第十章 变分推断（Variational Inference）

演讲者：本公子

本节目标：

1. **对变分推断有个简单了解**
2. **了解共轭分布**
3. **如何使用变分推断**
4. **了解指数族分布**
5. **将变分推断用于指数族分布**

1. 对变分推断有个简单了解

现有实验数据X和参数Z：

$$Z=\{z_1,z_2,z_3\},X=\{x_1,x_2,x_3\}$$

联合分布为：

$$P(X,Z)=P(x_1,x_2,x_3,z_1,z_2,z_3)$$

根据贝叶斯定理，我们可以得到后验概率为：

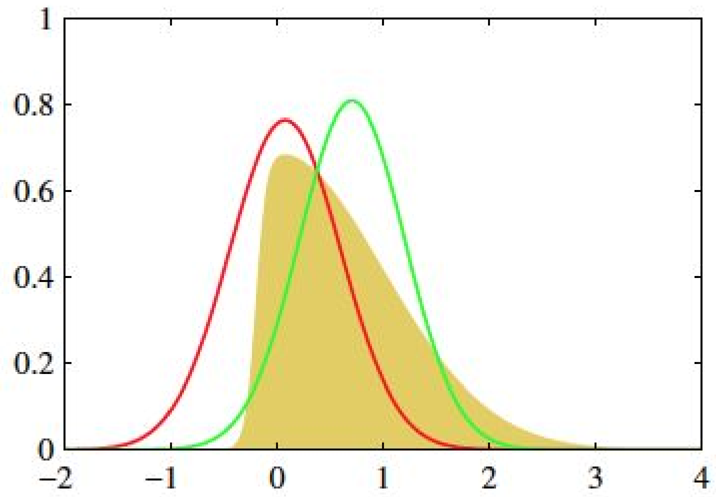
$$P(Z|X)=P(z_1,z_2,z_3|x_1,x_2,x_3)=\frac{P(x,z)}{P(x)}$$
$$p(Z|X)=\frac{P(x_1,x_2,x_3,z_1,z_2,z_3)}{\int_{z_1}\int_{z_2}\int_{z_3}P(x_1,x_2,x_3,z_1,z_2,z_3)dz_1dz_2dz_3}$$

问题来了：往往对分母积分是非常困难的，这时候就是我们**变分推断**上场了！

变分推断做什么？

> 任务：已知概率分布 $P(x_1,x_2,x_3,z_1,z_2,z_3)$，目标是求$P(z_1,z_2,z_3|x_1,x_2,x_3)$，**即希望根据已有的数据推断后验分布**。但是由于积分问题，直接求出后验分布是很困难的。因此我们希望**找出一个概率分布**$q(z_1,z_2,z_3)$**去近似后验概率分布**$P(z_1,z_2,z_3|x_1,x_2,x_3)$。

举个例子：



黄色的分布是我们的原始目标p，不好求。它看上去有点像高斯，那我们尝试从高斯分布中找一个红q和一个绿q，选更像p的q作为p的近似分布。

怎么做呢？

$$P(X)=\frac{P(X,Z)}{P(Z|X)}$$

$$\begin{aligned}lnP(X)&=lnP(X,Z)-lnP(Z|X)\\&=[lnP(X,Z)-lnq(Z)]-[lnP(Z|X)-lnq(Z)]\\&=ln\frac{P(X,Z)}{q(Z)}-ln\frac{P(Z|X)}{q(Z)}\end{aligned}$$

对左右两边求关于概率分布$q(Z)$的期望，得到：

$$\begin{aligned}E_{q(Z)}lnP(X)&=E_{q(Z)}lnP(X,Z)-E_{q(Z)}lnP(Z|X)\\\int lnP(X)q(Z)dZ&=\int q(Z)ln\frac{P(X,Z)}{q(Z)}dZ-\int\frac{P(Z|X)}{q(Z)}dZ\\lnP(X)&=\int q(Z)lnP(X,Z)dZ-\int q(Z)lnq(Z)dZ+(-\int q(Z)ln\frac{p(Z|X)}{q(Z)}dZ)\\lnP(X)&=\int q(Z)lnP(X,Z)dZ-\int q(Z)lnq(Z)dZ+KL(q(Z)||P(Z|X))\end{aligned}$$

我们发现了什么没有？

我们发现：

1. **我们的目标是找到一个$q(Z)$去近似$P(Z|X)$，在上式中即表现为$KL(q(Z)||P(Z|X))$，那么我们肯定希望KL散度越小越好（即q十分接近p）。**

2. **但是如果我们要直接优化KL散度也是不可行的，因为KL散度中包含了我们无法处理的$P(Z|X)$。（积分很困难嘛）**
3. **仔细观察上式，发现$P(X)$是一个恒定的常数，那么如果我们要最小化KL散度，即相当于最大化式子$\int q(Z)lnP(X,Z)dZ-\int q(Z)lnq(Z)dZ$。**

我们定义：

将$\int q(Z)lnP(X,Z)dZ-\int q(Z)lnq(Z)dZ$称为**ELOB**（Evidence Lower Bound），它会随着我们选择的$q(Z)$的变化而变化，因而ELOB是一个关于函数$q(Z)$的函数，即泛函。因此上述公式转变为：

$$\begin{aligned}lnP(X)&=ELOB+KL(q(Z)||P(Z|X))\\&=L(q)+KL(q(Z)||P(Z|X))\end{aligned}$$

所以现在我们的目标转变成了挑选一个q，使得ELOB最大化！！！

$$\begin{aligned}L(q)&=\sum_Z q(Z)ln\frac{P(X,Z)}{q(Z)}\\&=\sum_{z_1}\sum_{z_2}\sum_{z_3}q(z_1,z_2,z_3)ln\frac{P(x_1,x_2,x_3,z_1,z_2,z_3)}{q(z_1,z_2,z_3)}\end{aligned}$$

即，目标是选择$q(z_1,z_2,z_3)$来最大化$L(q)$，但实际操作**非常困难**！！！那么我们应该怎么办？？？

不妨对q函数做一个假设，假设$q(z_1,z_2,z_3)=q(z_1)q(z_2)q(z_3)$，即这个概率分布的各个变量之间都是独立的！

现在我们将q函数带入$L(q)$函数可得：

$$\begin{aligned}L(q)&=\sum_{z_1}\sum_{z_2}\sum_{z_3}q(z_1,z_2,z_3)ln\frac{P(x_1,x_2,x_3,z_1,z_2,z_3)}{q(z_1,z_2,z_3)}\\&=\sum_{z_1}\sum_{z_2}\sum_{z_3}q(z_1)q(z_2)q(z_3)ln\frac{P(X,Z)}{q(z_1)q(z_2)q(z_3)}\\&=\sum_{z_1}\sum_{z_2}\sum_{z_3}q(z_1)q(z_2)q(z_3)[lnP(X,Z)-lnq(z_1)q(z_2)q(z_3)]\\&=\sum_{z_1}\sum_{z_2}\sum_{z_3}q(z_1)q(z_2)q(z_3)[lnP(X,Z)-lnq(z_1)-lnq(z_2)-lnq(z_3)]\end{aligned}$$

让我们把$L(q)$拆成三部分！

$$\begin{aligned}
&\sum_{z_1}\sum_{z_2}\sum_{z_3} q(z_1)q(z_2)q(z_3)lnP(X,Z) \\
&-\sum_{z_1}\sum_{z_2}\sum_{z_3} q(z_1)q(z_2)q(z_3)lnq(z_1) \\
&-\sum_{z_1}\sum_{z_2}\sum_{z_3} q(z_1)q(z_2)q(z_3)[lnq(z_2)+lnq(z_3)]
\end{aligned}$$

我们的目标是寻找一个函数$q(z_1,z_2,z_3)$使得ELOB最大化，但是现在$q(z_1,z_2,z_3)$可被拆分成单独的三个小函数$q(z_1),q(z_2),q(z_3)$，所以我们的问题可以变成分别寻找到三个最优的小函数$q^*(z_1),q^*(z_2),q^*(z_3)$，使得ELOB最大化！

那么现在就让我们控制住$q(z_2),q(z_3)$这两个小函数，把它们两个当成常数，把函数$q(z_1)$函数当成变量，来求最优化问题！

看一下拆出来的第三个式子：

$$\begin{aligned}
-\sum_{z_1}\sum_{z_2}\sum_{z_3} q(z_1)q(z_2)q(z_3)[lnq(z_2)+lnq(z_3)] &= -\sum_{z_1} q(z_1)\sum_{z_2}\sum_{z_3} q(z_2)q(z_3)[lnq(z_2)+lnq(z_3)] \\
&= -\sum_{z_1} q(z_1)k
\end{aligned}$$

看一下拆出来的第二个式子：

$$\begin{aligned}
-\sum_{z_1}\sum_{z_2}\sum_{z_3} q(z_1)q(z_2)q(z_3)lnq(z_1) &= -\sum_{z_1} q(z_1)lnq(z_1)\sum_{z_2}\sum_{z_3} q(z_2)q(z_3) \\
&= -\sum_{z_1} q(z_1)lnq(z_1)
\end{aligned}$$

看一下拆出来的第一个式子：

$$\begin{aligned}
\sum_{z_1}\sum_{z_2}\sum_{z_3} q(z_1)q(z_2)q(z_3)lnP(x_1,x_2,x_3,z_1,z_2,z_3) &= \sum_{z_1} q(z_1)\sum_{z_2}\sum_{z_3} q(z_2)q(z_3)lnP(x_1,x_2,x_3,z_1,z_2,z_3) \\
&= \sum_{z_1} q(z_1)E_{q(z_2),q(z_3)}lnP(x_1,x_2,x_3,z_1,z_2,z_3)
\end{aligned}$$

注：$E[f(x)]=\sum f(x)p(x)$

我们将上述三个式子重新合并起来：

$$
\begin{aligned}
L &= \sum_{z_1} q(z_1) E_{q(z_2),q(z_3)} lnP(x_1,x_2,x_3,z_1,z_2,z_3) - \sum_{z_1} q(z_1) lnq(z_1) - \sum_{z_1} q(z_1) k \\
&= \sum_{z_1} q(z_1) [E_{q(z_2),q(z_3)} lnP(x_1,x_2,x_3,z_1,z_2,z_3) - k] - \sum_{z_1} q(z_1) lnq(z_1) \\
&= \sum_{z_1} q(z_1) [E_{q(z_2),q(z_3)} lnP(x_1,x_2,x_3,z_1,z_2,z_3) - k_1 - k_2] - \sum_{z_1} q(z_1) lnq(z_1)
\end{aligned}
$$

令

$$
\begin{aligned}
lnf(X,Z) &= E_{q(z_2),q(z_3)} lnP(x_1,x_2,x_3,z_1,z_2,z_3) - k_1 \\
f(X,Z) &= e^{-k_1} e^{E_{q(z_2),q(z_3)} lnP(x_1,x_2,x_3,z_1,z_2,z_3)} \\
&= ce^{E_{q(z_2),q(z_3)} lnP(x_1,x_2,x_3,z_1,z_2,z_3)}
\end{aligned}
$$

当k_1是一个合适的值时，c将成为归一化因子，使得$f(X,Z)$称为概率分布！

$$
\begin{aligned}
L &= \sum_{z_1} q(z_1)(lnf(X,Z) - k_2) - \sum_{z_1} q(z_1) lnq(z_1) \\
&= \sum_{z_1} q(z_1) lnf(X,Z) - \sum_{z_1} q(z_1) lnq(z_1) - k_2 \sum_{z_1} q(z_1) \\
&= \sum_{z_1} q(z_1) lnf(X,Z) - \sum_{z_1} q(z_1) lnq(z_1) - k_2 \\
&= \sum_{z_1} q(z_1) ln\frac{f(X,Z)}{q(z_1)} + const
\end{aligned}
$$

此时，最大化L相当于最小化$KL(q(z_1)||f(X,Z))$，即$q(z_1) = f(X,Z)$!!!!!

最终结论：

$$
\begin{aligned}
q(z_1) &= c_1 e^{\sum_{z_2}\sum_{z_3} lnP(x_1,x_2,x_3,z_1,z_2,z_3)} \\
q(z_2) &= c_1 e^{\sum_{z_1}\sum_{z_3} lnP(x_1,x_2,x_3,z_1,z_2,z_3)} \\
q(z_3) &= c_1 e^{\sum_{z_1}\sum_{z_2} lnP(x_1,x_2,x_3,z_1,z_2,z_3)}
\end{aligned}
$$

2. 了解共轭分布

定义：在贝叶斯理论中，如果后验概率分布$P(\theta|X)$与先验概率分布$P(\theta)$的概率分布是同一种形式，那么先验概率分布与后验概率分布就称为共轭分布，而先验概率分布就称为似然函数的共轭先验。

$$
P(\theta|X) \propto P(X|\theta)P(\theta)
$$

对应的似然与共轭先验

似然函数	共轭先验
Bernoulli	Beta
Binomial	Beta
Poisson	Gamma
Gaussian	Gaussian

举个例子：似然Possion，共轭先验Gamma

假设有一组观测样本$x_1, ..., x_n$独立同分布于泊松分布，即$x_i \sim Poisson(\lambda)$，则：

$$P(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

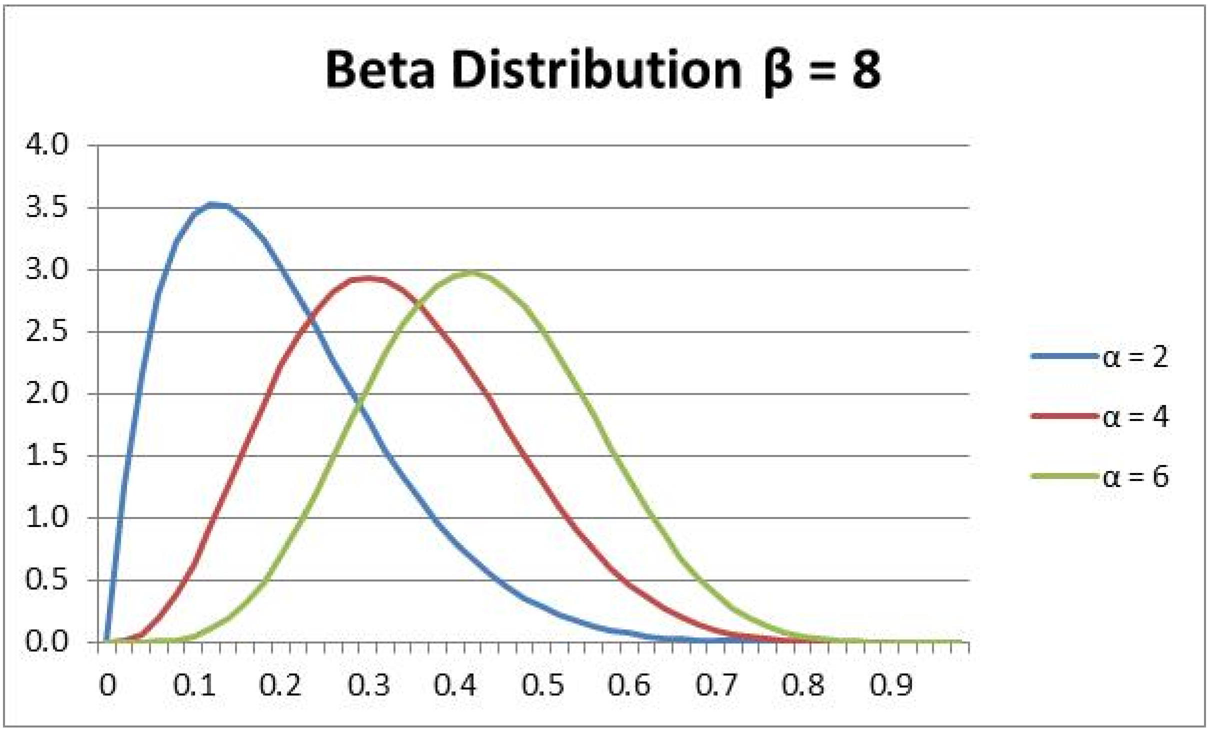
从而可以很轻松的写出相应的似然函数

$$\begin{aligned} L(x_1, ..., x_n|\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \end{aligned}$$

其中$\lambda > 0$是一个未知的参数。在贝叶斯的世界中，假设它服从Gamma分布，给定先验概率分布$\lambda \sim Gamma(\alpha, \beta)$，则

$$p(\lambda|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta\lambda} \lambda^{\alpha-1}$$

其中Gamma分布中的α表示形状参数，β表示比率参数。



根据贝叶斯公式，可以得到：

$$\begin{aligned} P(\lambda|x_1,\ldots,x_n,\alpha,\beta) &\propto P(\lambda|\alpha,\beta)L(x_1,\ldots,x_n|\lambda) \\ &\propto e^{-(\beta+n)\lambda}\lambda^{\alpha+\sum_{i=1}^n x_i-1} \end{aligned}$$

$$(\lambda|x_1,\ldots,x_n,\alpha,\beta) \sim Gamma(\alpha+\sum_{i=1}^n x_i, \beta+n)$$

所以，假设一组观测样本独立同分布于参数为λ的泊松分布，则Gamma分布是参数为λ的共轭先验分布。

3. 如何使用变分推断

已知现有数据为$X=(x_1,\ldots,x_n)$，则我们假设X服从**高斯分布**，高斯分布的两个参数的联合概率服从**高斯-伽马分布**，即：

$$
\begin{aligned}
P(X|\mu,\tau) &= \prod_{i=1}^{n}(\frac{\tau}{2\pi})^{\frac{1}{2}}exp(\frac{-\tau}{2}(x_i-\mu)^2) \\
&= (\frac{\tau}{2\pi})^{\frac{n}{2}}exp(\frac{-\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2) \\
P(\tau) \sim Gamma(\tau|\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\tau^{\alpha-1}e^{-\beta\tau} \\
P(\mu|\tau) \sim N(\mu_0,(\lambda_0\tau)^{-1}) &= (\frac{\lambda_0\tau}{2\pi})^{\frac{1}{2}}exp(-\frac{\lambda_0\tau(\mu-\mu_0)^2}{2}) \\
P(\mu,\tau) = P(\mu|\tau)P(\tau) \sim NormalGamma(\mu_0,\lambda_0,\alpha,\beta) &= \frac{\beta^\alpha\sqrt{\lambda_0}}{\Gamma(\alpha)\sqrt{2\pi}}\tau^{\alpha-\frac{1}{2}}e^{-\beta\tau}exp(-\frac{\lambda_0\tau(\mu-\mu_0)^2}{2})
\end{aligned}
$$

当(μ,τ)服从高斯-伽马分布且X服从高斯分布时，后验概率分布$(\mu,\tau|X)$将同样服从高斯-伽马分布（即共轭分布）。

$$
\begin{aligned}
P(\mu,\tau|X) \propto P(X|\mu,\tau)P(\mu,\tau) \sim NormalGamma(\mu_n,\lambda_n,\alpha_n,\beta_n) \\
\mu_n &= \frac{\lambda_0\mu_0+n\bar{x}}{\lambda_0+n} \\
\lambda_n &= \lambda_0+n \\
\alpha_n &= \alpha_0+\frac{n}{2} \\
\beta_n &= \beta_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{\lambda_0 n(\bar{x}-\mu_0)^2}{2(\lambda_0+n)}
\end{aligned}
$$

虽然共轭分布的后验概率分布十分好求，但现在就假设我们不知道这后验概率分布，让我们使用变分推断来进行推导吧!

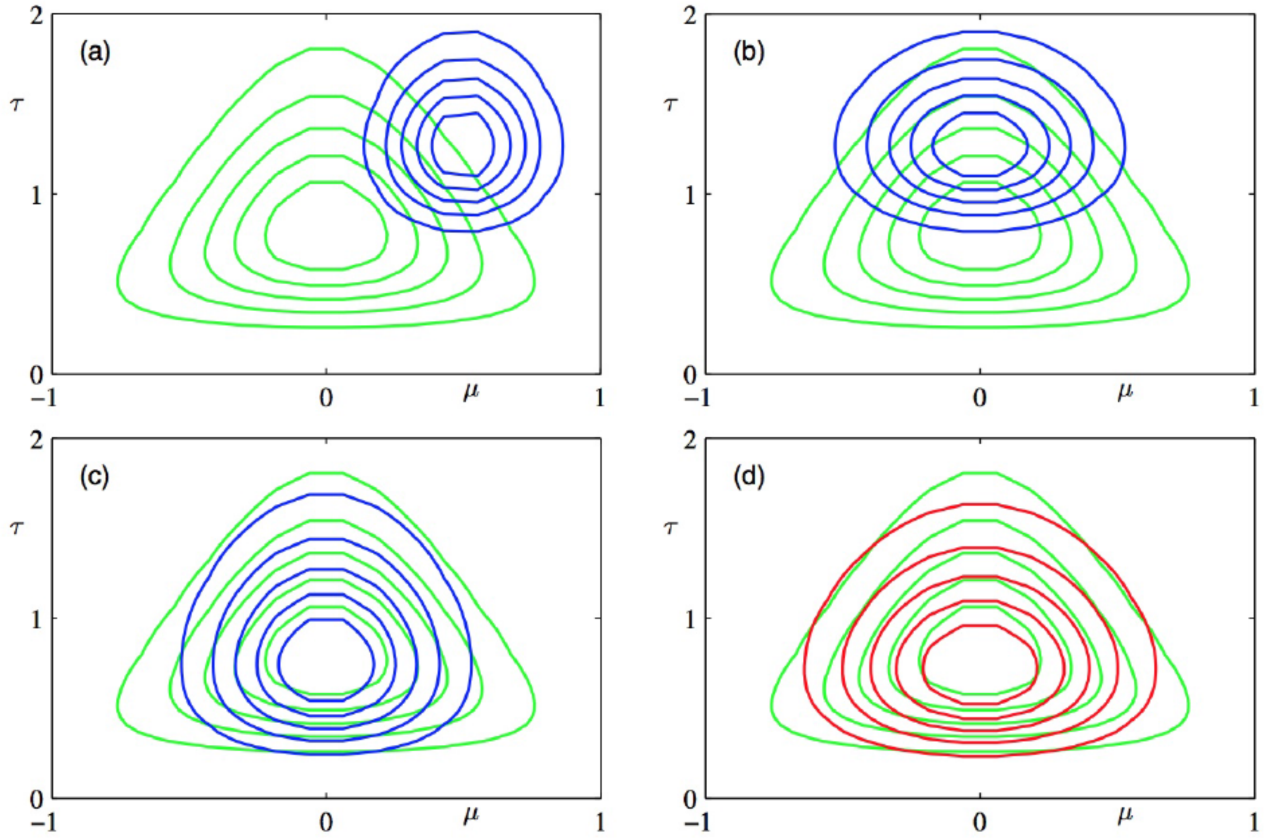
对高斯-伽马分布进行变分推断

我们希望找到一个概率分布函数$q(\mu,\tau)$去近似后验概率分布$P(\mu,\tau|X)$。

首先我们假设$q(\mu,\tau)=q(\mu)q(\tau)$，此时我们的目标就是找到这两个最优的小函数$q(\mu),q(\tau)$，假设我们先求$q(\mu)$。根据已有公式我们可得:

$$
\begin{aligned}
lnq_{\mu}^{*}(\mu) &= \int_{\tau} logP(X,\mu,\tau)q_{\tau}(\tau)d\tau \\
&= \int_{\tau} [logP(X|\mu,\tau) + logP(\mu|\tau) + logP(\tau)]q_{\tau}(\tau)d\tau \\
&= \int_{\tau} [\frac{n}{2}log\frac{\tau}{2\pi} + log\frac{\beta^{\alpha}\sqrt{\lambda_0}}{\Gamma(\alpha)\sqrt{2\pi}} + (\alpha - \frac{1}{2})log\tau - \beta\tau - \frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{\lambda_0\tau}{2}(\mu-\mu_0)^2]q_{\tau}(\tau)d\tau \\
&= \int_{\tau} [-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{\lambda_0\tau}{2}(\mu-\mu_0)^2]q_{\tau}(\tau)d\tau + const \\
&= \int_{\tau} -\frac{\tau}{2}[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2]q_{\tau}(\tau)d\tau + const \\
&= -\frac{\int_{\tau}\tau q_{\tau}(\tau)d\tau}{2}[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2] + const \\
&= -\frac{\int_{\tau}\tau q_{\tau}(\tau)d\tau}{2}[\sum_{i=1}^{n}x_i^2 - 2n\mu\bar{x} + n\mu^2 + \lambda_0\mu^2 - 2\lambda_0\mu_0\mu + \lambda_0\mu_0^2] + const \\
&= -\frac{\int_{\tau}\tau q_{\tau}(\tau)d\tau}{2}[n\mu^2 - 2n\mu\bar{x} + \lambda_0\mu^2 - 2\lambda_0\mu_0\mu] + const \\
&= -\frac{\int_{\tau}\tau q_{\tau}(\tau)d\tau}{2}(n+\lambda_0)(\mu - \frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0})^2 + const \\
&= -\frac{(n+\lambda_0)\int_{\tau}\tau q_{\tau}(\tau)d\tau}{2}(\mu - \frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0})^2 + const \\
&= N(\frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0}, (n+\lambda_0)\int_{\tau}\tau q_{\tau}(\tau)d\tau)
\end{aligned}
$$

得到$q(\mu)$之后，我们可以用同样的方法得到$q(\tau)$，最后不断循环迭代，即采用数值方法不断精确解。



指数族分布

1. 指数族分布的pdf / pmf可以表示成：

$$p(x|\eta) = h(x)exp(T(x)^T\eta - A(\eta))$$

其中，$T(x)$、$h(x)$只是包含x的函数, $A(\eta)$是只包含η的函数。$T(x)$叫做sufficient statistics。$A(\eta)$叫做log-normalizer。在变分推断中，$A(\eta)$起到很重要的作用。

$$\frac{\int h(x)exp(T(x)^T\eta)dx}{exp(A(\eta))} = 1$$

$$A(\eta) = log\int h(x)exp(T(X)^T\eta)dx$$

2. 举高斯分布为例子

$$p(x|\theta) = p(x|\mu, \sigma^2) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}exp - \frac{(x-\mu)^2}{2\sigma^2}$$

3. 我们学到的很多分布都是指数族分布，比如：

Normal, beta, Poisson, gamma, Bernoulli, chi-squared, geometric, exponential, categorical...

4. 例子：怎样把高斯分布写成指数族分布的形式，就是怎样把均值和方差这两个参数替换成η_1, η_2。

$$
\begin{aligned}
N(x|\mu,\ \sigma^2) &= (2\pi\sigma^2)^{-\frac{1}{2}} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \\
&= exp(-\frac{x^2-2x\mu+\mu^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2) \\
&= exp(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)) \\
&= exp(\begin{bmatrix} x \\ x^2 \end{bmatrix}^T \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2))
\end{aligned}
$$

这里，我们得到：

$$
\begin{aligned}
T(x) &= \begin{bmatrix} x \\ x^2 \end{bmatrix} \\
\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} &= \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} \\
\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} &= \begin{bmatrix} \frac{-\eta_1}{2\eta_2} \\ \frac{-1}{2\eta_2} \end{bmatrix} \\
A(\eta) &= \frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}ln(-2\eta_2)
\end{aligned}
$$

所以均值和方差可以表示为：

$$
\eta_2 = -\frac{1}{2\sigma^2} \Rightarrow \sigma^2 = -\frac{1}{2\eta_2}
$$

$$
\mu = \eta_1\sigma^2 = \eta_1\frac{-1}{2\eta_2} = -\frac{\eta_1}{2\eta_2}
$$

5. 指数族分布有什么好处呢？

- 如果一个条件概率可以写成上面的形式，很多问题的求解变得简单。
- 比如：求解$\underset{\theta}{argmax}[logp(X|\eta)]$：

$$
\begin{aligned}
\underset{\eta}{argmax}[logp(X|\eta)] &= \underset{\eta}{argmax}[log\prod_{i=1}^{N}p(x_i|\eta)] \\
&= \underset{\eta}{argmax}\sum_{i=1}^{N}[logh(x_i) + T(x_i)^T\eta - A(\eta)] \\
&= \underset{\eta}{argmax}\sum_{i=1}^{N}T(x_i)^T\eta - NA(\eta)
\end{aligned}
$$

令上式为$L(\eta)$，则

$$\frac{\partial L(\eta)}{\partial \eta}=\sum_{i=1}^{N}T(x_i)-NA'(\eta)=0$$

即：

$$A'(\eta)=\frac{\sum_{i=1}^{N}T(x_i)}{N}$$

6. 共轭：

$$p(\beta|x)\propto p(x|\beta)p(\beta)$$

如果似然函数和先验是共轭的，则后验和先验是同一种分布。

如果似然函数是指数族分布，理论上一定可以找到一个与之共轭的先验分布（也是指数族分布）。

7. 一个结论：$A_l'(\beta)=E_{p(x|\beta)}[T(x)]$

证明：

$$p(x|\beta)=h(x)exp(T(x)^T\beta-A_l(\beta))$$
$$\because \int p(x|\beta)dx=1$$

$$\begin{aligned}
\therefore \frac{\partial\int p(x|\beta)dx}{\partial\beta}&=\frac{\partial\int h(x)exp(T(x)^T\beta-A_l(\beta))dx}{\partial\beta}=0\\
&=\int_x\frac{\partial[h(x)exp[T(x)^T\beta-A_l(\beta)]}{\partial\beta}dx\\
&=\int_x h(x)exp[T(x)^T\beta-A_l(\beta)](T(x)-A_l'(\beta))dx\\
&=\int_x h(x)exp[T(x)^T\beta-A_l(\beta)]T(x)dx-\int_x h(x)exp[T(x)^T\beta-A_l(\beta)]A_l'(\beta))dx\\
&=E_{p(x|\beta)}[T(x)]-A_l'(\beta)=0
\end{aligned}$$

8. 数据集合X，隐变量集合Z，参数集合β。

后验概率分布：

$$\begin{aligned}
p(\beta,Z|X)&=p(\beta|Z,X)p(Z|X)\\
&=p(Z|\beta,X)p(\beta|X)
\end{aligned}$$

$p(\beta|Z,X)$和$p(Z|\beta,X)$，这两个后验分布都是指数族分布。

则：

$$p(\beta|Z,X)=h(\beta)exp(T(\beta)^T\eta(Z,X)-A_l(\eta(Z,X)))$$

在做变分推断时，希望用函数$q(\beta|\lambda)$去近似$p(\beta|Z,X)$，即：

$$p(\beta|Z,X)\approx q(\beta|\lambda)=h(\beta)exp(T(\beta)^T\lambda-A_g(\lambda))$$

接下来，就要不断地调整λ，使得$q(\beta|\lambda)$越来越接近于$p(\beta|Z,X)$，即增大$ELOW$函数。

同样的，对于$p(Z|\beta,X)$也是如此：

$$\begin{aligned}p(Z|\beta,X)&=h(Z)exp(T(Z)^T\eta(\beta,X)-A_l(\eta(\beta,X)))\\&\approx q(Z|\phi)=h(Z)exp(T(Z)^T\phi-A_g(\phi))\end{aligned}$$

$ELOB$函数如下：

$$L(q)=E_{q(Z,\beta)}[logp(X,Z,\beta)]-E_{q(Z,\beta)}[logq(Z,\beta)]$$

现在，$ELOB$函数可以写成：

$$L(\lambda,\phi)=E_{q(Z,\beta)}[logP(X,Z,\beta)]-E_{q(Z,\beta)}[logq(Z,\beta)]$$

目标：找到一个λ和ϕ，使得$ELOB$函数最大化。

方法：

- **先固定一个参数，对另一个参数优化**

具体做法：

- 固定ϕ，优化λ
- $$\begin{aligned}L(\lambda,\phi)&=E_{q(Z,\beta)}[logp(X,Z,\beta)]-E_{q(Z,\beta)}[logq(Z,\beta)]\\&=E_{q(Z,\beta)}[logp(\beta|X,Z)+logp(Z|X)]-E_{q(Z,\beta)}[logq(\beta)]-E_{q(Z,\beta)}[logq(Z)]\\&=E_{q(Z,\beta)}[logp(\beta|X,Z)]-E_{q(Z,\beta)}[logq(\beta|\lambda)]\end{aligned}$$
- 将$p(\beta|Z,X)$和$q(\beta|\lambda)$代入上式
- $$\begin{aligned}L(\lambda,\phi)&=E_{q(Z,\beta)}[logh(\beta)]+E_{q(Z,\beta)}[T(\beta)^T\eta(Z,X)]-E_{q(Z,\beta)}[A_g(\eta(X,Z))]-E_{q(Z,\beta)}[logh(\beta)]-E_{q(Z,\beta)}[(T(\beta)^T\lambda)]+E_{q(Z,\beta)}[A_g(\lambda)]\\&=E_{q(\beta)}[T(\beta)^T]\cdot E_{q(Z)}[\eta(Z,X)]-E_{q(Z)}[A_g(\eta(X,Z))]-E_{q(\beta)}[(T(\beta)^T\lambda)]+A_g(\lambda)\\&=A_g'(\lambda)^TE_{q(Z)}[\eta(Z,X)]-\lambda A_g'(\lambda)^T+A_g(\lambda)\end{aligned}$$
- 上式对λ求导
- $$\begin{aligned}\frac{\partial L(\lambda,\phi)}{\partial\lambda}&=A_g''(\lambda)^T\cdot E_{q(Z)}[\eta(Z,X)]-A_g'(\lambda)^T-\lambda A_g''(\lambda)^T+A_g'(\lambda)\\&=A_g''(\lambda)^T(E_{q(Z)}[\eta(Z,X)]-\lambda)=0\end{aligned}$$
- 如果$A_g''(\lambda)^T\neq0$，则

  $$\lambda=E_{q(Z|\phi)}[\eta(Z,X)]$$

  同样

  $$\phi=E_{q(\beta|\lambda)}[\eta(X,\beta)]$$