EM算法

VCG

数学工具和记号

期望

- 连续型随机变量 x，概率密度 $p(x)$。函数 $f(x)$ 的期望
  - $\mathbb{E}_{p(x)}[f(x)] = \int f(x)p(x)\mathrm{d}x$
    - 清晰地表明它是关于 $p(x)$ 的期望值
    - 关于概率分布 $q(x)$ 的期望值

$$\mathbb{E}_{q(x)}[f(x)] = \int f(x)q(x)\mathrm{d}x$$

- 参数 θ 的概率分布 $p(x;\theta)$，也可以写成 $p_\theta(x)$

之前		之后
$\mathbb{E}[f(x)]$	$\longrightarrow$	$\mathbb{E}_{p(x)}[f(x)]$
$p(x;\theta)$	$\longrightarrow$	$p_\theta(x)$

KL 散度

- KL散度$D_{\mathrm{KL}}(p \parallel q)$：用于衡量概率分布 $p(x)$ 和 $q(x)$之间差异
  - x 为连续型随机变量

$$D_{\mathrm{KL}}(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$$

    - $\int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x = \int p(x)(\log(-q(x)) - \log(-p(x)))\mathrm{d}x$【相当于：平均信息差】
  - x 为离散型随机变量

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

- KL 散度特性
  - 两个概率分布的差异越大，KL 散度的值就越大。
  - KL 散度的值大于或等于 0，且仅当两个概率分布相同时，其值才为 0
  - KL 散度非对称，$D_{\mathrm{KL}}(p \parallel q)$ 和 $D_{\mathrm{KL}}(q \parallel p)$ 的值不同

最大似然估计 (Maximum Likelihood Estimation)

最大似然估计

- 概率分布 p，由参数 θ决定：$p(x;\theta)$；例如，正态分布参数 $\theta=\{\mu,\sigma\}$
- 样本 $\mathcal{D}=\left\{x^{(1)},x^{(2)},\cdots,x^{(N)}\right\}$，数据基于概率分布 $p(x;\theta)$ 独立生成
  - 当参数为 θ 时，获得样本 $\mathcal{D}$ 的概率密度

$$L(\theta)=p(\mathcal{D};\theta)=p\left(x^{(1)};\theta\right)p\left(x^{(2)};\theta\right)\cdots p\left(x^{(N)};\theta\right)=\prod_{n=1}^{N}p\left(x^{(n)};\theta\right)$$

  - $L(\theta)$ 称为似然 (likelihood) 或似然函数 (likelihood function)
    - 以参数 θ 为参数的函数,表示在给定参数 θ 的情况下，样本 $\mathcal{D}$ 出现的概率密度
- 最大似然估计：找到使似然 $p(\mathcal{D};\theta)$ 最大的参数 θ

$$\hat{\theta}=\underset{\theta}{\operatorname{argmax}}\prod_{n=1}^{N}p\left(x^{(n)};\theta\right)$$

    - $\hat{\theta}$ 为观测到样本的概率最大，模型最拟合样本
  - 常用对数似然 $\log p(\mathcal{D};\theta)$ 的最大化 $\hat{\theta}=\underset{\theta}{\operatorname{argmax}}\sum_{n=1}^{N}\log p_\theta\left(x^{(n)}\right)$

似然函数优化与ELBO

优化目标：似然函数

- 似然函数的基础项

$$\log p_\theta = \log \sum_{z} p_\theta(x, z)$$

- 偏导

$$\frac{\partial \log p_\theta}{\partial \theta} = \frac{1}{\sum_z p_\theta(x, z)} \sum_{z} \frac{\partial}{\partial \theta} p_\theta(x, z)$$

- log-sum形式，优化存在问题
  - 各个参数项紧密耦合，$\frac{1}{\sum_z p_\theta(x,z)}$为包含了所有的 $p_\theta(x, z)$项共同的分母项
    - 意味着参数 θ 对$p_\theta(x, z)$的微小改变，其对总体梯度的贡献，会受到其他项$p_\theta(x, z)$的影响

优化形式：sum-log与log-sum

sum-log与log-sum

- 优化通常通过参数（θ）偏导并令其为零来实现，两种形式
- $L(\theta) = \sum_i \log f_i(\theta)$【sum-log】
  - $\frac{\partial L}{\partial \theta} = \sum_i \frac{\partial}{\partial \theta} \log f_i(\theta) = \sum_i \frac{1}{f_i(\theta)} \frac{\partial f_i(\theta)}{\partial \theta}$
  - 每个 $f_i(\theta)$ 的导数独立计算，然后简单相加。可以分别处理每一项
- $L(\theta) = \log\left(\sum_i f_i(\theta)\right)$【log-sum】
  - $\frac{\partial L}{\partial \theta} = \frac{1}{\sum_j f_j(\theta)} \cdot \left(\sum_i \frac{\partial f_i(\theta)}{\partial \theta}\right)$
  - 包含了所有的 $f_i(\theta)$ 项共同的分母项
    - 意味着参数 θ 对$f_i(\theta)$的微小改变，其对总体梯度的贡献，会受到其他项$f_j(\theta)$的影响

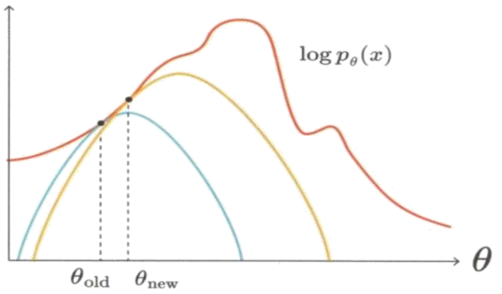
sum-of-log与log-of-sum

- 实际上，似然函数中还有其它项，即
  - $\frac{\partial L}{\partial \theta} = \sum_i \frac{1}{f_i(\theta)} \frac{\partial f_i(\theta)}{\partial \theta}$+(其它项)【sum-log】
  - $\frac{\partial L}{\partial \theta} = \frac{1}{\sum_j f_j(\theta)} \cdot \left(\sum_i \frac{\partial f_i(\theta)}{\partial \theta}\right)$+(其它项)【log-sum】
- log-of-sum的问题
  - 优化时，令$\frac{\partial L}{\partial \theta} = 0$，对于 log-of-sum 结构，得到一个复杂方程，其中所有参数都通过分母紧密地耦合在一起
    - 通常，这个log-sum方程没有解析解（closed-form solution）

似然函数的推导

EM算法的变分推断视角

- 【优化目标】对数似然函数$\log p_\theta$ 。其中，θ是要学习的模型参数
- 【思想】引入隐变量**z**及其变分分布$q(\mathbf{z})$，将对 θ 的优化转为对 (θ, q) 的联合优化
- 【策略】最大化对数似然转为最大化它的下界(ELBO)。通过交替优化ELBO，单调提升$\log p_\theta$的值，对应着EM算法的E步和M步
  - E步(变分步 / Expectation Step)
    - 固定当前模型参数 θ，将 $q(\mathrm{z})$ 调整到最优。即令 $q(\mathrm{z}) = p(\mathrm{z}|\mathbf{x}, \theta)$
      - 此时， ELBO被提升至紧贴真实的对数似然函数值，得到当前参数 θ 的最优下界
      - 等价于构建了完整数据的对数似然期望（Q函数）
    - 【当前模型参数下的，隐变量后验分布$p(\mathrm{z}|\mathbf{x}, \theta)$的最优估计；故称变分步】
  - M步 (最大化步 / Maximization Step)
    - 固定上一步得到的最优分布 $q(z)$，调整模型参数 θ ，最大化当前的ELBO
      - 得到新的模型参数
      - 由于下界被抬高，真实对数似然也随之被保证单调提升
    - 【当前隐函数分布下的，对模型参数的最优估计】



变分分布 $q(z)$的分析

【优化项】$\log p_\theta(x) = \log\frac{p_\theta(x,z)}{q(z)} + \log\frac{q(z)}{p_\theta(z|x)}$

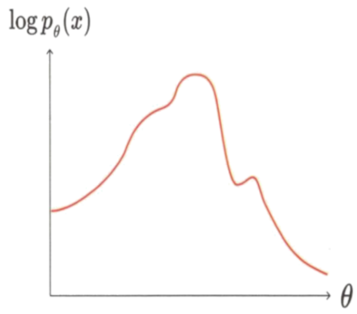
【分析】用辅助变分分布 $q(z)$对后验分布$p_\theta(z \mid x)$（ $p_\theta(z \mid x)$给定观测数据x，猜测其潜在变量z分布）进行最优估计。因此，需要拼凑出 $\frac{q(z)}{p_\theta(z|x)}$ 项！

$\log p_\theta(x) = \log\frac{p_\theta(x,z)}{p_\theta(z|x)}$【引入数据到隐变量的条件后验概率 $p_\theta(z \mid x)$ 】

$= \log\frac{p_\theta(x,z)}{p_\theta(z|x)}\frac{q(z)}{q(z)}\left(\text{乘以 } \frac{q(z)}{q(z)} = 1\right)$【引入任意的隐变量分布$q(z)$ 】

$= \log\frac{p_\theta(x,z)}{q(z)} + \log\frac{q(z)}{p_\theta(z|x)}$【引入隐变量分布和后验分布的差异】

于是， $\log p_\theta(x) = \log\frac{p_\theta(x,z)}{q(z)} + \log\frac{q(z)}{p_\theta(z|x)}$



ELBO

【优化项】$\log p_\theta(x) = \log\frac{p_\theta(x,z)}{q(z)} + \log\frac{q(z)}{p_\theta(z|x)}$

$\sum_z q\,(z) = 1$，则$\mathbb{E}_{z\sim q(z)}1 = 1$，而$\log p_\theta(x) = \log\frac{p_\theta(x,z)}{q(z)} + \log\frac{q(z)}{p_\theta(z|x)}$与$z$无关，于是

$\log p(x) = \mathbb{E}_{z\sim q(z)}[\log p(x)]$【期望项与$z$无关，当常数看】

$= \mathbb{E}_{z\sim q(z)}\left[\log\frac{p_\theta(x,z)}{q(z)} + \log\frac{q(z)}{p_\theta(z|x)}\right]$【拆分期望项】

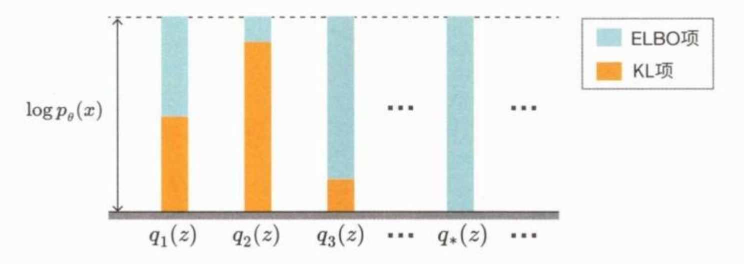
$= \mathbb{E}_{z\sim q(z)}\left[\log\frac{p_\theta(x,z)}{q(z)}\right] + \mathbb{E}_{z\sim q(z)}\left[\log\frac{q(z)}{p_\theta(z|x)}\right]$【期望的线性性质】

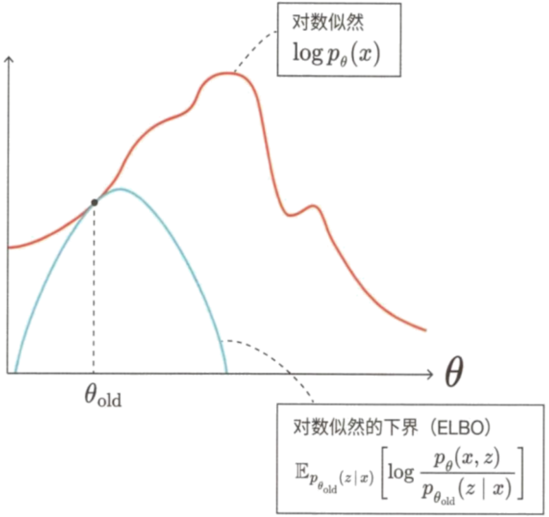
$= \mathrm{ELBO}(x;q,\theta) + D_{\mathrm{KL}}\big(q(z)\parallel p_\theta(z\mid x)\big)$

其中， $\mathrm{ELBO}(x;q,\theta) \stackrel{\mathrm{def}}{=} \mathbb{E}_{z\sim q(z)}\left[\log\frac{p_\theta(x,z)}{q(z)}\right]$【sum-log形式，可以解析！！】

似然函数优化 – E步

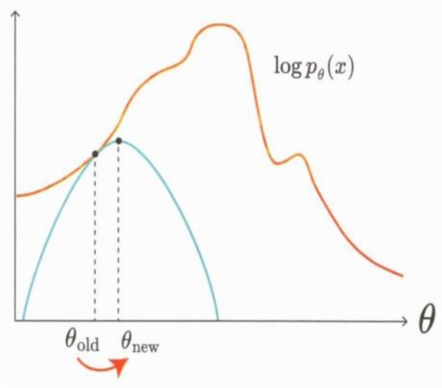
- 【优化项】$\log p(x) = \mathrm{ELBO}(x; q, \theta) + D_{\mathrm{KL}}\big(q(z) \parallel p_\theta(z \mid x)\big)$
- E步：固定模型参数$\theta = \theta_{old}$，优化$q(z)$
  - $\log p(x)$与z无关，不能更大。因此，优化目标转为原目标函数的下界ELBO，让D_{KL}最小即可
  - 于是，$D_{\mathrm{KL}} = 0$；从而，$q(z) = p_{\theta_{old}}(z \mid x)$
    - 用后验概率 $p_\theta(z \mid x)$ 作为对隐变量分布的最佳猜测 $q(z)$
    - 即，现有模型给出了缺失信息的一个最合理猜测
  - ELBO更新
    - $\mathrm{ELBO}\big(x; q = p_{\theta_{old}}, \theta\big) = \mathbb{E}_{p_{\theta_{old}}(z|x)}\left[\log \frac{p_\theta(x,z)}{p_{\theta_{old}}(z|x)}\right]$

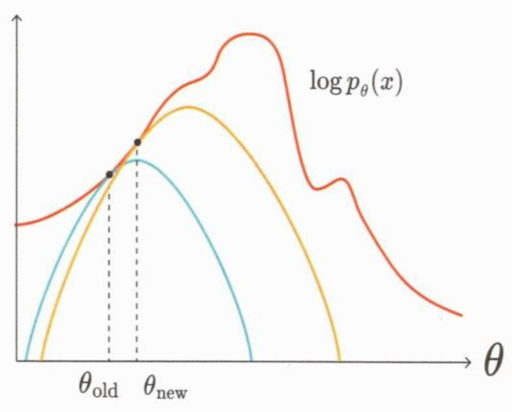




似然函数优化 – M步

- 【优化项】$\log p(x) = \mathrm{ELBO}(x; q, \theta) + D_{\mathrm{KL}}\big(q(z) \parallel p_\theta(z \mid x)\big)$
- M步：固定隐变量分布$q(z) = q_{old}(z)$，优化θ
  - D_{KL}与θ无关，因此，最大化$\mathrm{ELBO}(x; q, \theta)$即可
  - 通过ELBO更新参数 θ
    - 实际上此步骤为标准的最大似然，于是，新模型更适应当前的隐变量分布$q_{old}(z)$
- 新一轮E步和M步





单调有界

- ELBO单调上升，且有上界，故EM算法收敛