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

COMP 5212
Machine Learning
Lecture 12

Expectation Maximization

Junxian He
Oct 17, 2024

Midterm Exam

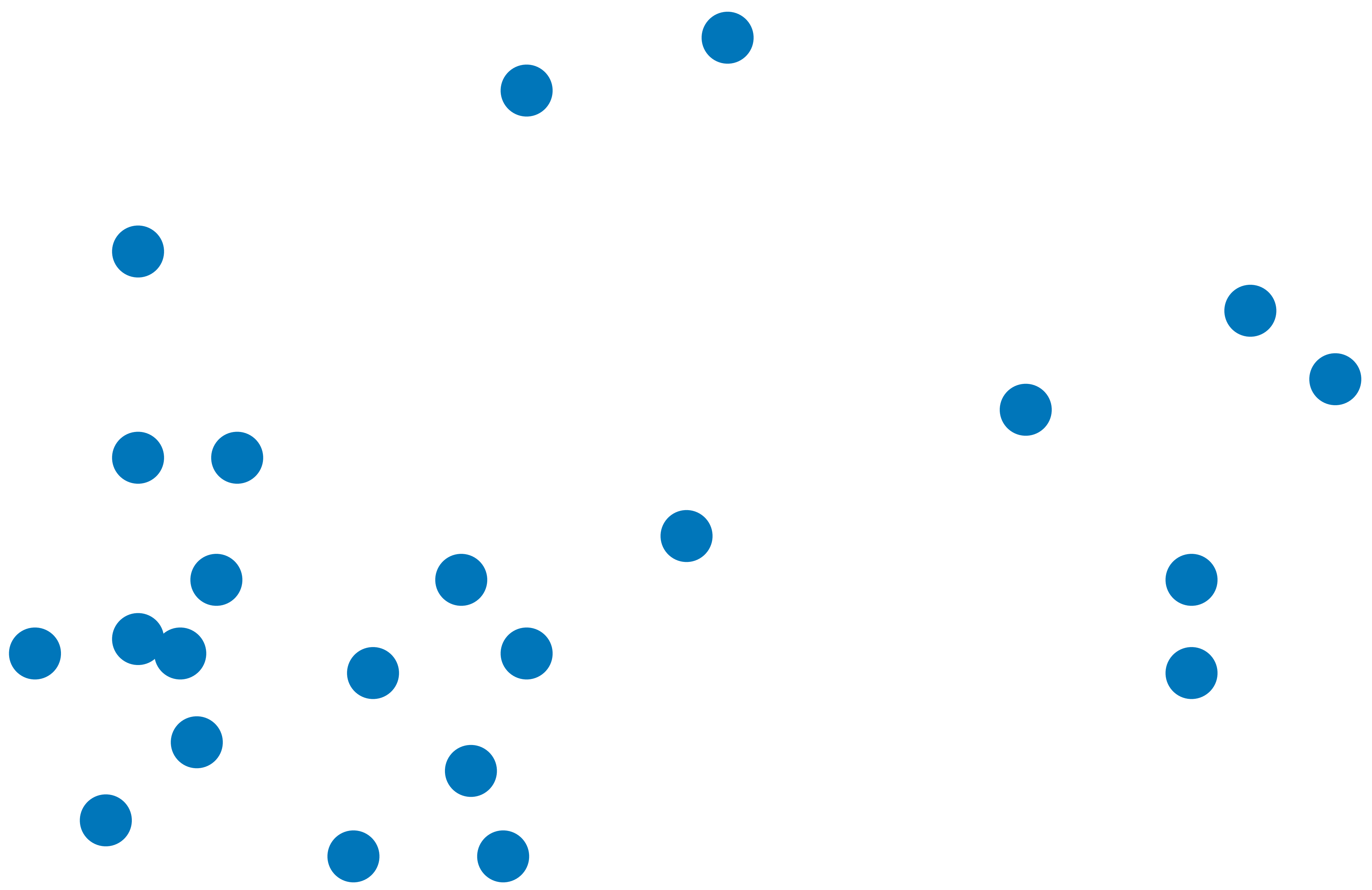
Next Thursday (Oct 24), 120pm-240pm, one A4-size double-sided cheetsheet is allowed (either printing or handwriting is fine)

We have two rooms for the exam for sparse seat plans:

1. For SIS ID ending with an even digit: Room 2303
2. For SIS ID ending with an odd digit: Room 2504

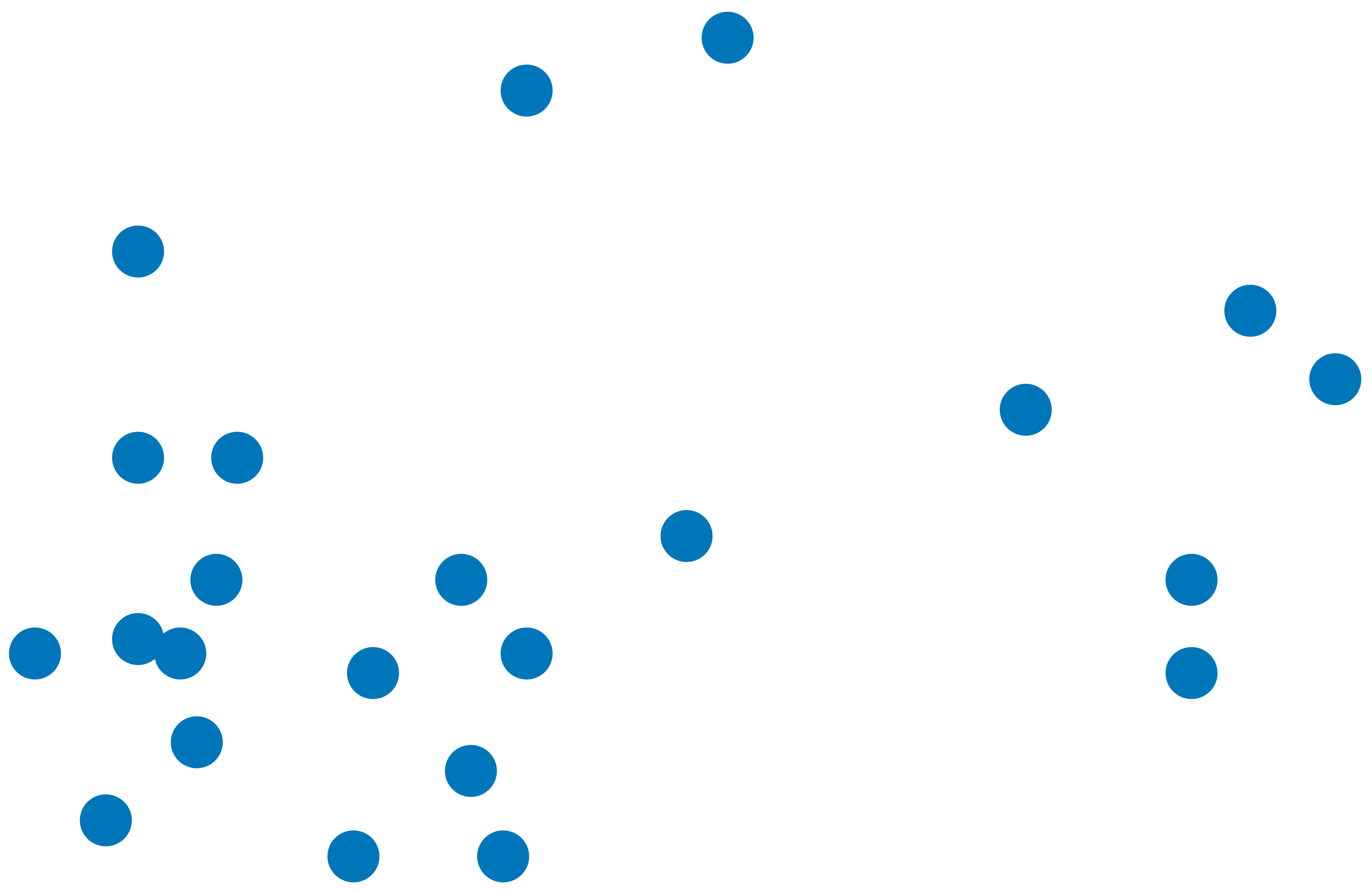
Recap: Generative Models

Recap: Generative Models



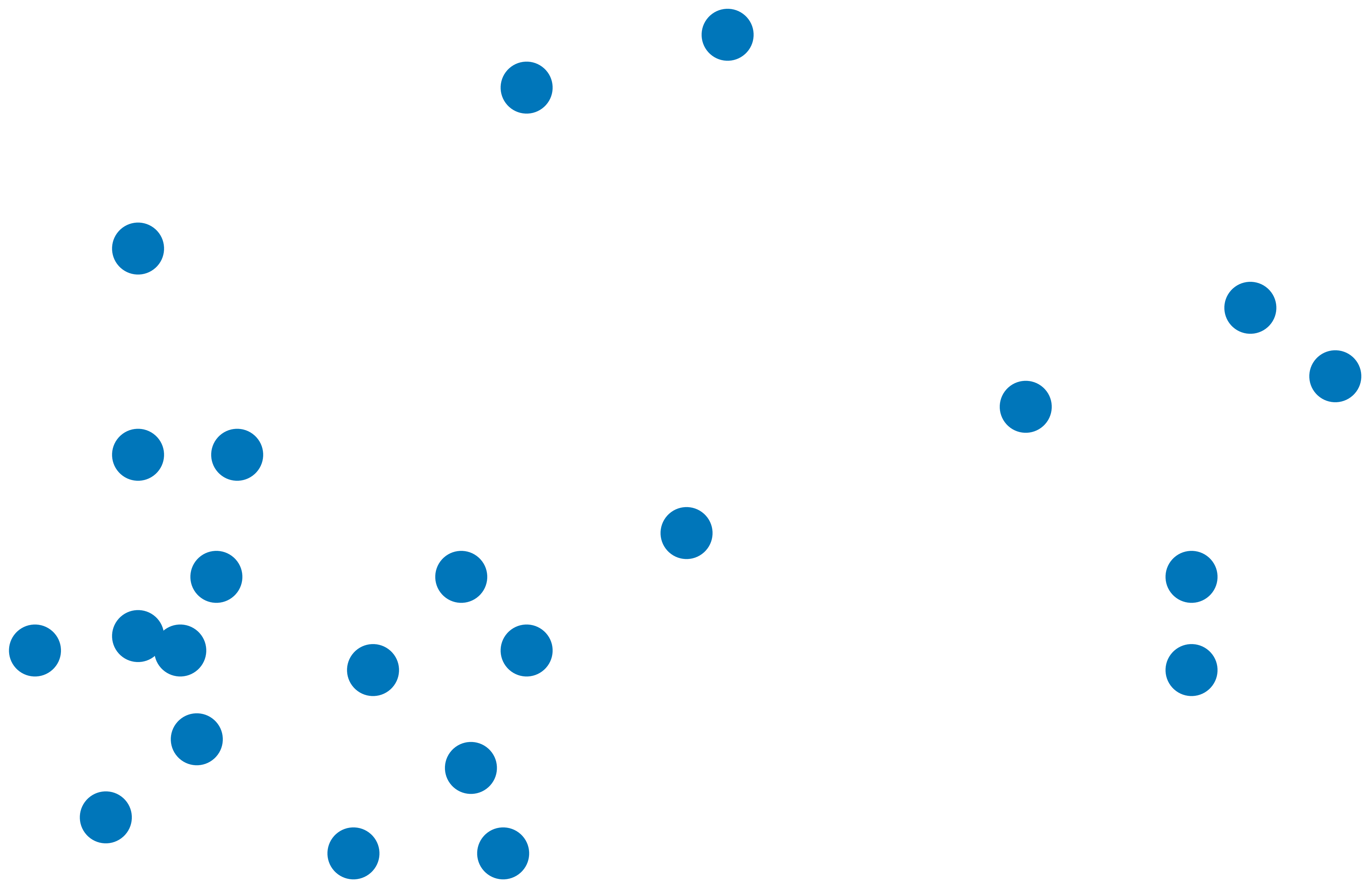
Recap: Generative Models

We want to model $p(x)$



Recap: Generative Models

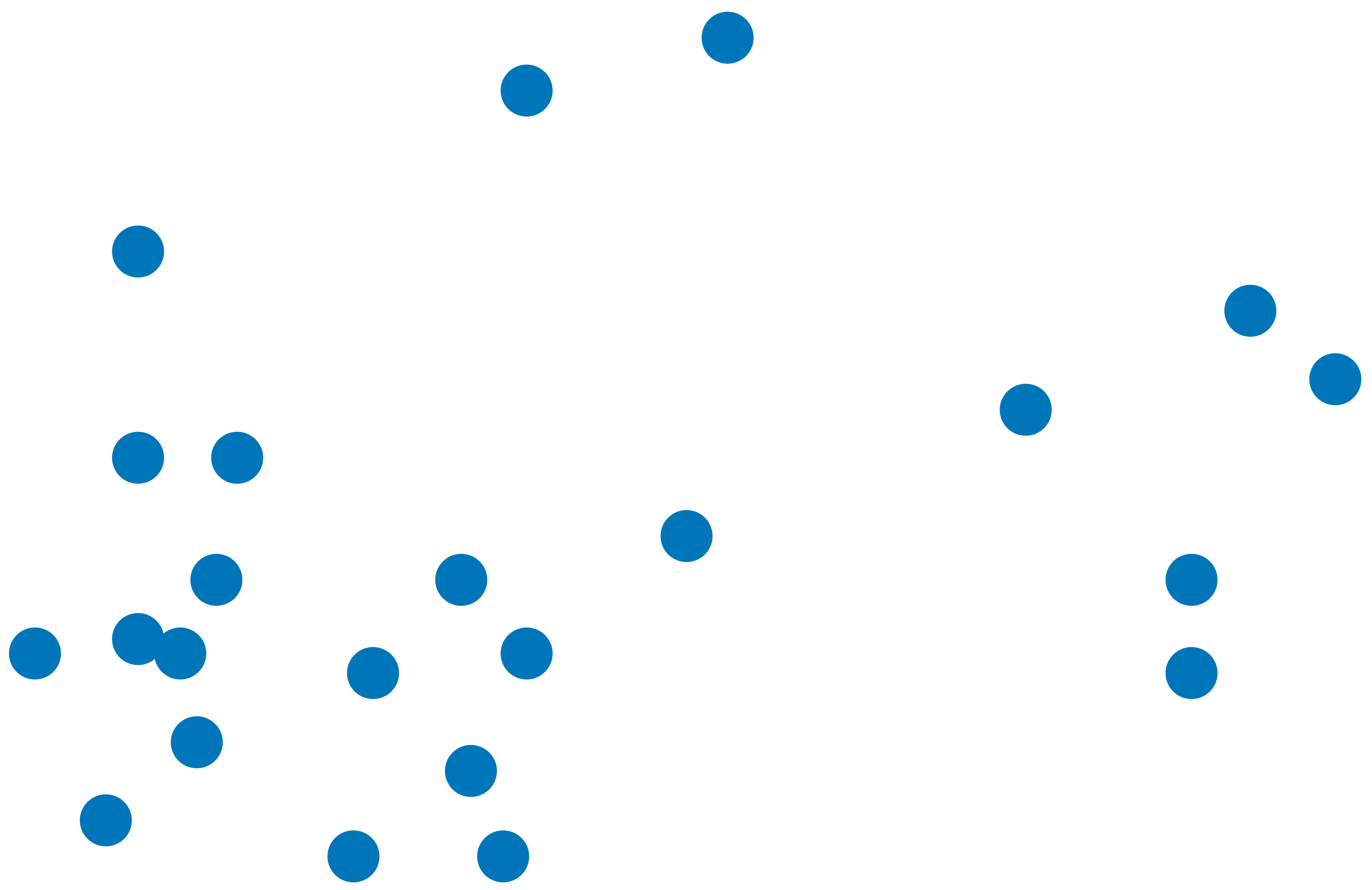
We want to model $p(x)$



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 $p(x)$



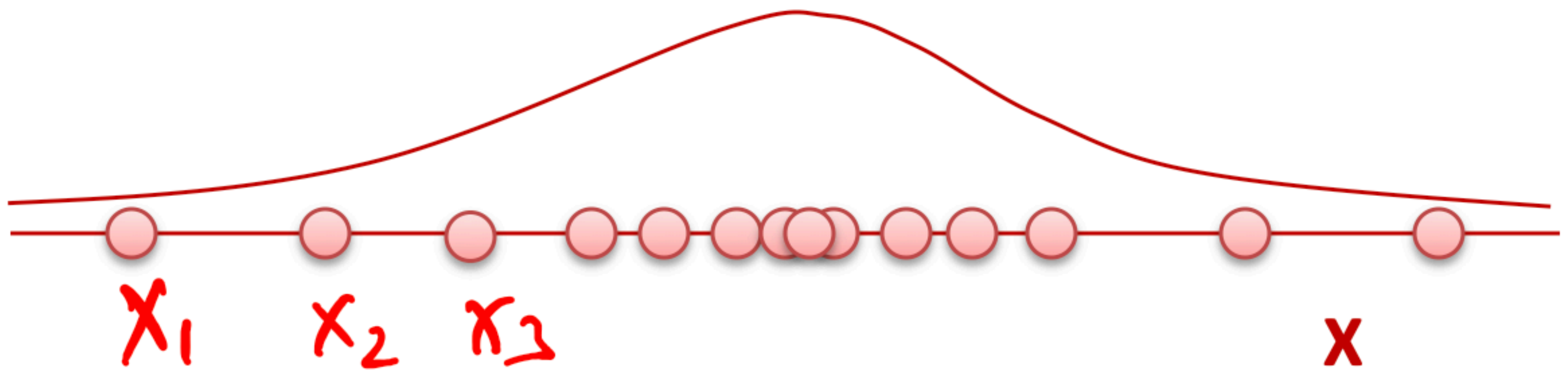
In discriminative models, we need to "design" model to make assumption about the function: linear regression, logistic regression, kernel methods ….

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

Recap: Generative Models

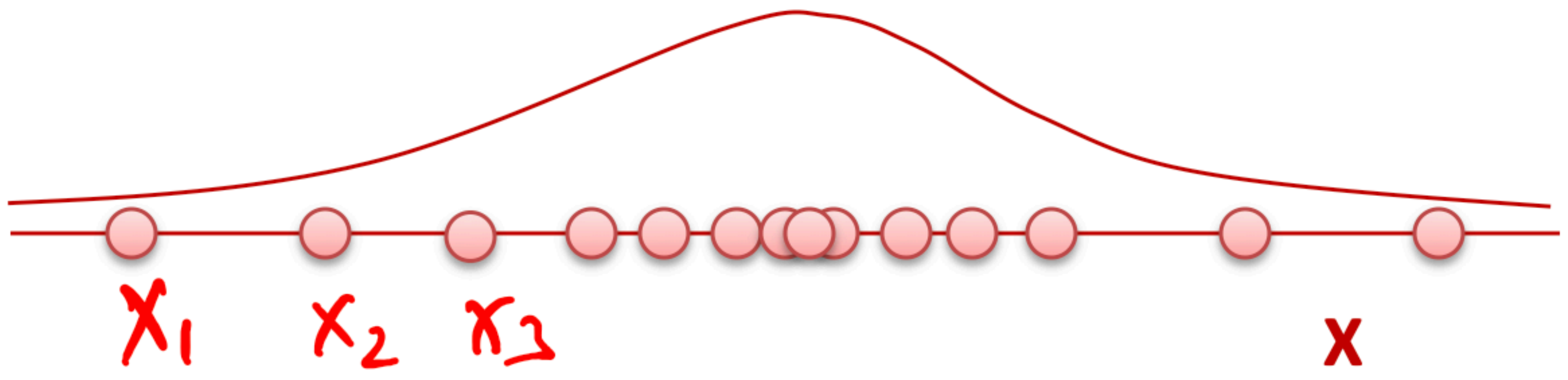
Recap: Generative Models

Data, D =



Recap: Generative Models

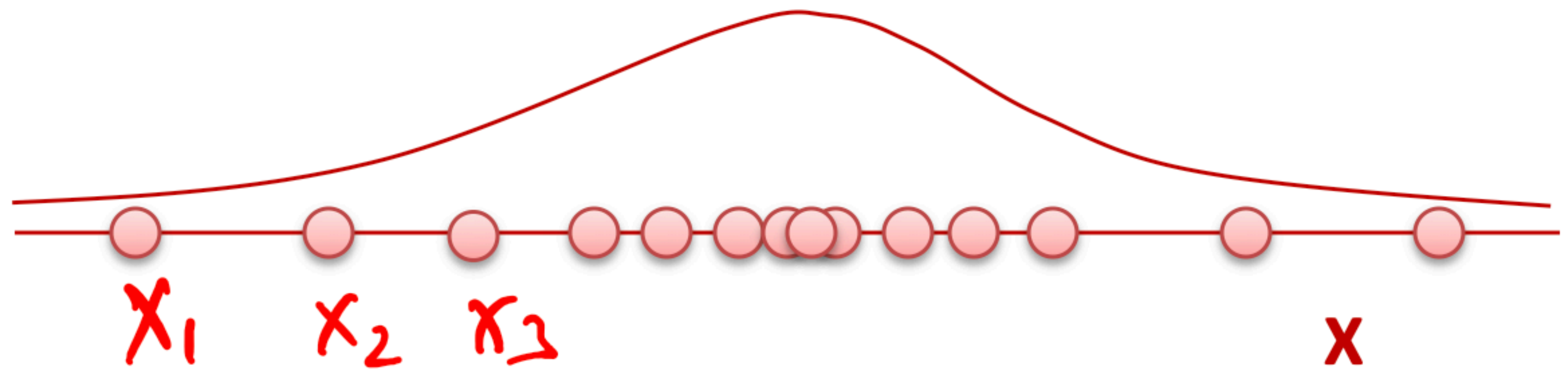
Data, D =



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

Recap: Generative Models

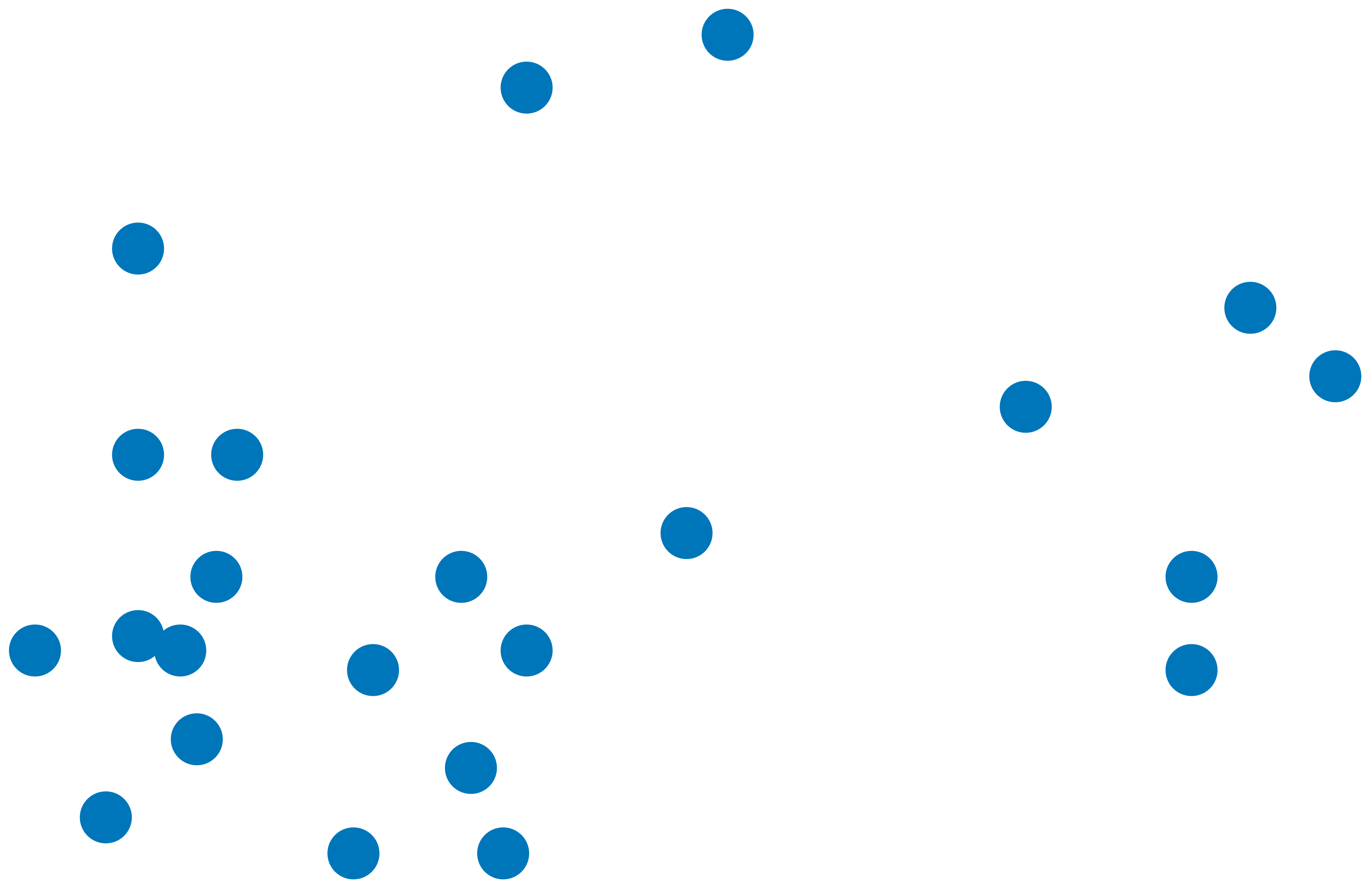
Data, D =



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

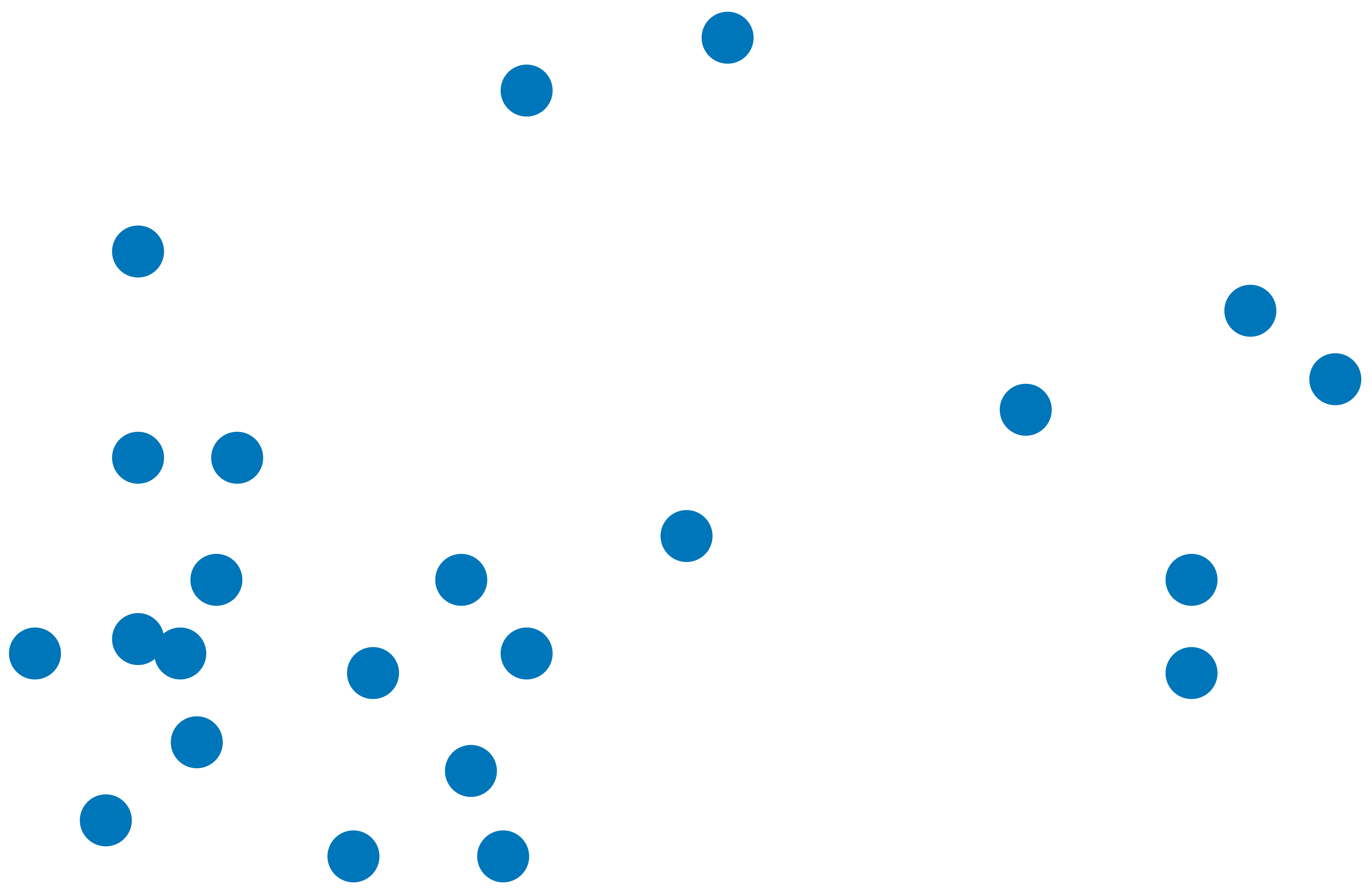
By varying the parameters (μ, Σ), the model represents different distributions that belong to the Gaussian family

Recap: Generative Models



How to construct more complex distribution family?

Recap: Generative Models

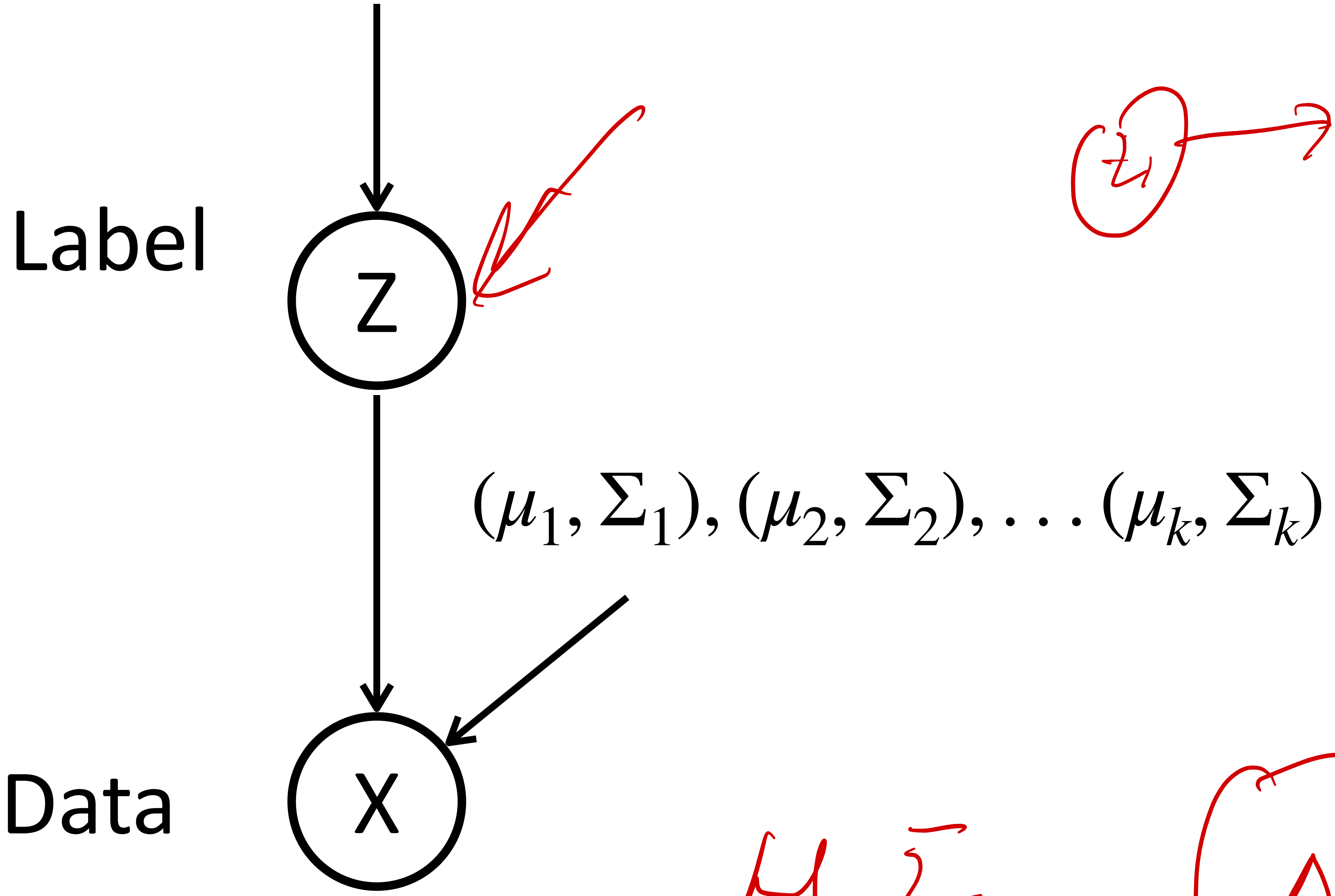


How to construct more complex distribution family?

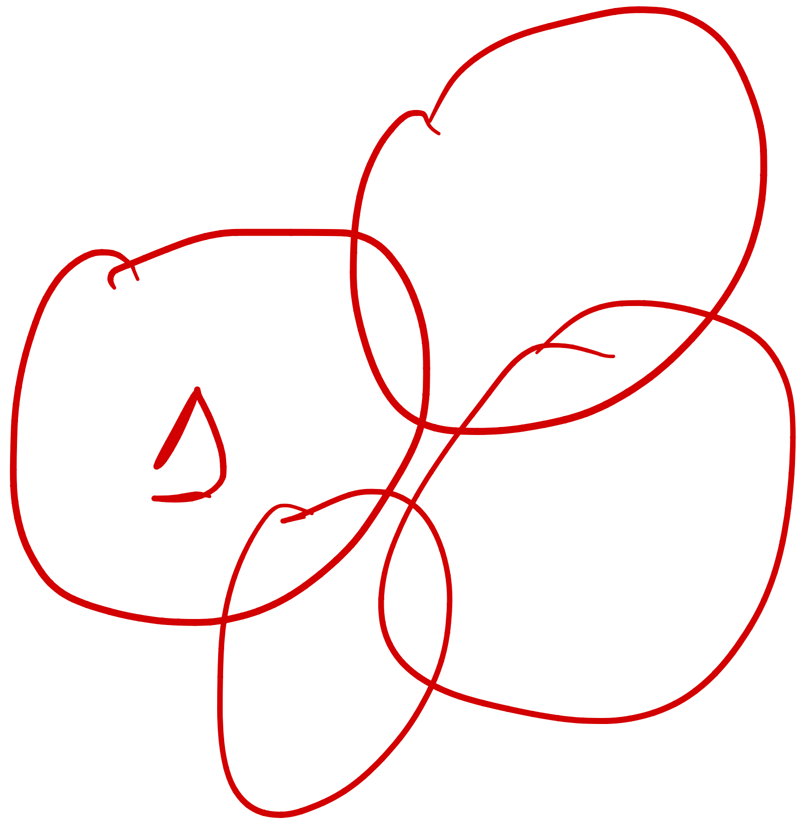
Introducing more latent variables

Recap: Gaussian Mixture Model

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

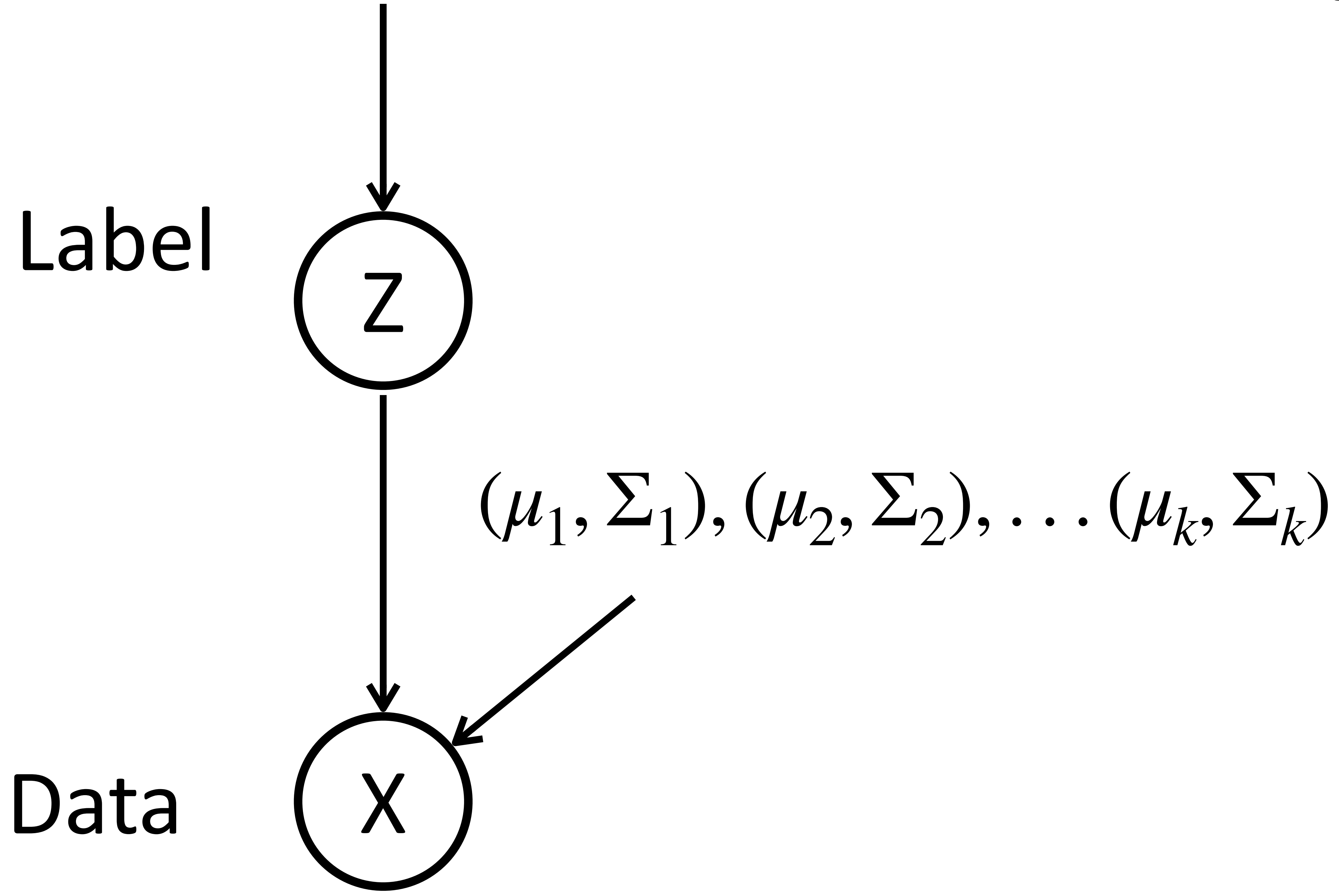


μ, Σ



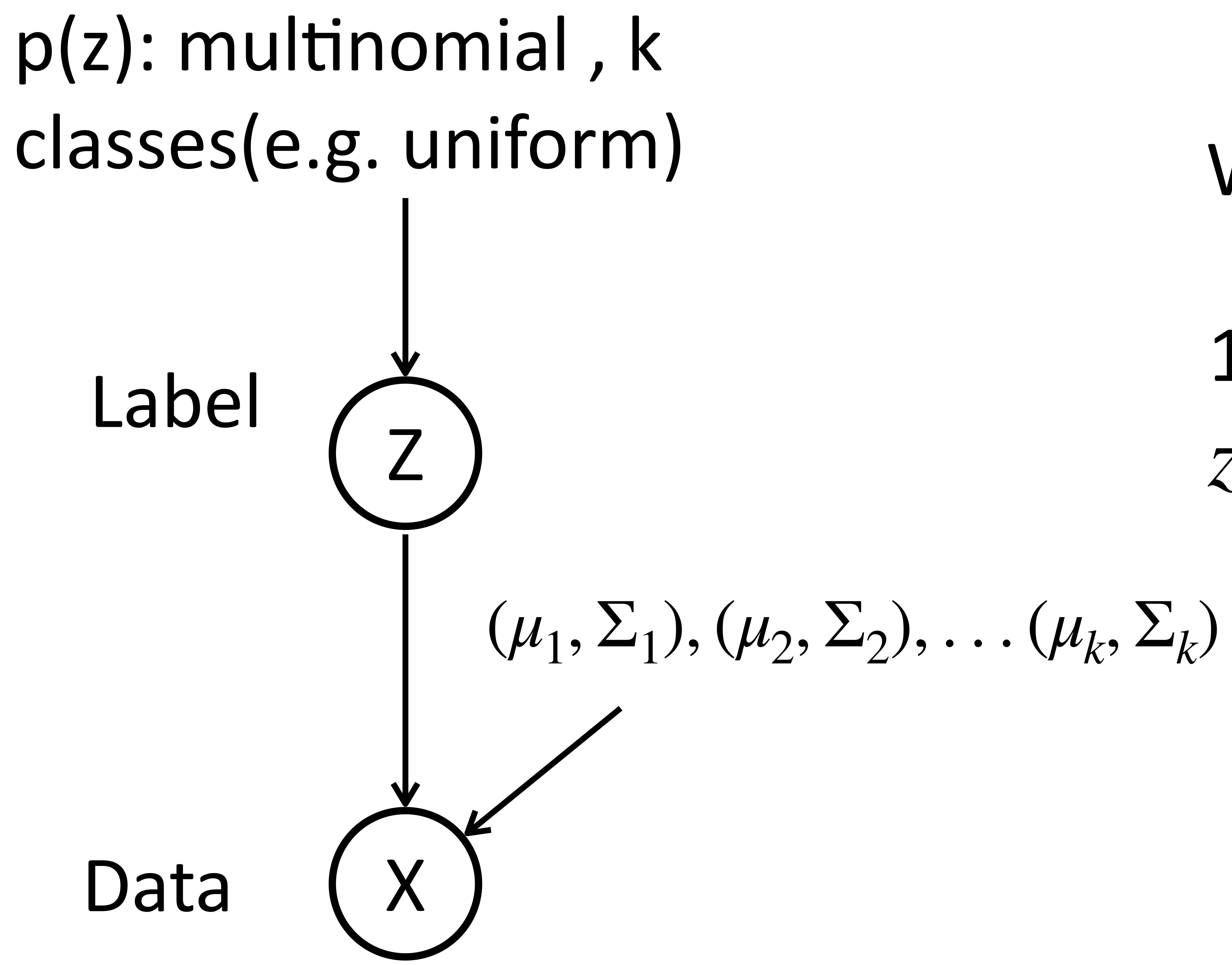
Recap: Gaussian Mixture Model

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



We assume the generative process as:

Recap: Gaussian Mixture Model

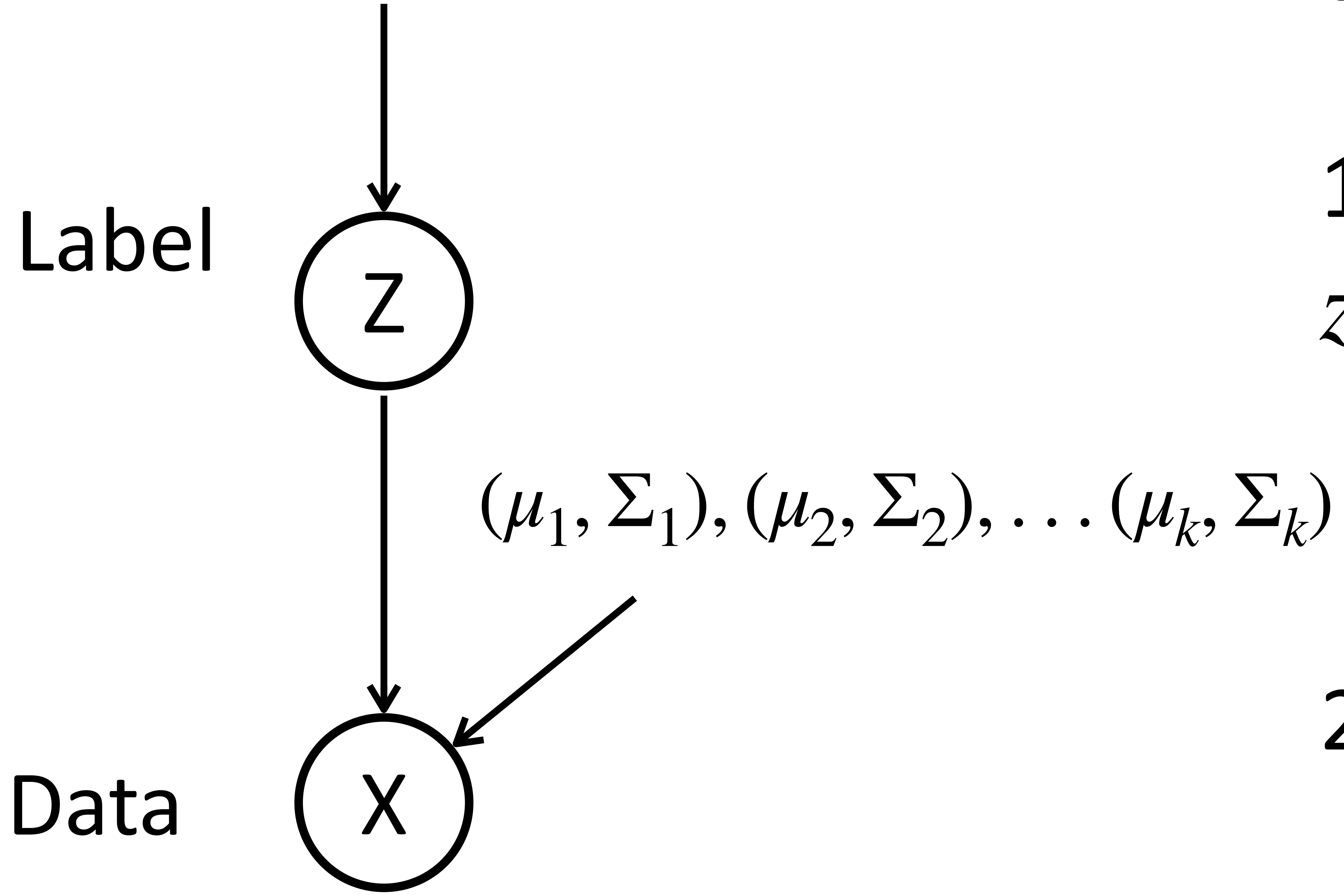


We assume the generative process as:

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)



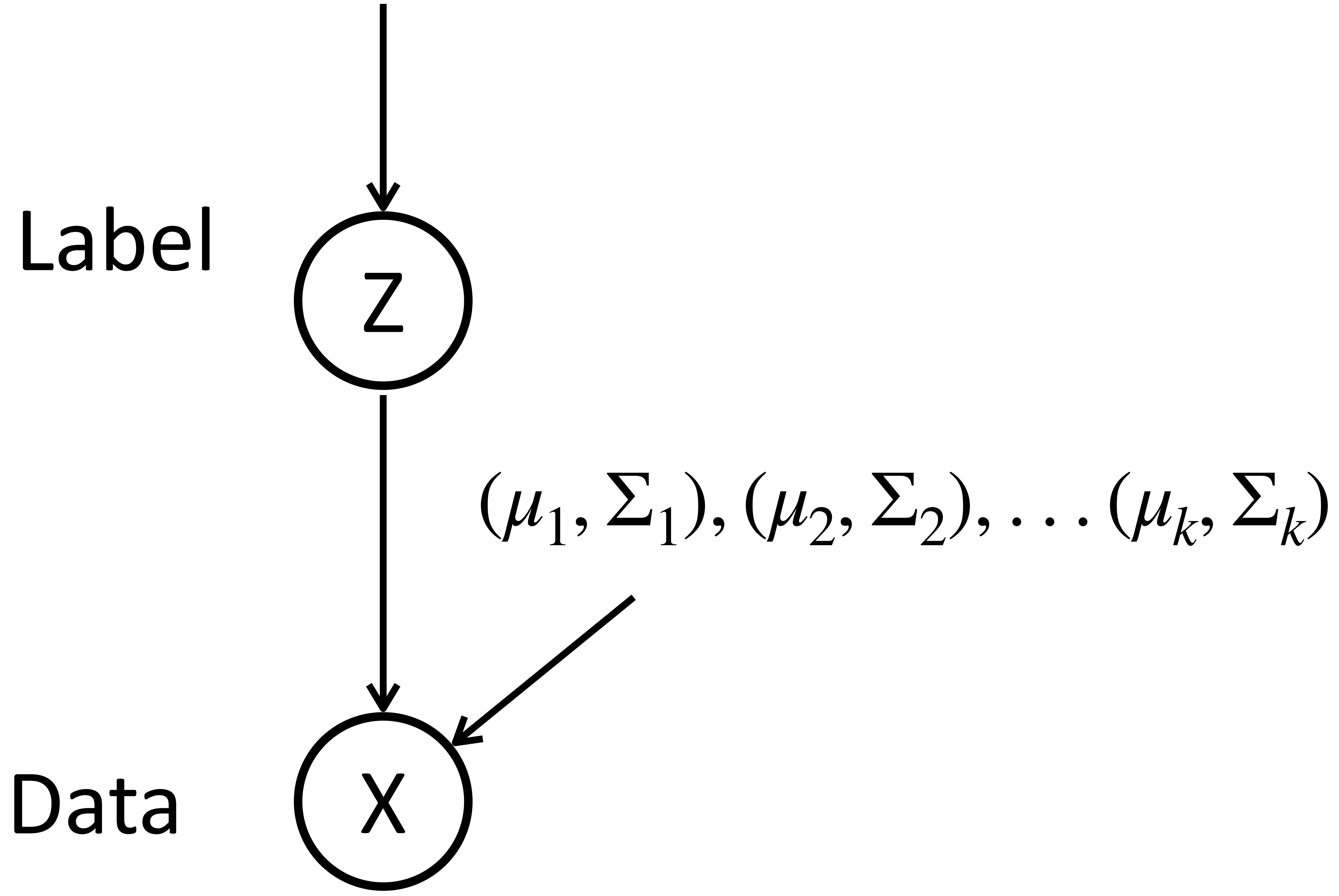
We assume the generative process as:

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)



$\phi \rightarrow p(z)$

Unsupervised:

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

→ hard

How to compute this?

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

easy

Recap: MLE for GMM

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

marginalization

Recap: MLE for GMM

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

z continuoses $\log \int_z \cdots$

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

Recap: MLE for GMM

$$\sum_z f \cdot p(z) = E_{z\sim p(z)} f$$

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

$$\log \left[E_{z\sim p(z)}\, p(x|z) \right]$$

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

Recap: MLE for GMM

log P(x)

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

GANs VAEs diffusion

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

log p(x, z)

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

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 iteratively

Expectation Maximization for GMM

Repeat until convergence:

{

}

Expectation Maximization for GMM

Repeat until convergence:

{

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

inference

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

}

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?

is EM still MLE?

- 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

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

$$\underbrace{z_1 \to z_2 \to z_3 \to z_4}_{Z}$$

General EM Algorithm

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

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

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

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

Jensen inequality

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

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

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

ELBO

This lower bound holds for any Q(z)

Let Q to be a distribution over z

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

tight

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:

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

$$f(P_1 x_1 + P_2 x_2 + \cdots P_k x_k) \leq \sum P_i f(x_i)$$

$$\sum P_i = 1$$

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:

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

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

Log P(x)

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

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

max log P(x)

max 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

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

EM

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)

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

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?

X is constant

Evidence Lower Bound (ELBO)

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

When is the lower bound tight?

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

$f(x)$

no matter what z is

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

$$\propto p(x,z)$$

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

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?

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

Evidence Lower Bound (ELBO)

Evidence Lower Bound (ELBO)

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

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

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

$$= \log P(\cancel{x}) \sum_z P(z|x) = 1$$

$$= \log P(x)$$

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)

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

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

$Q(z) = ?$

$\log P(x) \geq \mathrm{ELBO}(x; Q, \theta)$

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

$\log P(x)$ is constant varying $Q(z)$

$\log P(x)$

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

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

}

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

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

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

}

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

Q(x) is fixed

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

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

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

}

The General EM Algorithm

$\nearrow$ E-step

ELBO$(\theta, Q(z))$

$\searrow$ M-step

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

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

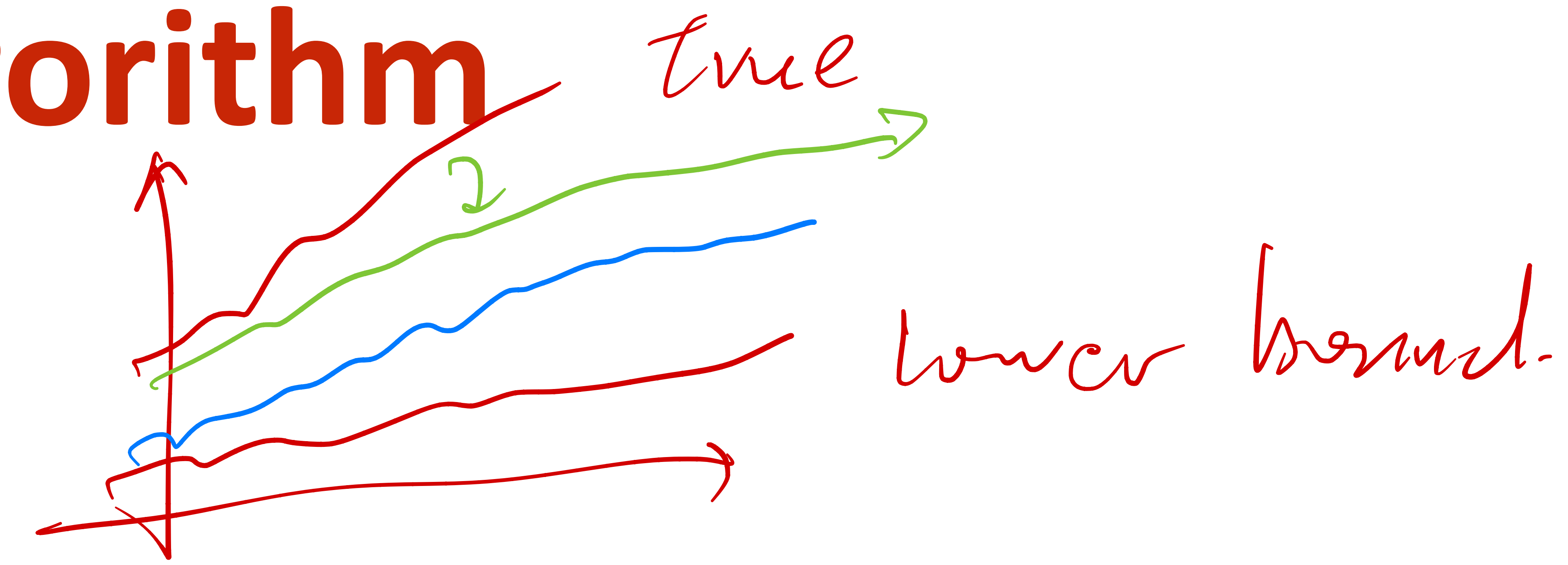
$$= \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}.$$

$Q(z)$ 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

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

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

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

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

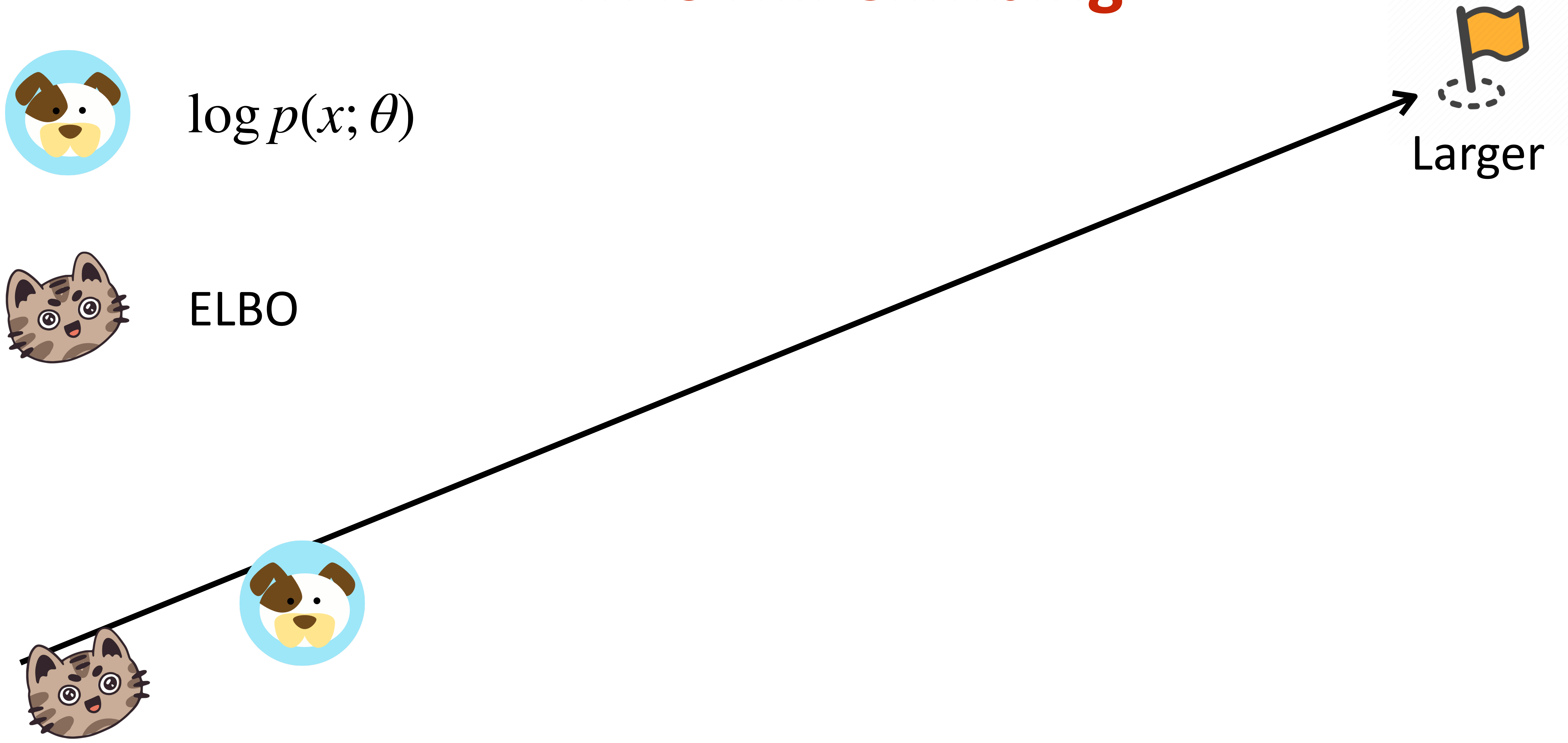
}

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

Why is maximizing lower-bound sufficient?

$\boxed{\log P(x)}$

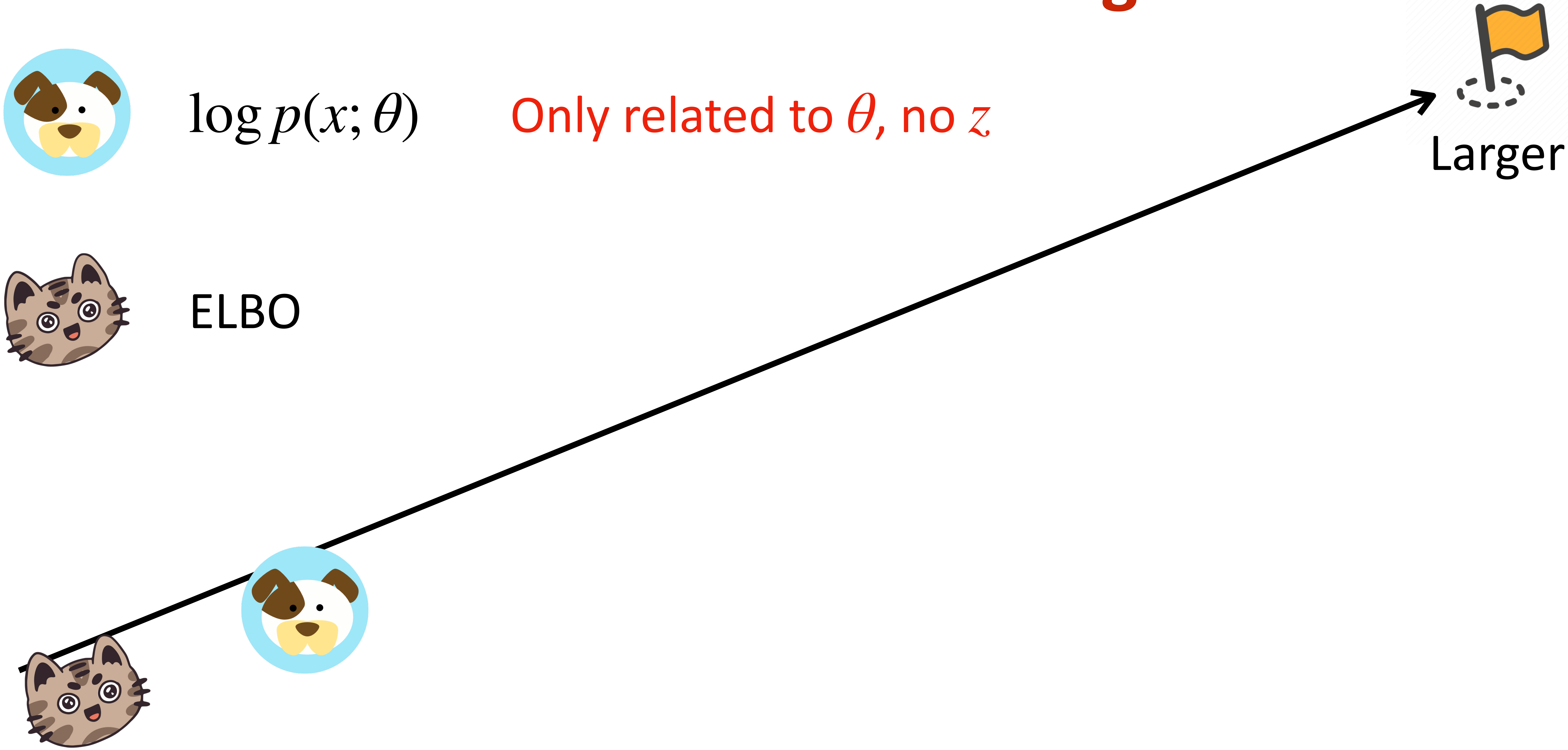
EM is Hill Climbing



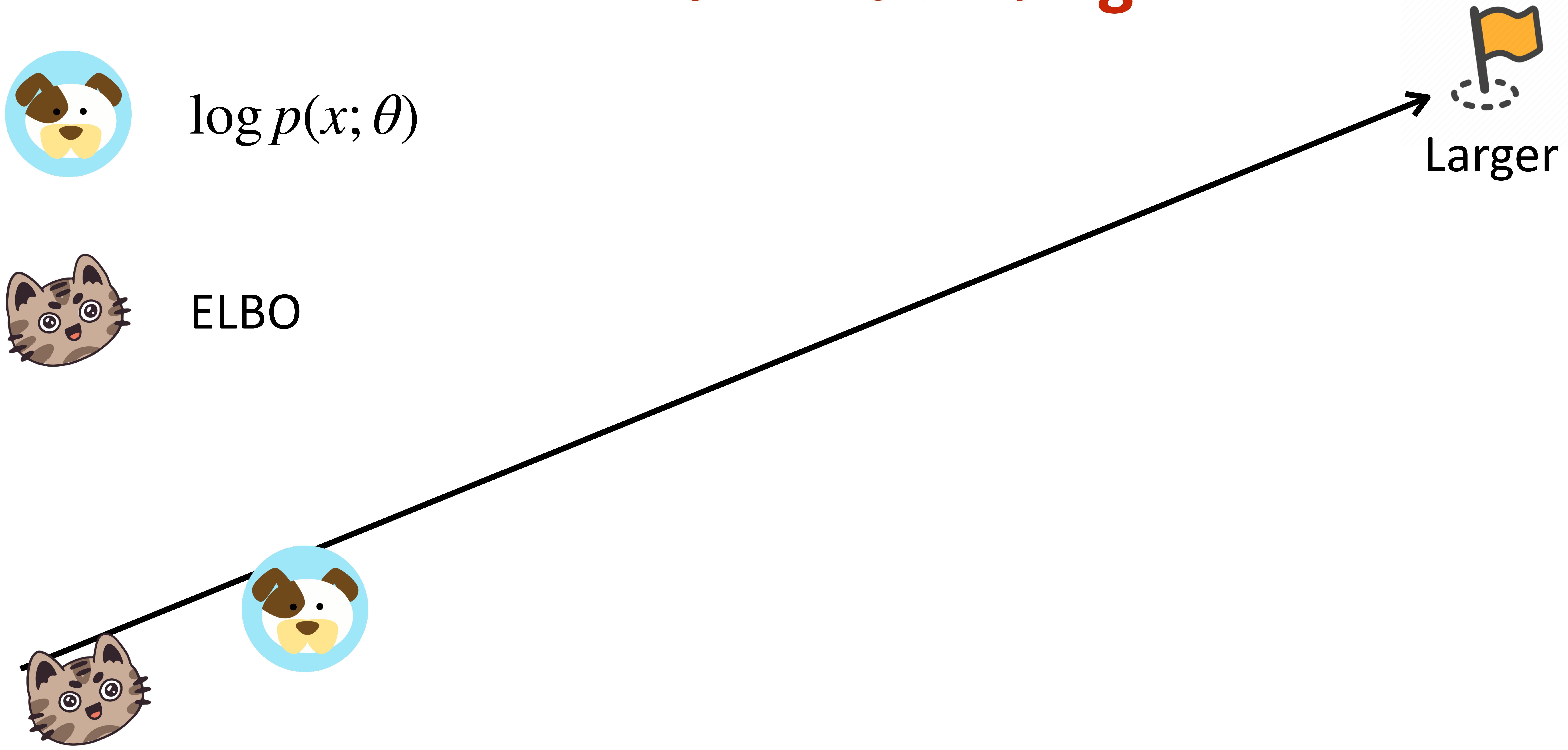
EM is Hill Climbing

$\log p(x;\theta)$ Only related to θ, no z

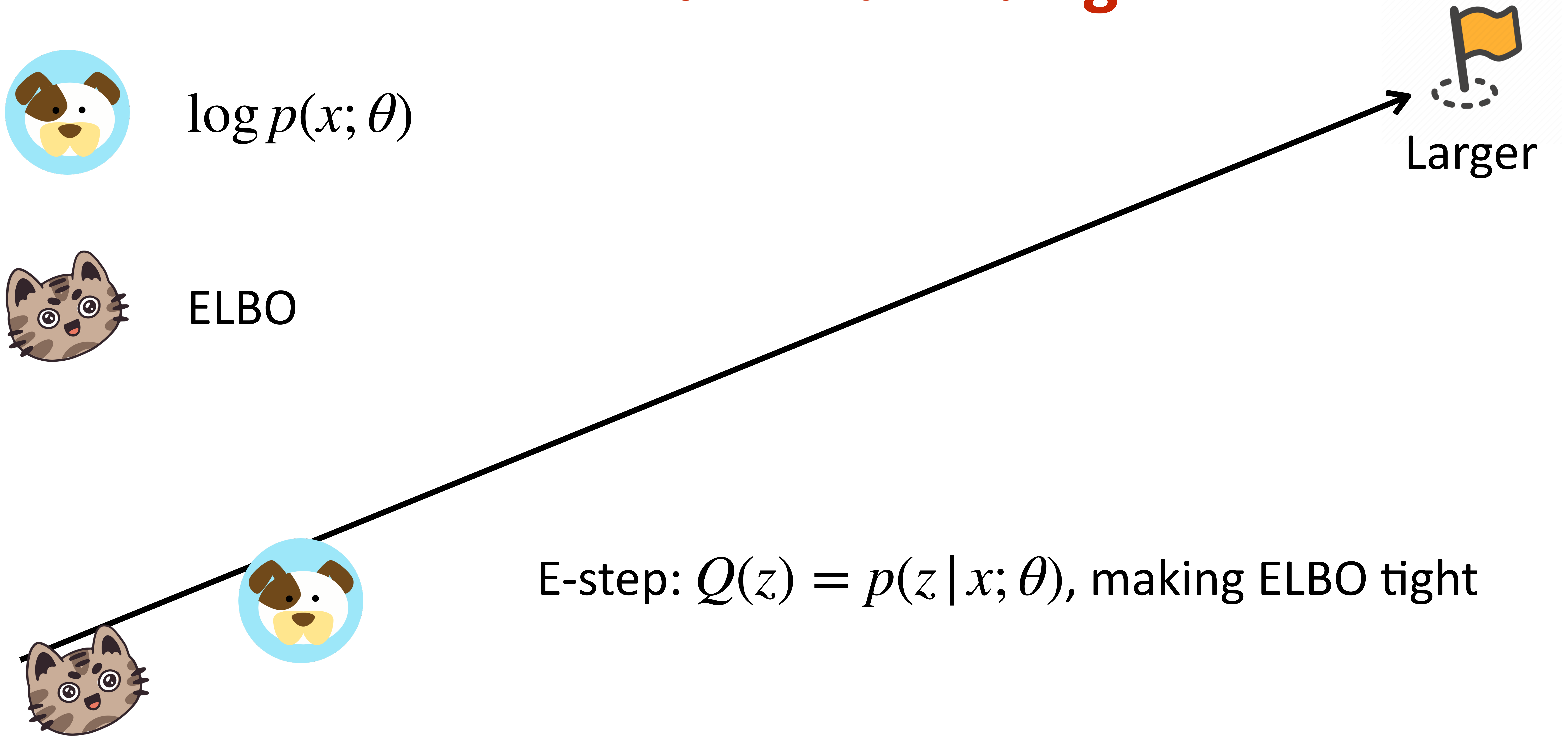
ELBO



EM is Hill Climbing



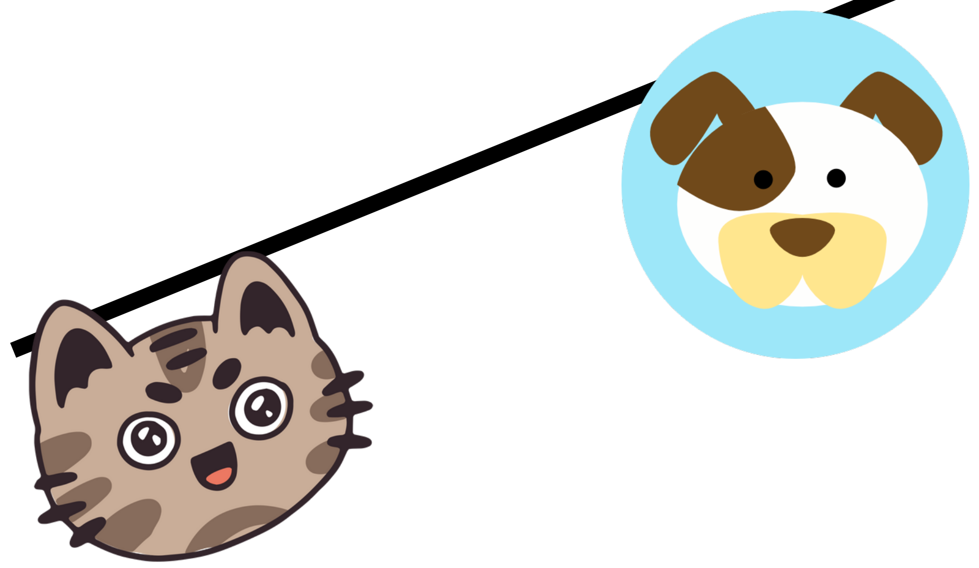
EM is Hill Climbing



EM is Hill Climbing

$\log p(x;\theta)$

ELBO



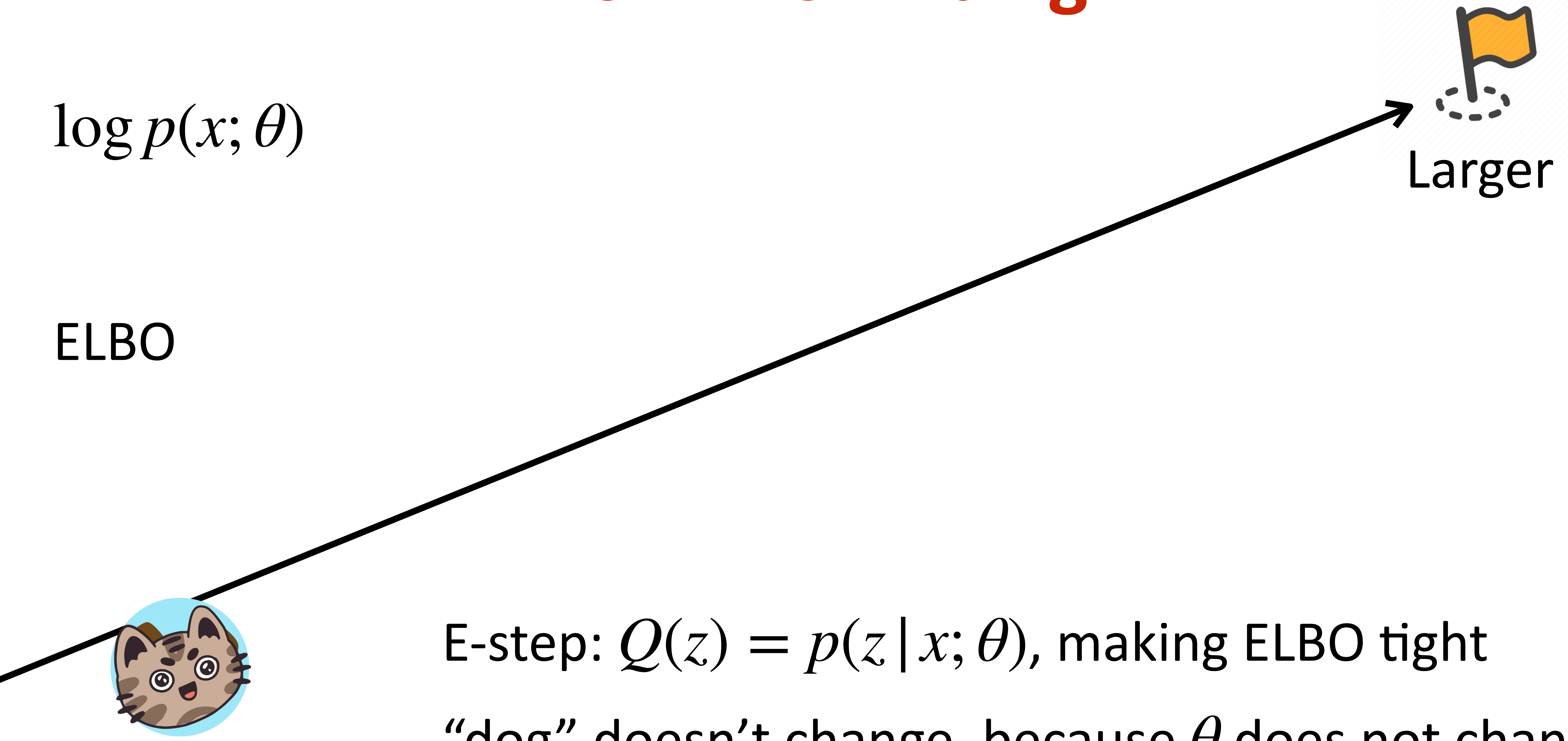
E-step: $Q(z) = p(z \mid x;\theta)$, making ELBO tight

“dog” doesn’t change, because θ does not change

EM is Hill Climbing

$\log p(x;\theta)$

ELBO



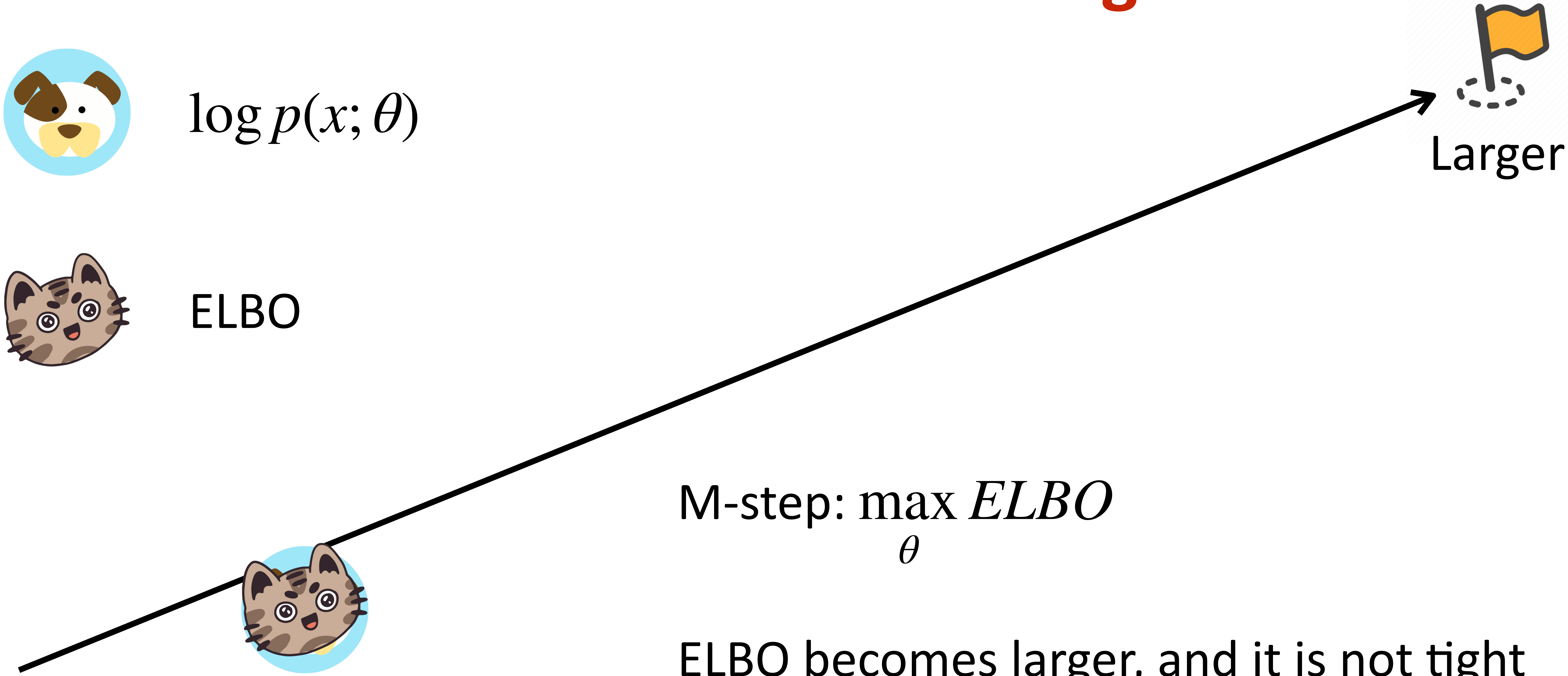
E-step: $Q(z) = p(z \mid x;\theta)$, making ELBO tight

“dog” doesn’t change, because θ does not change

EM is Hill Climbing

$\log p(x;\theta)$

ELBO



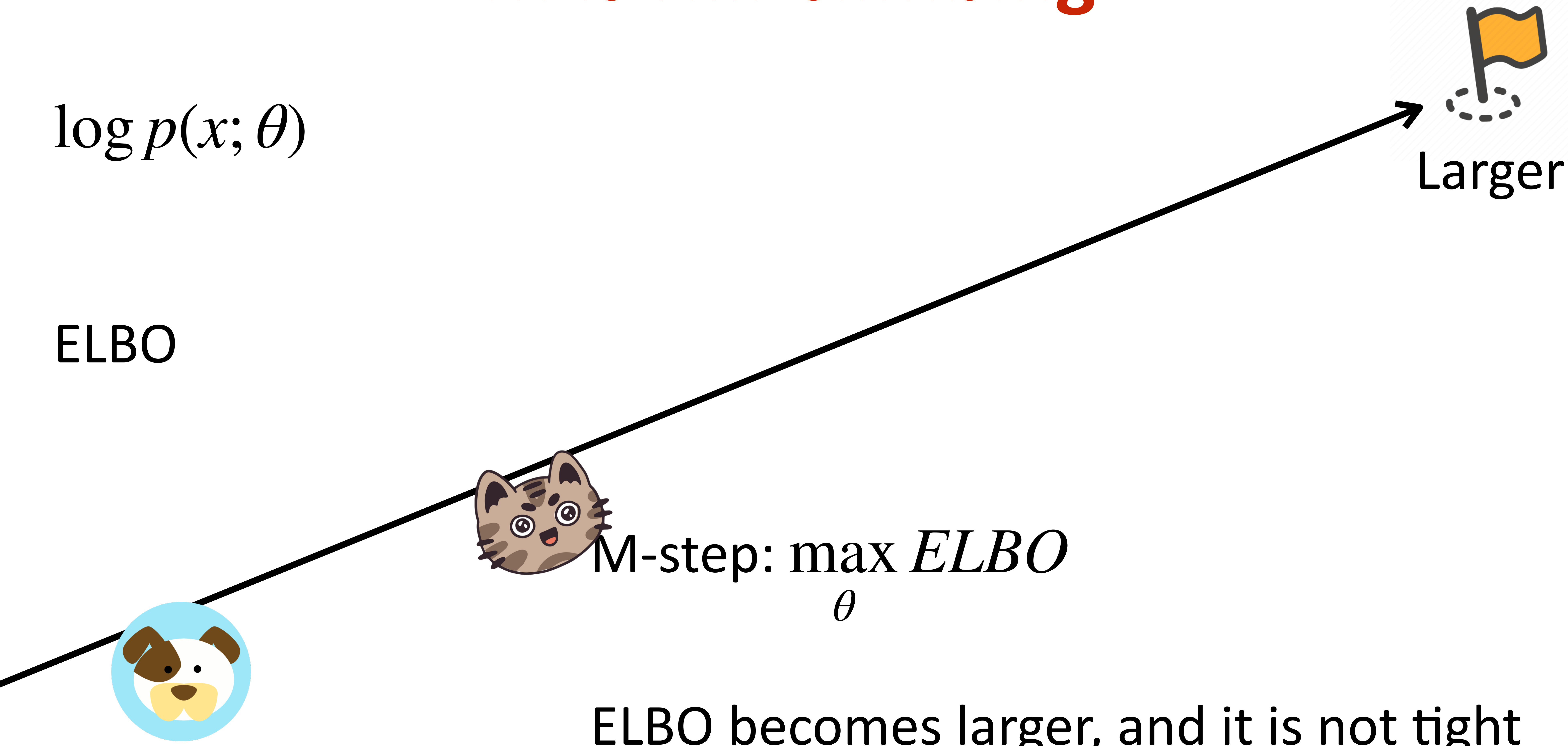
M-step: $\max_{\theta} ELBO$

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

EM is Hill Climbing

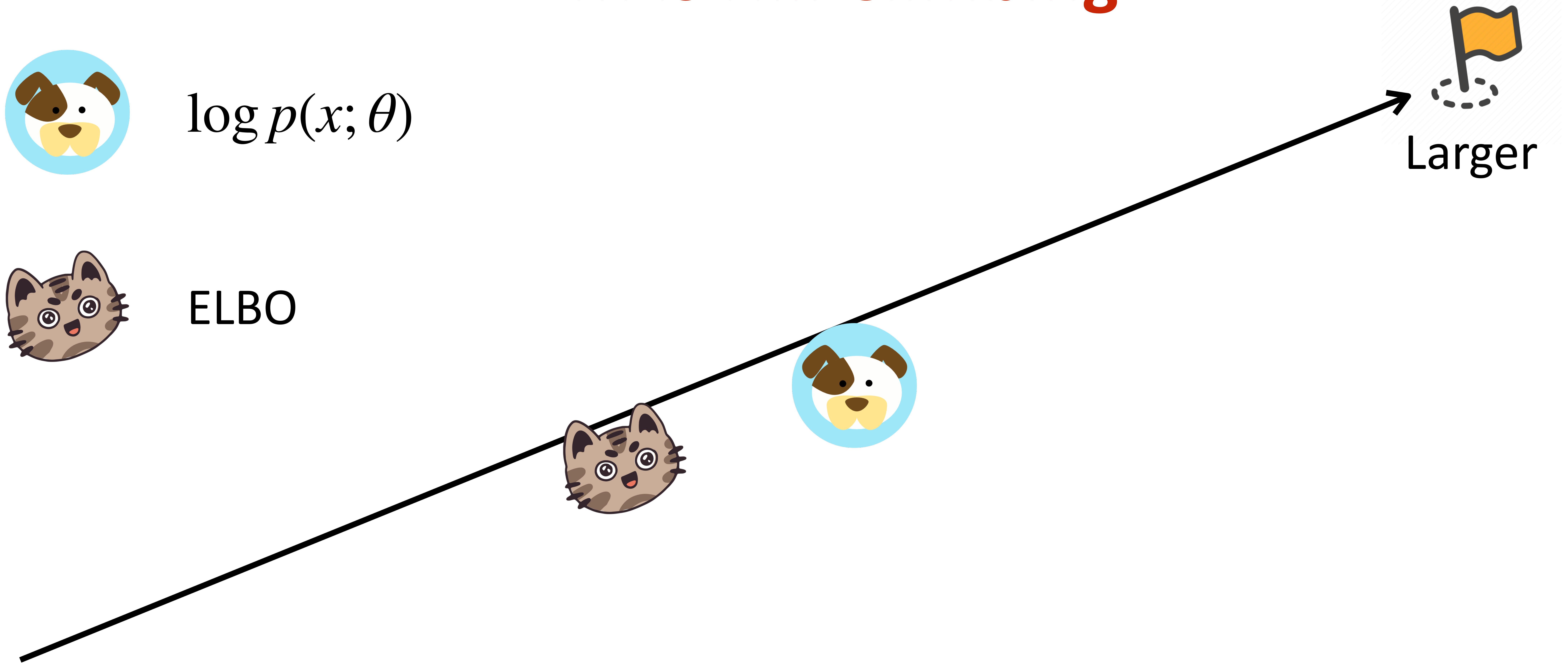
$\log p(x;\theta)$

ELBO

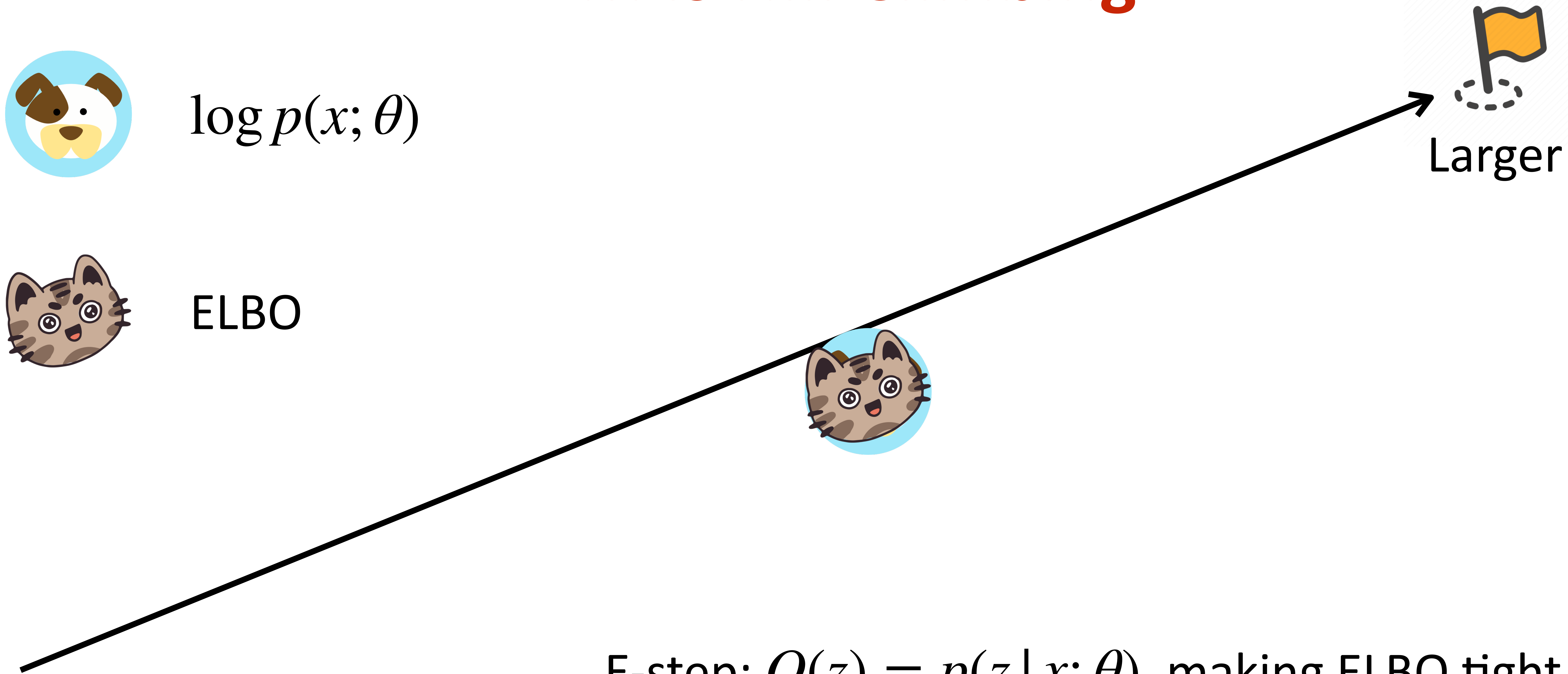
M-step: $\max_{\theta} ELBO$

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

EM is Hill Climbing



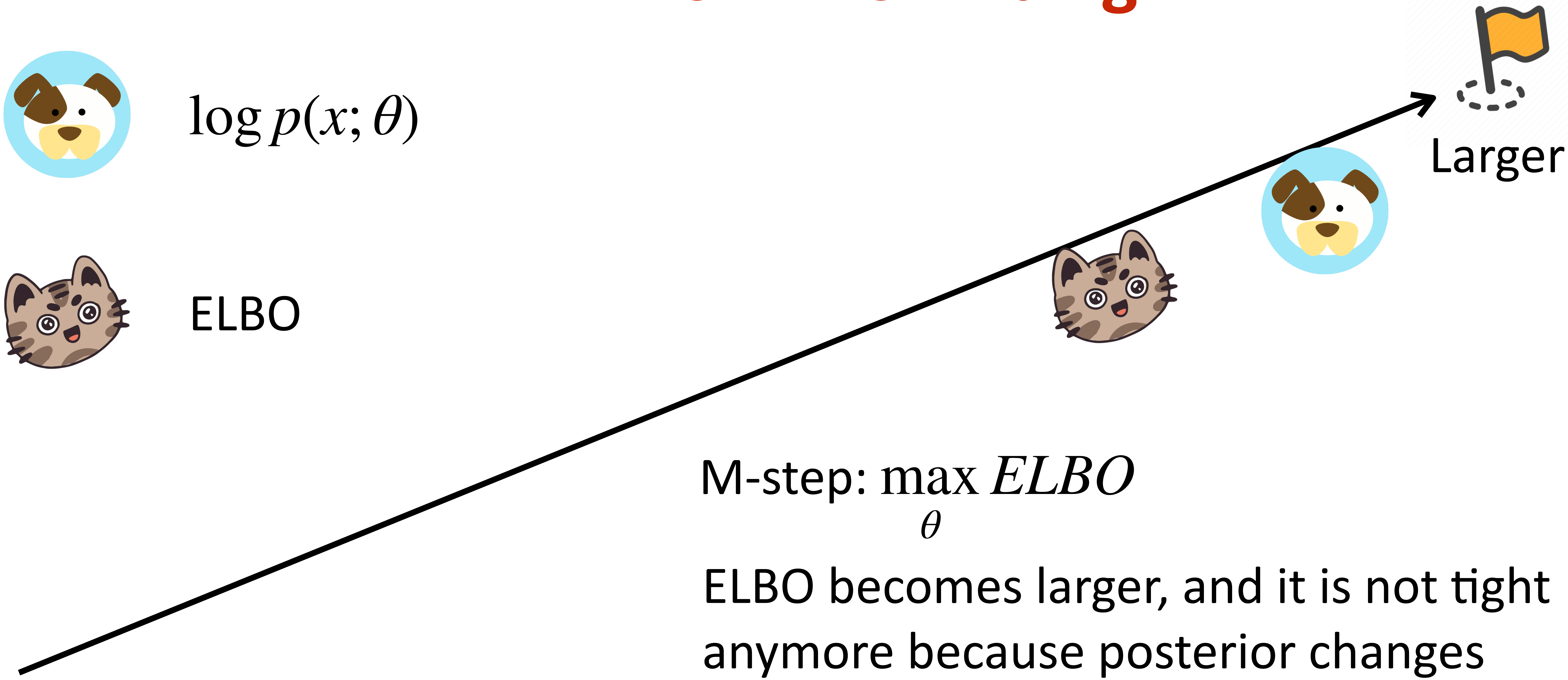
EM is Hill Climbing



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

M-step: $\max\limits_{\theta} ELBO$

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

Larger

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

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

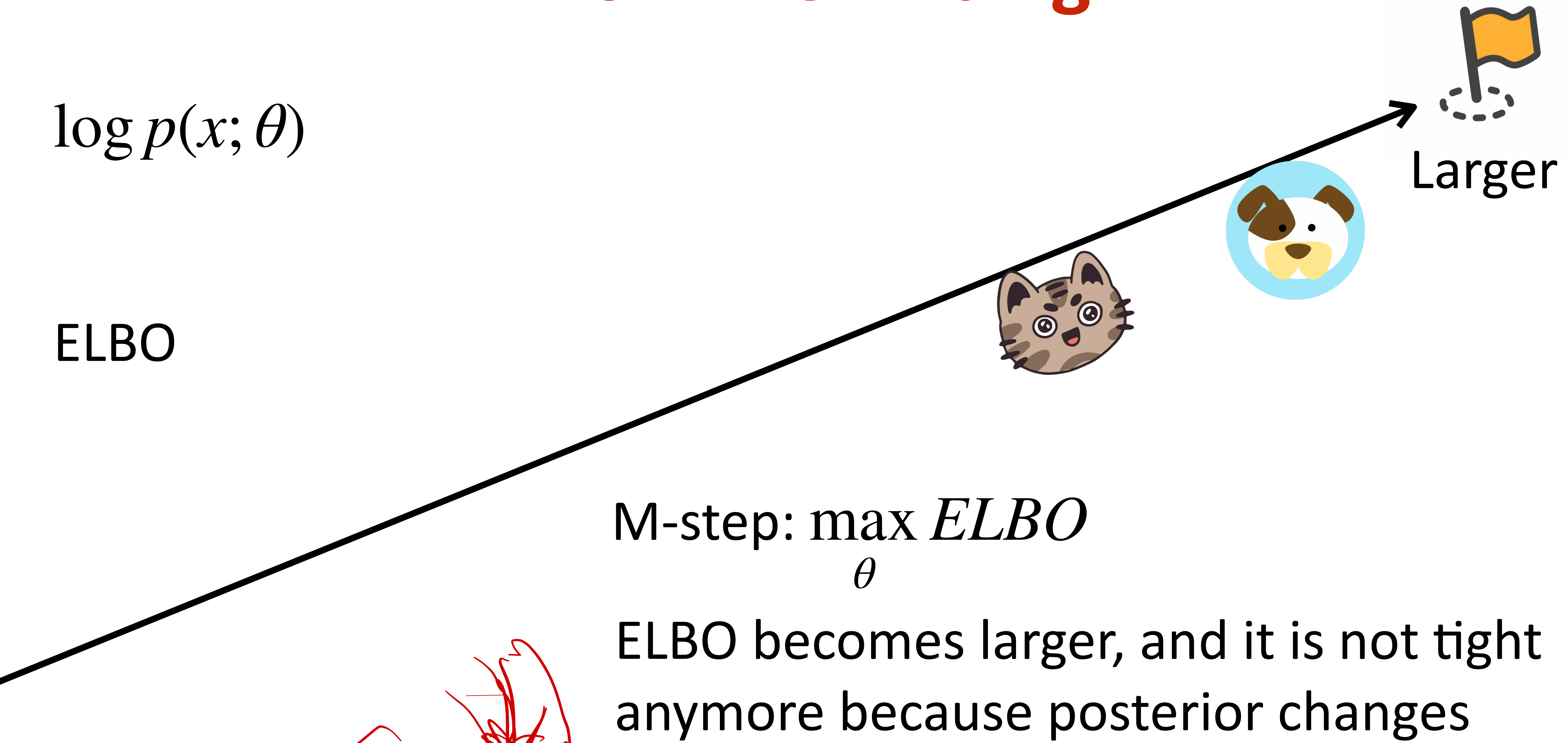
$\log p(x;\theta)$ is monotonically increasing!

We are doing MLE implicitly!

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

ELBO loss function

VAE

ELBO

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

E-step

maximize ELBO until convergene
$Q(z)$

m-step

maximize ELBO
θ

ELBO loss function

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

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

}

Revisit the E-Step

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

Repeat until convergence {

(E-step) For each i, set

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

(M-step) Set

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

}

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

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

}

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

$P(z)$

z

x

family

$P(z|x) \equiv P(z)$

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

$E_{z\sim Q(z)} \log \frac{P(x,z)}{Q(z)}$

$Q(z)$ easy to sample from

Revisit the M-Step

Revisit the M-Step

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

$\cancel{\theta}$

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

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

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

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:

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

$\times$

M-Step in EM:

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

$\checkmark$

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

x is given

Direct maximization:

$\leftarrow p(x|z)$

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

$z \sim p(z)$

$Q(z) = p(z|x)$

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?

$p(x,z)$

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

$$E_{z\sim q(z)} f(z) = \sum_i q(z_i) f(z_i)$$

$f(z_i)$ is large

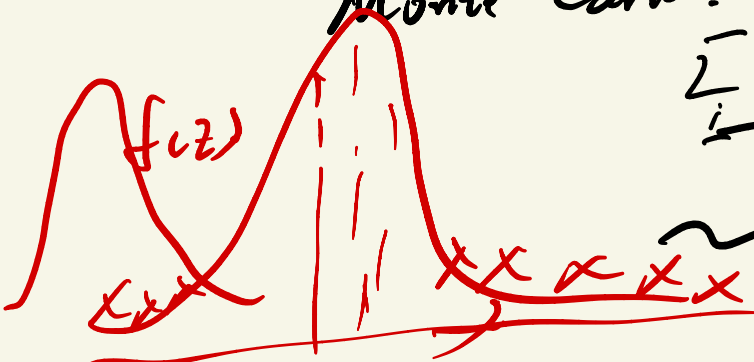
$q(z_i)$

$q(z)$

Monte Carlo:

$$\frac{\sum_i f(z^{(i)})}{N}$$

$z^{(i)} \sim q(z)$



$z^{(i)}$ where $f(z_i)$ and

$q(z)$ uniform 100 sample $q(z_i)$ are large

$f(z_i)$

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)

$p(z)$

reconstru

constant

AE

$p(z|x)$

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

$D_{KL} \geq 0$

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

$Q = p(z|x)$

Further Questions

Further Questions

E step $Q = p(z|x)$

- What if we do not have closed-form model posterior? $p(z|x)$

Further Questions

inference

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

$p(z|x)$

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

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