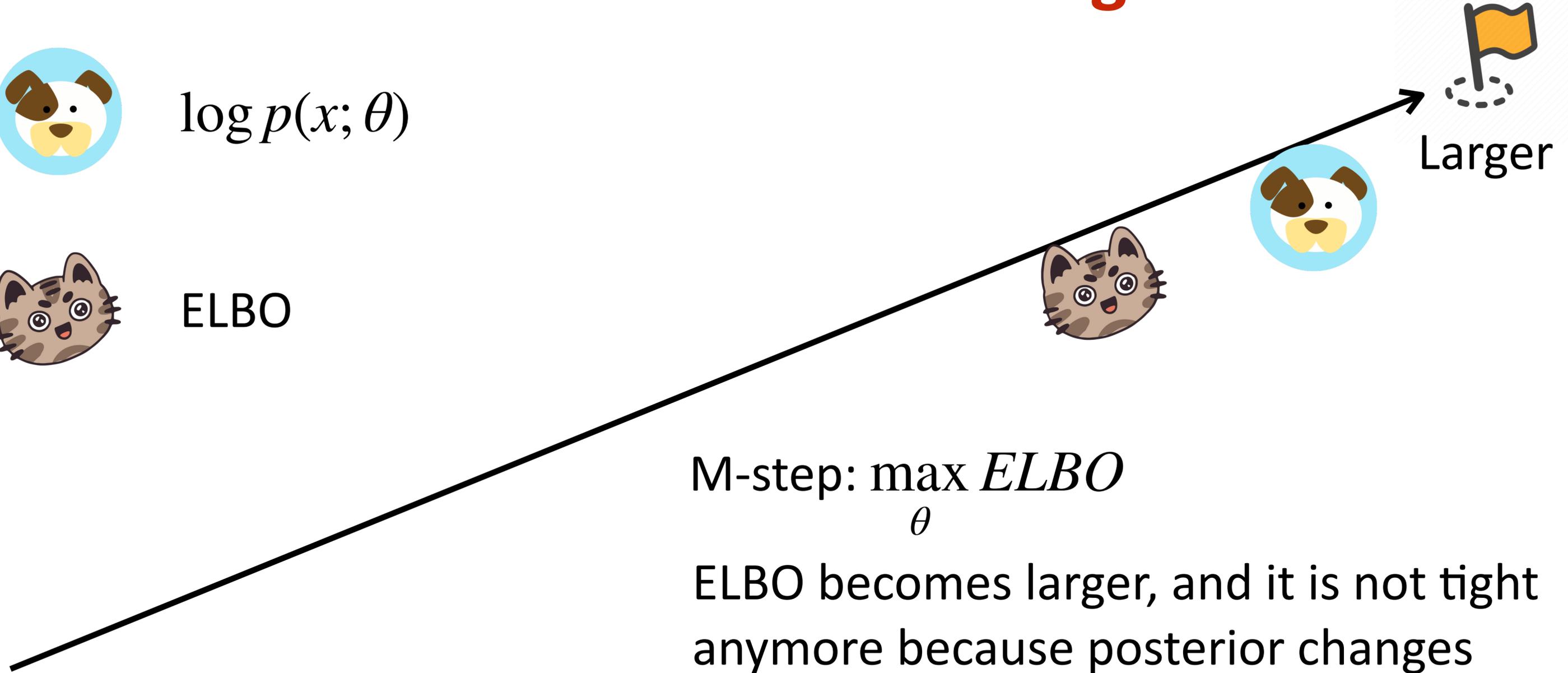
EM is Hill Climbing



$\log p(x; \theta)$



ELBO



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

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

$\log p(x; \theta)$ 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

$$\begin{aligned} \theta &:= \arg \max_{\theta} \sum_{i=1}^n \text{ELBO}(x^{(i)}; Q_i, \theta) \\ &= \arg \max_{\theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}. \end{aligned}$$

}

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

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

Sometimes the sum is computable, but sometimes not

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

We can use Monte-Carlo sampling to approximate the expectation

Comparing Direct Maximization and EM

Direct maximization:

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

M-Step in EM:

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

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

Other Interpretations of ELBO

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

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

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

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

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