Approximate Inference

用KL Divergence在本次计算中很方便。所以用KL Divergence去度量两个分布之间的距离。

- VB is fast, streaming and distributed (millions of observations)
- **Variational Bayes:** To find the $q^*(z)$ with the smallest KL divergence to $p(z|x)$

$$
\begin{aligned}
q^*(z) &= \underset{q(z)\in Q}{\arg\min}\, \mathrm{KL}\big(q(z)||p(z|x)\big) \\
&= \underset{q(z)\in Q}{\arg\min} - \int_z q(z)\log\left[\frac{p(z \mid x)}{q(z)}\right] dz
\end{aligned}
$$

- $q^*(z)$ is the best approximation to $p(z|x)$ within Q
- $q^*(z) = p(z \mid x)$ if there is no constraint, but $p(z \mid x)$ itself is unknown

现在产生了一个问题，首先后验分布是不知道的，所以要通过算 q(z) 去近似 后验分布，可是要算 q(z) 又要知道后验分布，变成一个循环论证的问题。所以我们不能直接用KL Divergence去计算，而是要进行某种转换。所以我们把它看作优化问题，什么时候我们能取到理论上的最优值，答案是$q^*(z) = p(z \mid x)$。我们可以把KL Divergence转化，找到一个可以计算的项。

$$
\begin{aligned}
\mathrm{KL}(q(z)||p(z \mid x)) &= -\int_z q(z)\log\left[\frac{p(z \mid x)}{q(z)}\right] dz \\
&= \int_z q(z)\log q(z)dz - \int_z q(z)\log p(z \mid x)dz \\
&= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z \mid x)] \\
&= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q\left[\log\left[\frac{p(x,z)}{p(x)}\right]\right] \\
&= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(x,z)] + \mathbb{E}_q[\log p(x)] \\
&= \underbrace{\mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(x,z)]}_{-\mathrm{ELBO}} + \log p(x)
\end{aligned}
$$

- Note: $\mathbb{E}_q\big[\log p(x)\big] = \log p(x)$

将积分表达式 ∫q(z) log q(z) dz转化成期望 E_q[log q(z)] 的过程涉及到概率密度函数 q(z) 和随

机变量 z 的期望值。数学原理如下：
首先，∫q(z) log q(z) dz表示对 q(z) 关于变量 z 的积分。这通常用于信息论、熵和期望的计算。要将它转化为期望的形式，我们可以将 q(z) 视为一个概率密度函数，并将 log q(z) 视为一个关于 z 的函数。然后，我们可以将期望的定义应用到这个表达式中。期望 E_q[log q(z)] 表示在概率密度函数 q(z) 下，对 log q(z) 进行加权平均。具体来说，它计算了 log q(z) 乘以 q(z) 并对所有可能的 z 值进行积分。这两个表达式实际上是等价的。将积分形式转化为期望形式是因为它更适合描述概率分布和期望值的计算。

此时，变成了三项，观察各项，发现第三项里面 log p(x) 与期望的对象 q(z) 是无关的，所以期望符号可以直接去掉。