中山大学
SUN YAT-SEN UNIVERSITY

隐变量模型、EM算法和变分自编码器

主讲人： 林倞

无监督概率建模

- 在监督学习中，回归和分类都可以理解为是在学习条件概率分布

$$p(y|\mathbf{x};\mathbf{w})$$

> 在回归任务中，条件概率密度函数的形式通常假设为

$$p(y|\mathbf{x};\mathbf{w}) = \mathcal{N}(y;\mathbf{w}^T\mathbf{x},\ \sigma^2)$$

> 在分类任务中，条件概率密度函数的形式通常假设为

$$\text{二分类：}\quad p(y|\mathbf{x}) = (\sigma(\mathbf{x}\mathbf{w}))^y \cdot (1-\sigma(\mathbf{x}\mathbf{w}))^{1-y}$$

$$\text{多分类：}\quad p(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^{K}[softmax_k(\mathbf{W}\mathbf{x})]^{y_k}$$

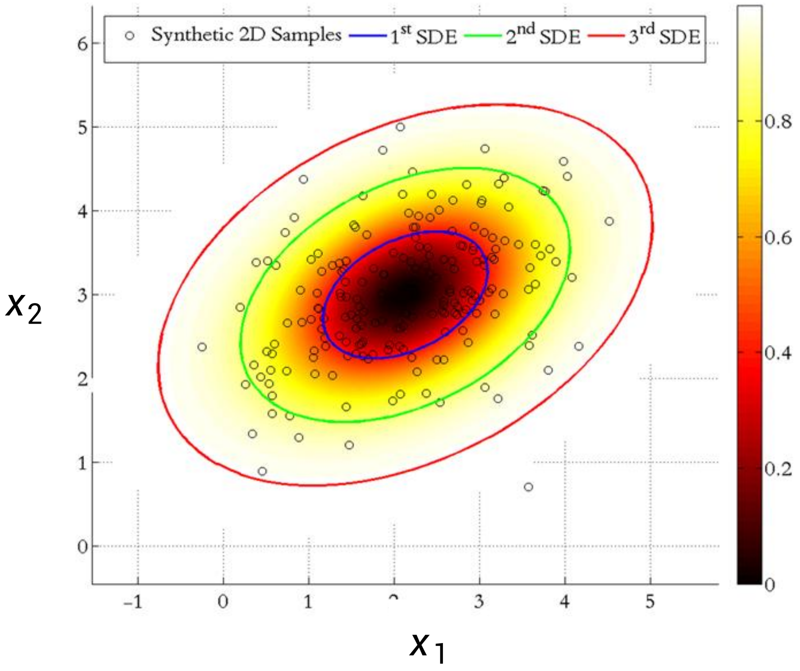
- 无监督学习也可以从学习概率分布的角度来理解。 但它只关注输入数据 $\boldsymbol{x}$ 的分布

$$p(\boldsymbol{x};\boldsymbol{w})$$

- 建模 $\boldsymbol{x}$ 要比标签 y困难得多。 一种朴素的方法是将 $p(\boldsymbol{x};\boldsymbol{w})$ 限制为高斯形式

$$p(\boldsymbol{x};\ \boldsymbol{w}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\ \boldsymbol{\Sigma})$$

  - 通过优化 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 来最优地描述数据点$\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$

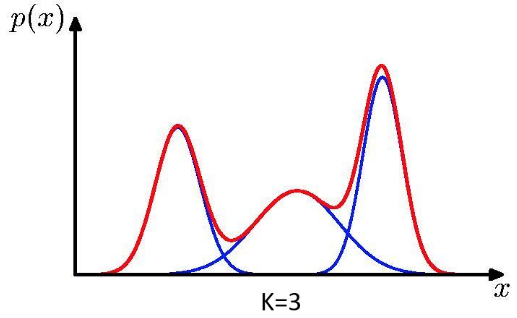


显然，该模型的表达能力非常有限

为什么需要隐变量？

- 理由1：通过组合简单模型构建更具表达能力的模型
  - 假设存在一个简单的类别分布 $p(z) = Cat(K, \boldsymbol{\pi})$ 和一个高斯分布 $p(x) = \mathcal{N}(x|\mu, \sigma^2)$
  - 如果单独使用它们，只能对简单的统计关系进行建模
  - 但如果我们将它们组合成 $p(x, z) = p(x|z)p(z)$ ，所得到的边缘分布 $p(x)$ 表达能力强得多

$$p(x) = \sum_{z} p(x|z)p(z) = \sum_{k=1}^{K} \pi_z \mathcal{N}(x|\mu_z, \sigma_z^2)$$



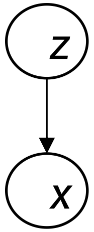
  - 理论上，它能够表示任何复杂的分布

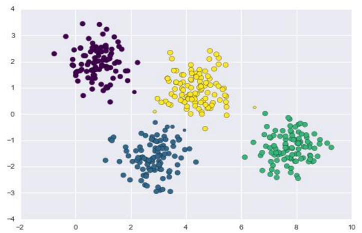
- 理由2：数据中隐藏的结构

1) 具有隐藏簇状结构的数据

$$z_n \sim Cat(K, \boldsymbol{\pi})$$

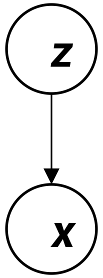
$$x_n \sim \mathcal{N}(x|\boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}_{z_n})$$

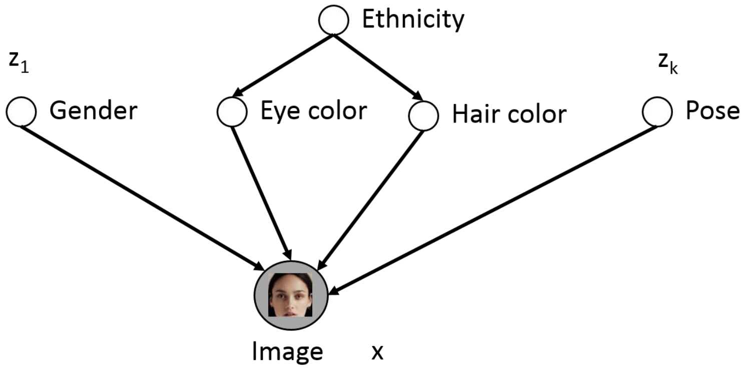




2) 面向文档的主题模型

3) 图像建模





➢ 在上面的例子中，隐变量 $\boldsymbol{z}$ 通常对应于高层级特征

➢ *如果考虑到这种隐藏结构，就可以获得更具可解释性的模型*

隐变量模型的一般形式

- 隐变量模型（LVMs）：含有隐变量的概率模型

$$p(\mathbf{x}, \mathbf{z})$$

  - $\mathbf{x}$ 是我们感兴趣的随机变量
  - $\mathbf{z}$ 是隐变量（Latent Variable）

➢ 有时可能存在多个隐变量 $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_K$

$$p(\mathbf{x}, \mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_K)$$

- 关于我们感兴趣的变量 $\mathbf{x}$ 的概率模型是

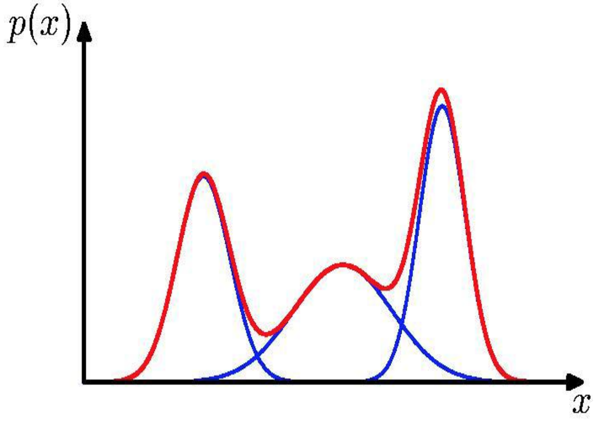
$$p(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) d\mathbf{z} \quad \text{或} \quad p(\mathbf{x}) = \int_{\mathbf{z}_1 \cdots \mathbf{z}_K} p(\mathbf{x}, \mathbf{z}_1, \cdots, \mathbf{z}_K) d\mathbf{z}_1 \cdots d\mathbf{z}_K$$

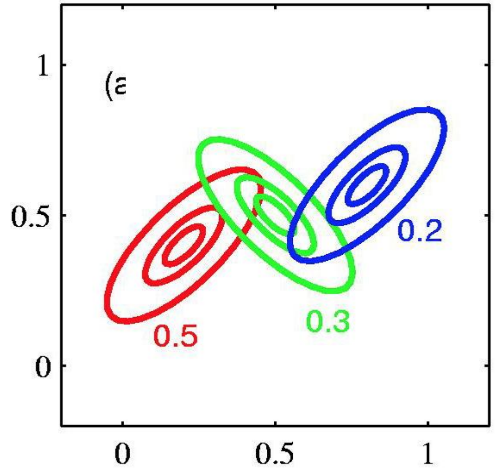
高斯混合分布

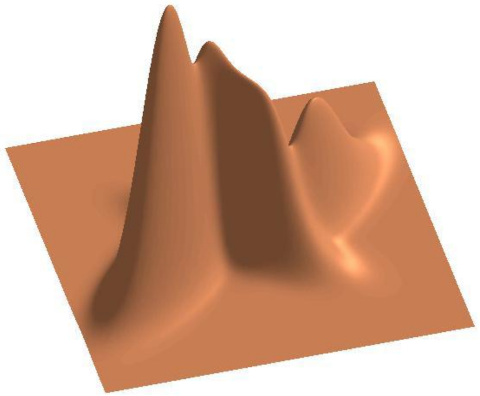
- 高斯混合分布的表达式

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

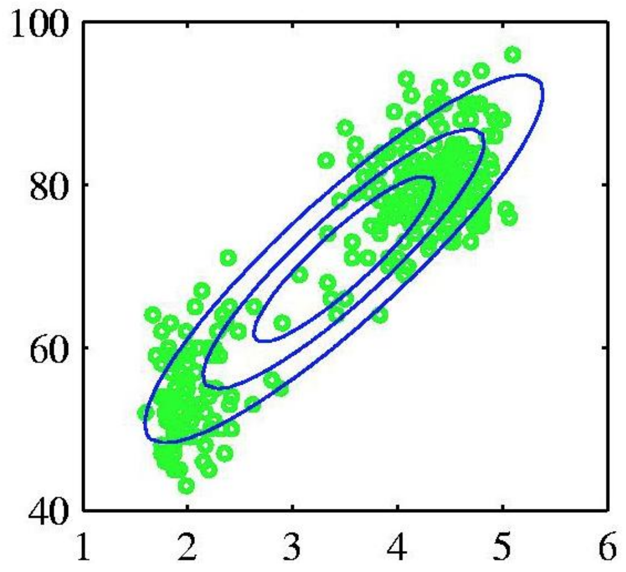
  - K表示高斯分布的数量
  - π_k 是第 k 个分布的权重，满足 $\sum_{k=1}^{K} \pi_k = 1$
  - $\boldsymbol{\mu}_k$ 和 $\boldsymbol{\Sigma}_k$ 分别是第 k 个高斯分布的均值向量和协方差矩阵



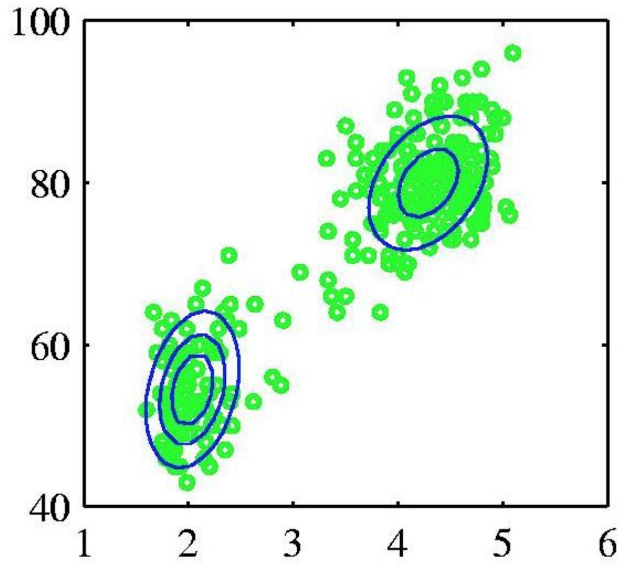




- 这些绿色的点很难用单个高斯分布建模



- 但如果用两个高斯分布的混合来建模，效果会好得多



将高斯混合分布表示为隐变量模型

- 对于一个隐变量模型 $p(\boldsymbol{x}, \boldsymbol{z})$， 如果我们将其条件分布 $p(\boldsymbol{x}|\boldsymbol{z})$ 和先验分布 $p(\boldsymbol{z})$ 设为

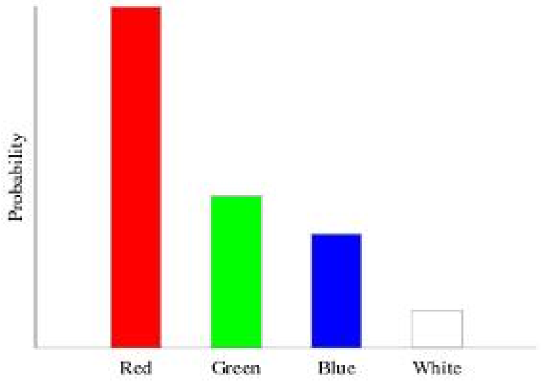
$$p(\boldsymbol{x}|\boldsymbol{z}=\boldsymbol{1}_k) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\boldsymbol{z}=\boldsymbol{1}_k) = \pi_k$$

  - $\boldsymbol{z}$ 只能是一个 one-hot 向量， 其中 $\boldsymbol{1}_k$ 表示第 k 个元素为1
  - 由于$p(\boldsymbol{z}=\boldsymbol{1}_k)=\pi_k$， $p(\boldsymbol{z})$ 实际上表示一个类别分布， 即

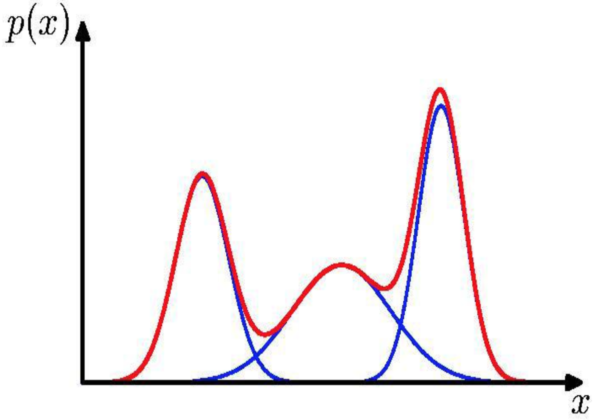
$$p(\boldsymbol{z}) = Cat(\boldsymbol{z}; \boldsymbol{\pi})$$

其中 $Cat(\boldsymbol{z}=\boldsymbol{1}_k; \boldsymbol{\pi}) = \pi_k$ 且 $\boldsymbol{\pi} = [\pi_1, \pi_2, \cdots, \pi_K]$



- 由于 $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$，我们可以很容易得到

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$



这正是高斯混合分布

- 因此，高斯混合分布也可以等价用隐变量模型来$p(\mathbf{x}, \mathbf{z})$表示，其中$p(\mathbf{x}|\mathbf{z} = \mathbf{1}_k) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$，$p(\mathbf{z} = \mathbf{1}_k) = \pi_k$

高斯混合分布的隐变量分布$p(\mathbf{x}, \mathbf{z})$可以写成如下形式

$$p(\mathbf{x}, \mathbf{z}) = \prod_{k=1}^{K} [\pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_k}$$

训练目标：最大化边缘概率分布

- 给定一组训练数据 $\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$， 目标是学习分布的参数

$$\{\pi_k,\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K} \triangleq \boldsymbol{\theta}$$

- 假设数据点 $\boldsymbol{x}^{(n)}$ 是独立同分布的， 因此我们可以将联合分布写为

未使用隐变量形式

$$p(\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}) = \prod_{n=1}^{N} \underbrace{\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{p(\boldsymbol{x}^{(n)})}$$

- 对于概率模型， 训练目标是*最大化对数似然函数*， 即

$$\boxed{\log p(\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

针对问题的一般形式

- 给定联合分布

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

其中 $\boldsymbol{x}$ 是观测变量，$\boldsymbol{z}$ 是隐变量，我们需要最大化关于 $\boldsymbol{\theta}$ 的对数似然，即

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x};\boldsymbol{\theta}),$$

其中

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

我们拥有的是 联合概率密度函数 $p(\boldsymbol{x},\ \boldsymbol{z};\boldsymbol{\theta})$，但需要优化的是边缘概率密度函数 $p(\boldsymbol{x};\boldsymbol{\theta})$

一般隐变量模型的训练

- 根据边缘分布与联合分布的关系，可知

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

训练目标就是最大化对数似然函数

$$\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_n)$$

  - 但是，由于积分的存在，很多时候，获得$p_{\boldsymbol{\theta}}(\boldsymbol{x}_n)$的表达式是很困难的

- 因此，对于一般隐变量模型，计算对数似然的梯度$\frac{d \log p_{\boldsymbol{\theta}}(\boldsymbol{x})}{d\boldsymbol{\theta}}$是很困难的，使用梯度下降法来训练模型就变得非常具有挑战性

对数似然的重表示

- 对数似然可以重写为

$\forall$ 分布 $q(\boldsymbol{z})$

$$\log p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log p(\boldsymbol{x})$$

$$= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})q(\boldsymbol{z})}$$

$$= \underbrace{\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z})}}_{\mathfrak{l}(q,\boldsymbol{\theta})} + \underbrace{\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})}}_{KL(q||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}))}$$

$$\boxed{= \mathfrak{l}(q,\boldsymbol{\theta}) + KL(q||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})), \quad \text{for } \forall\ \boldsymbol{\theta},\ q(\boldsymbol{z})}$$

注: KL 散度用于衡量两个分布 q 和 p 之间的距离，定义为

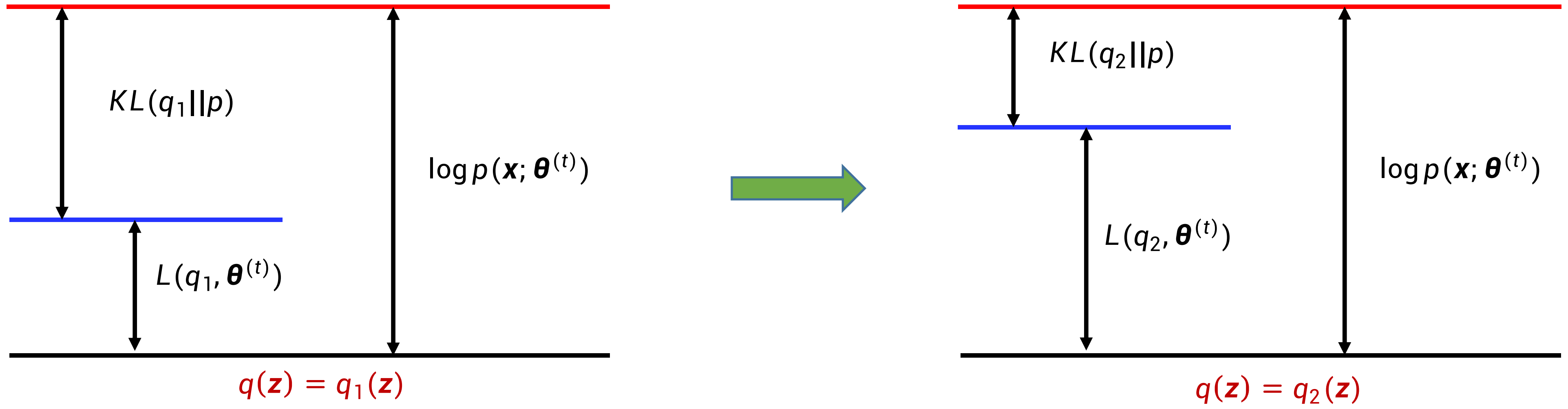
$$KL(q||p) \triangleq \int q(z) \log \frac{q(z)}{p(z)} dz \geq 0$$

- 因此，我们将第 t 次迭代的参数记为 $\boldsymbol{\theta}^{(t)}$，可得

$$\log p\left(\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right) = \mathbb{L}\left(q,\boldsymbol{\theta}^{(t)}\right) + KL\left(q||p\left(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)\right)$$

该等式对任何分布$q(\boldsymbol{z})$都成立

- 不同的 $q(\boldsymbol{z})$ 会导致 $\log p\left(\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)$ 的不同分解



EM 算法的理论依据

$$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z})\log\frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^{(t)})}{q(\boldsymbol{z})} + \sum_{\boldsymbol{z}} q(\boldsymbol{z})\log\frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}$$

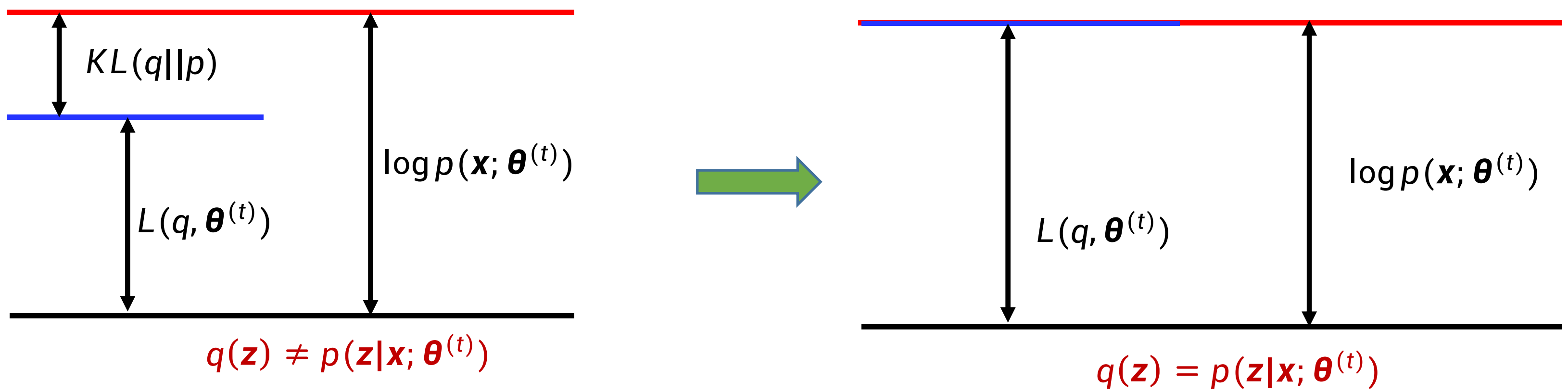
- 如果我们设 $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$，那么可以得到

$$KL(q||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})) = 0$$

因此，我们有

$$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)}) = \mathbb{L}(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta}^{(t)})$$

$$= \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})\log\frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^{(t)})}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}$$



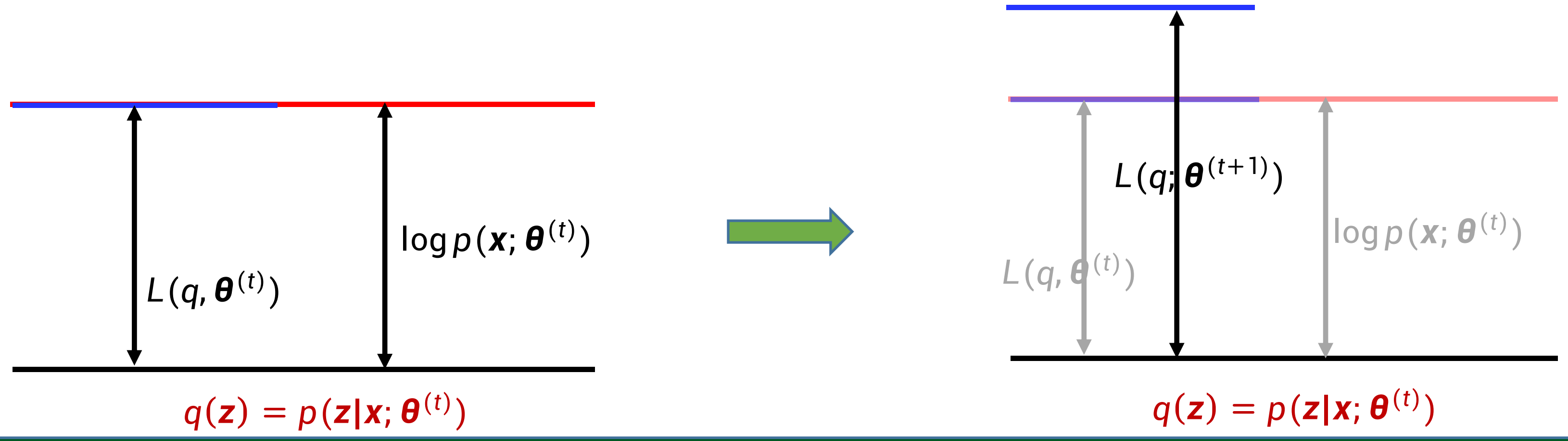
$$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)}) = \ell\left(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta}^{(t)}\right) = \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}) \log \frac{p(\boldsymbol{x},\ \boldsymbol{z};\boldsymbol{\theta}^{(t)})}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}$$

- 如果我们更新 $\boldsymbol{\theta}$ 为

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \ell\left(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta}\right),$$

那么我们必然得到如下关系

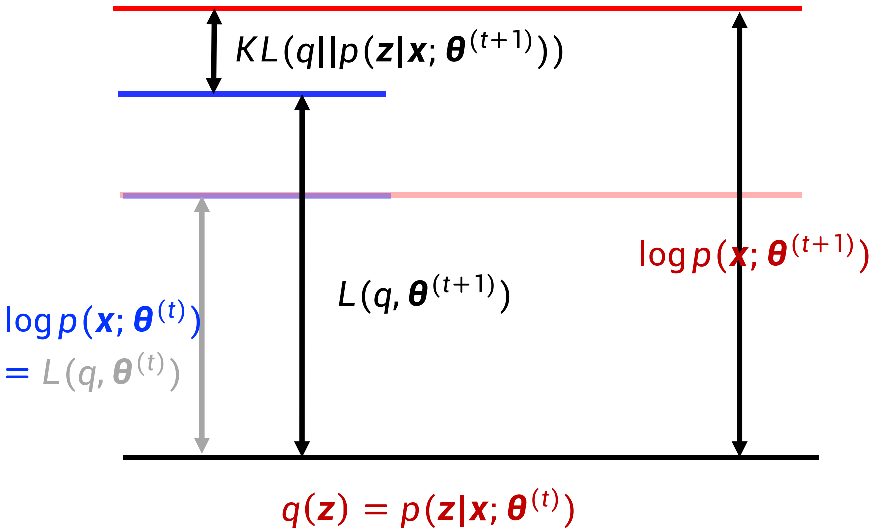
$$\ell\left(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta}^{(t+1)}\right) \geq \underbrace{\ell\left(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta}^{(t)}\right)}_{=\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})}$$



- 通过设置 $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(t)})$，我们得到

$$\log p(\mathbf{x};\boldsymbol{\theta}^{(t+1)}) = \underbrace{\mathcal{L}(p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta}^{(t+1)})}_{\geq \log p(\mathbf{x};\boldsymbol{\theta}^{(t)})} + \underbrace{KL(p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(t)})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(t+1)}))}_{\geq 0}$$

KL 散度始终为非负

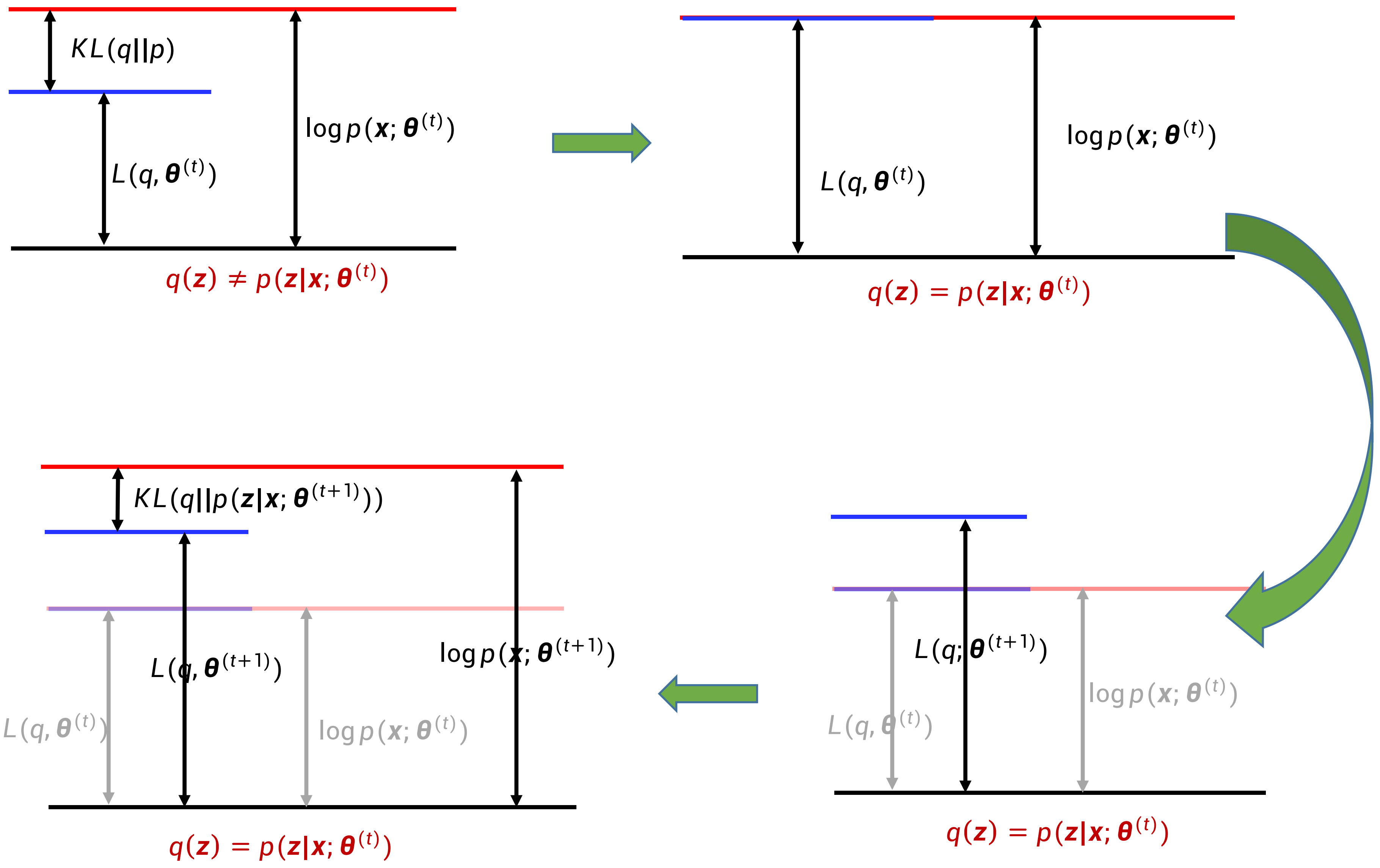


- 因此，我们可以看到

$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t+1)}) \geq \log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})$

KL 散度始终为非负

$\max_{\boldsymbol{\theta}} \mathcal{l}(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta})$ *能够保证每一步的似然函数值都会增加*



$KL(q||p)$
$L(q, \boldsymbol{\theta}^{(t)})$
$\log p(\mathbf{x}; \boldsymbol{\theta}^{(t)})$
$q(\mathbf{z}) \neq p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(t)})$
$L(q, \boldsymbol{\theta}^{(t)})$
$\log p(\mathbf{x}; \boldsymbol{\theta}^{(t)})$
$q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(t)})$
$L(q; \boldsymbol{\theta}^{(t+1)})$
$\log p(\mathbf{x}; \boldsymbol{\theta}^{(t)})$
$L(q, \boldsymbol{\theta}^{(t)})$
$q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(t)})$
$KL(q||p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(t+1)}))$
$L(q, \boldsymbol{\theta}^{(t+1)})$
$\log p(\mathbf{x}; \boldsymbol{\theta}^{(t+1)})$
$L(q, \boldsymbol{\theta}^{(t)})$
$\log p(\mathbf{x}; \boldsymbol{\theta}^{(t)})$
$q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(t)})$

EM 算法

- 算法

  E 步： 计算期望

$$\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x},\ \boldsymbol{z};\boldsymbol{\theta})]$$

  M 步： 更新参数　*EM 算法可以保证每一步的似然函数值都会增加*

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})$$

- EM 算法的关键要素

  1) 后验分布 $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(t)})$
  2) 联合分布对数 $\log p(\boldsymbol{x},\ \boldsymbol{z};\boldsymbol{\theta})$ 关于后验概率的期望
  3) 最大化

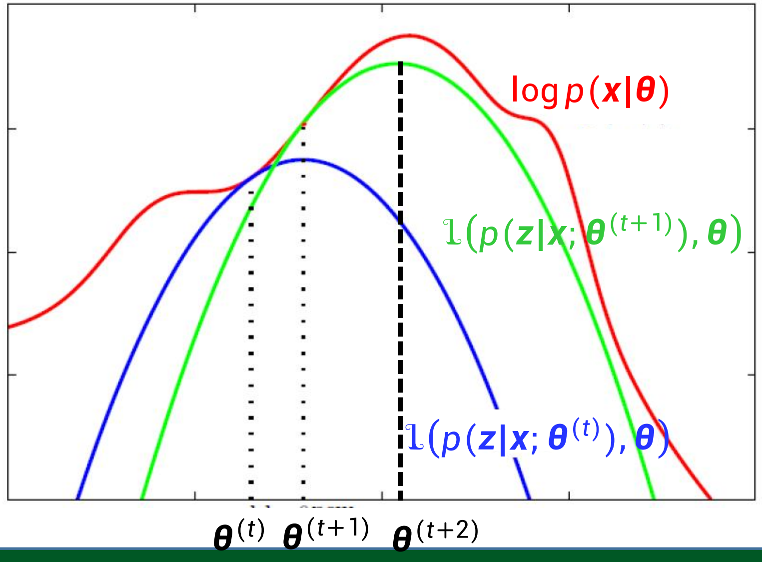
参数空间视角

1)E 步 (t)：给定模型参数 $\boldsymbol{\theta}^{(t)}$，推导出 $\mathbb{l}(p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta})$ 的表达式

2)M 步 (t)： 计算最优值 $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{l}(p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta})$

3)E 步 $(t+1)$：给定模型参数 $\boldsymbol{\theta}^{(t+1)}$， 推导出 $\mathbb{l}(p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(t+1)}),\boldsymbol{\theta})$ 的表达式

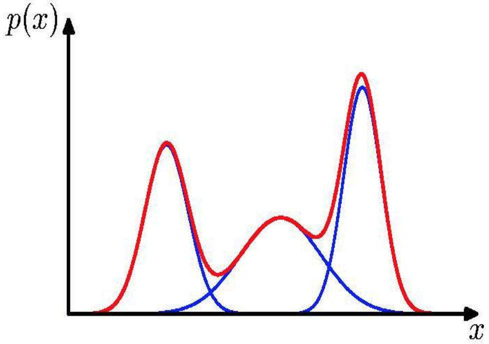
4)重复以上过程直至收敛



高斯混合模型回顾

- 对于一个高斯混合分布，即

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$



它可以表示为联合分布的边缘分布

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$$

$$= \prod_{k=1}^{K} [\pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_k}$$

  - $\boldsymbol{z} = [z_1, z_2, \cdots, z_K]$ 服从参数为 π 的类别分布

EM 算法的两个步骤

- 这是一个隐变量模型，因此我们可以使用 EM 算法来优化它

注： $\max_{\boldsymbol{\theta}} \mathbb{l}(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta})$ 等价于 $\max_{\boldsymbol{\theta}} \mathbb{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})$

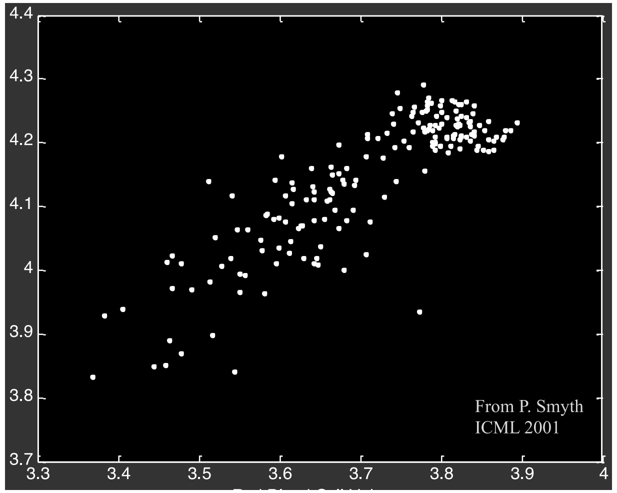
- 回顾： EM 算法的关键要素

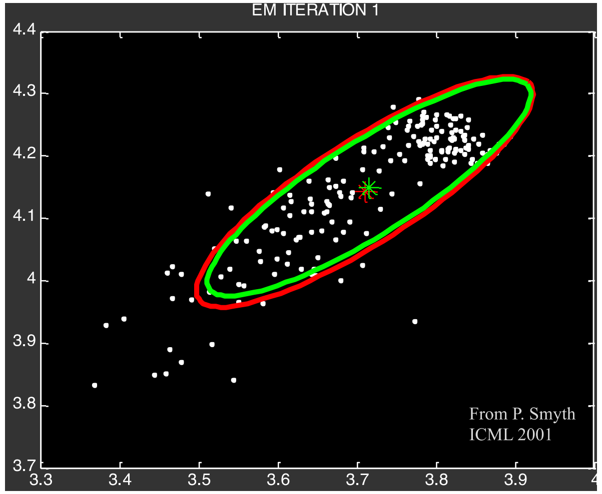
  - E 步：计算关于后验概率 $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$ 的期望

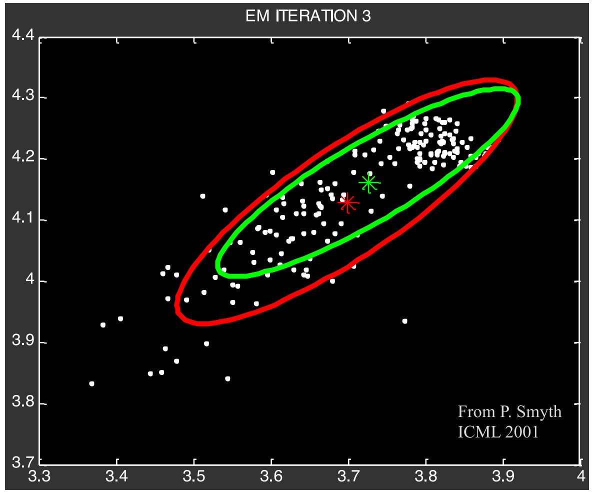
$$\mathbb{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{p(\boldsymbol{z}^{(n)}|\boldsymbol{x}^{(n)};\boldsymbol{\theta}^{(t)})}\left[\log p(\boldsymbol{x}^{(n)}, \boldsymbol{z}^{(n)};\boldsymbol{\theta})\right]$$

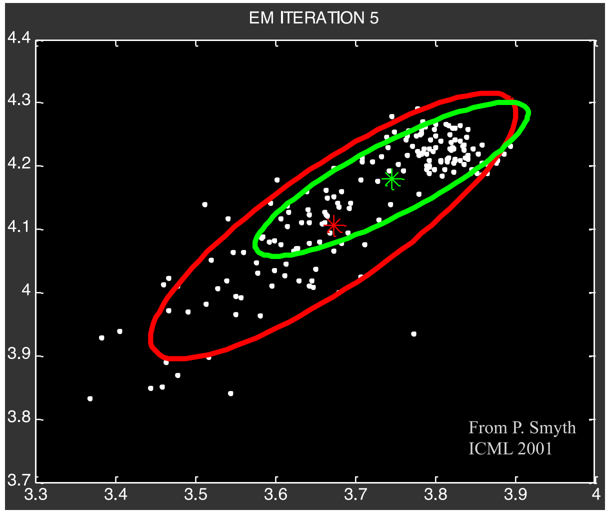
  - M 步：最大化

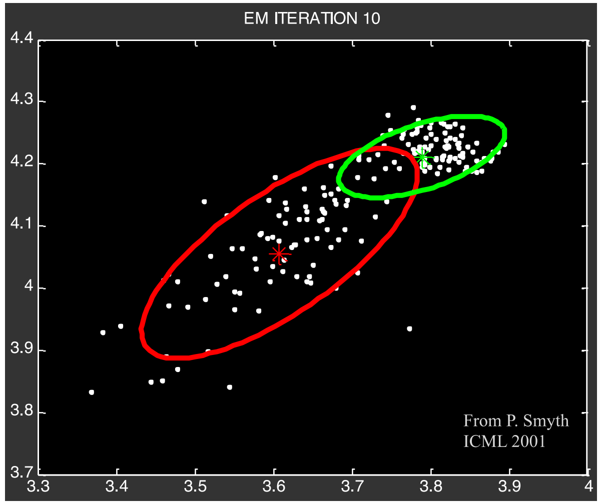
$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})$$

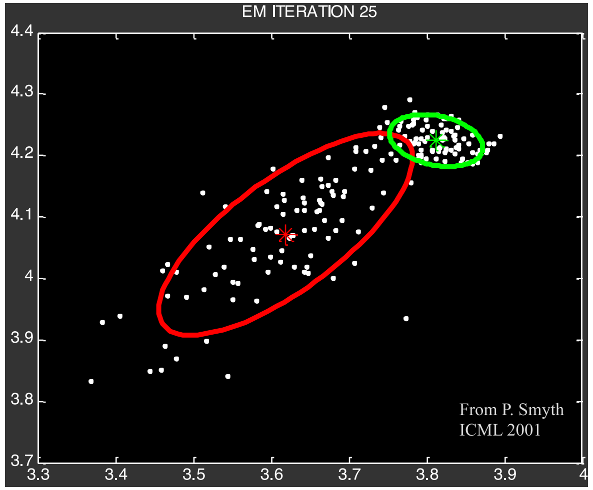

From P. Smyth
ICML 2001


EM ITERATION 1
From P. Smyth
ICML 2001


EM ITERATION 3
From P. Smyth
ICML 2001


EM ITERATION 5
From P. Smyth
ICML 2001


EM ITERATION 10
From P. Smyth
ICML 2001

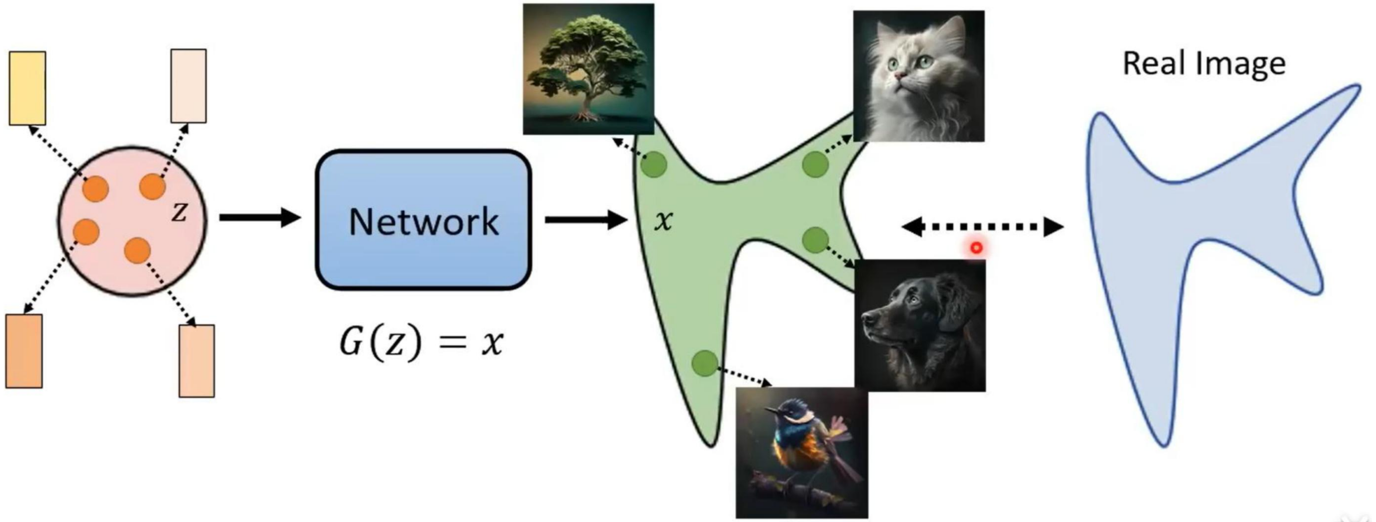

EM ITERATION 25
From P. Smyth
ICML 2001

从 EM 算法到变分自编码器

- 变分自编码器（Variational AutoEncoder，VAE）的直接数学目标就是：

  输入一个向量z，经过模型得到$G(z) = x$，希望这个得到的x的分布$p_\theta(x)$

  与实际已有数据的分布$p_{data}(x)$尽可能接近



- 优化目标：

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} KL(p_\theta(x), p_{data}(x)) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{x\sim p(\boldsymbol{data})}[p_\theta(x)]$$

从 EM 算法到变分自编码器

- 优化目标:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{x\sim p(\boldsymbol{data})}[p_\theta(x)]$$

- EM算法最核心的地方在于，为隐变量z引入了一个分布$q(z)$：

$$L(\theta) = logp_\theta(x) = \int_z q(z)logp_\theta(x|\theta)dz$$

$$= \int_z q(z)\log(\frac{p(x,z|\theta)}{p(z|x,\theta)}\cdot\frac{q(z)}{q(z)})dz$$

$$= \int_z q(z)\log\frac{p(x|z,\theta)\cdot p(z)}{q(z)}dz$$

$$+ \int_z q(z)\log\frac{q(z)}{p(z|x,\theta)}dz$$

$$= \underbrace{\mathbb{E}_{z\sim\boldsymbol{q}(\boldsymbol{z})}[\boldsymbol{log}p_\theta(x|z)] - KL(q(z)\|p(z))}_{\boldsymbol{ELBO}} + \underbrace{KL(q(z)\|p_\theta(z|x))}$$

从 EM 算法到变分自编码器

- 优化目标：

$$L(\theta) = \underbrace{\mathbb{E}_{z\sim \boldsymbol{q}(\boldsymbol{z})}[\boldsymbol{log}\, p_\theta(x|z)] - KL(q(z)\|p(z))}_{\boldsymbol{ELBO}} + \underbrace{KL(q(z)\|p_\theta(z|x))}$$

- EM算法的基本思路：先让KL散度取0，这样只剩一个ELBO项，最后让ELBO项最大。不断迭代直到收敛

E 步: 令$q^{(t+1)}(z) = p_\theta(z|x)$，计算期望：

$$\mathbb{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) = \mathbb{E}_{z\sim p_\theta(z|x)}[\boldsymbol{log}\, p_\theta(x|z)]$$

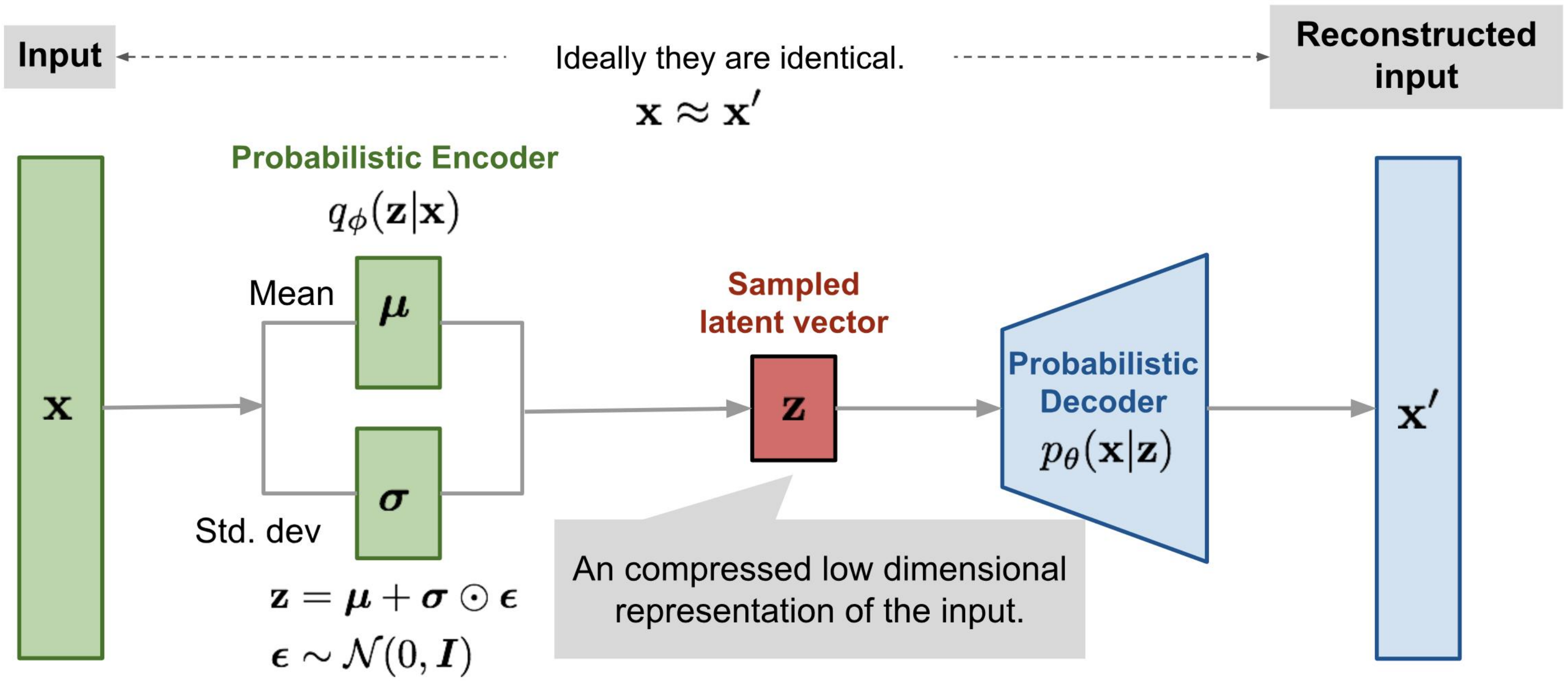
M步: 更新参数：

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$$

变分自编码器

$$L(\theta) = \underbrace{\mathbb{E}_{z\sim \boldsymbol{q}(\boldsymbol{z})}[\boldsymbol{log} p_\theta(x|z)] - KL(q(z)\|p(z))}_{ELBO} + \underbrace{KL(q(z)\|p_\theta(z|x))}$$

- 第一个改进 ： 合并E-step和M-step，同时优化$q(z),\theta$来达到一步到位$max_{q,\theta}ELBO$
- 第二个改进 ： 对于$q(z)$这个很难预测的分布，用深度学习模型$q_\phi(z|x)$来进行控制



变分自编码器

- 变分自编码器的优化目标：

$$L(\theta) = \underbrace{\mathbb{E}_{z\sim q_\Phi(z|x)}[\log p_\theta(x|z)]}_{\text{重建项}} + \underbrace{KL(q_\Phi(z|x)\|p(z))}_{\text{正则项}}$$