变分推断

我们已经知道概率模型可以分为，频率派的优化问题和贝叶斯派的积分问题。从贝叶斯角度来看推断，对于 $\hat{x}$ 这样的新样本，需要得到：

$$p(\hat{x}|X)=\int_\theta p(\hat{x},\theta|X)d\theta=\int_\theta p(\theta|X)p(\hat{x}|\theta,X)d\theta \tag{1}$$

如果新样本和数据集独立，那么推断就是概率分布依参数后验分布的期望。

我们看到，推断问题的中心是参数后验分布的求解，推断分为：

1. 精确推断
2. 近似推断-参数空间无法精确求解
    1. 确定性近似-如变分推断
    2. 随机近似-如 MCMC，MH，Gibbs

基于平均场假设的变分推断

我们记 Z 为隐变量和参数的集合，Z_i 为第 i 维的参数，于是，回顾一下 EM 中的推导：

$$\log p(X)=\log p(X,Z)-\log p(Z|X)=\log\frac{p(X,Z)}{q(Z)}-\log\frac{p(Z|X)}{q(Z)} \tag{2}$$

左右两边分别积分：

$$\begin{aligned}&Left:\int_Zq(Z)\log p(X)dZ=\log p(X)\\&Right:\int_Z[\log\frac{p(X,Z)}{q(Z)}-\log\frac{p(Z|X)}{q(Z)}]q(Z)dZ=ELBO+KL(q,p)\end{aligned} \tag{3}$$

第二个式子可以写为变分和 KL 散度的和：

$$L(q)+KL(q,p) \tag{4}$$

由于这个式子是常数，于是寻找 $q\simeq p$ 就相当于对 $L(q)$ 最大值。

$$\hat{q}(Z)=\mathop{argmax}_{q(Z)}L(q) \tag{5}$$

假设 $q(Z)$ 可以划分为 M 个组（平均场近似）：

$$q(Z)=\prod_{i=1}^Mq_i(Z_i) \tag{6}$$

因此，在 $L(q)=\int_Zq(Z)\log p(X,Z)dZ-\int_Zq(Z)\log q(Z)$ 中，看 $p(Z_j)$ ，第一项：

$$
\begin{aligned}
\int_Z q(Z)\log p(X,Z)dZ &= \int_Z \prod_{i=1}^M q_i(Z_i)\log p(X,Z)dZ \\
&= \int_{Z_j} q_j(Z_j)\int_{Z-Z_j}\prod_{i\neq j} q_i(Z_i)\log p(X,Z)dZ \\
&= \int_{Z_j} q_j(Z_j)\mathbb{E}_{\prod\limits_{i\neq j} q_i(Z_i)}[\log p(X,Z)]dZ_j
\end{aligned} \tag{7}
$$

第二项：

$$
\int_Z q(Z)\log q(Z)dZ = \int_Z \prod_{i=1}^M q_i(Z_i)\sum_{i=1}^M \log q_i(Z_i)dZ \tag{8}
$$

展开求和项第一项为：

$$
\int_Z \prod_{i=1}^M q_i(Z_i)\log q_1(Z_1)dZ = \int_{Z_1} q_1(Z_1)\log q_1(Z_1)dZ_1 \tag{9}
$$

所以：

$$
\int_Z q(Z)\log q(Z)dZ = \sum_{i=1}^M \int_{Z_i} q_i(Z_i)\log q_i(Z_i)dZ_i = \int_{Z_j} q_j(Z_j)\log q_j(Z_j)dZ_j + Const \tag{10}
$$

两项相减，令 $\mathbb{E}_{\prod\limits_{i\neq j} q_i(Z_i)}[\log p(X,Z)] = \log\hat{p}(X,Z_j)$ 可以得到：

$$
-\int_{Z_j} q_j(Z_j)\log\frac{q_j(Z_j)}{\hat{p}(X,Z_j)}dZ_j \leq 0 \tag{11}
$$

于是最大的 $q_j(Z_j)=\hat{p}(X,Z_j)$ 才能得到最大值。我们看到，对每一个 q_j，都是固定其余的 q_i，求这个值，于是可以使用坐标上升的方法进行迭代求解，上面的推导针对单个样本，但是对数据集也是适用的。

基于平均场假设的变分推断存在一些问题：

1. 假设太强，Z 非常复杂的情况下，假设不适用
2. 期望中的积分，可能无法计算

SGVI

从 Z 到 X 的过程叫做生成过程或译码，反过来的额过程叫推断过程或编码过程，基于平均场的变分推断可以导出坐标上升的算法，但是这个假设在一些情况下假设太强，同时积分也不一定能算。我们知道，优化方法除了坐标上升，还有梯度上升的方式，我们希望通过梯度上升来得到变分推断的另一种算法。

我们的目标函数：

$$
\hat{q}(Z) = \underset{q(Z)}{argmax}\, L(q) \tag{12}
$$

假定 $q(Z)=q_\phi(Z)$，是和 ϕ 这个参数相连的概率分布。于是 $argmax_{q(Z)}\, L(q) = argmax_\phi\, L(\phi)$，其中 $L(\phi)=\mathbb{E}_{q_\phi}[\log p_\theta(x^i,z)-\log q_\phi(z)]$，这里 x^i 表示第 i 个样本。

$$
\begin{aligned}
\nabla_\phi L(\phi) &= \nabla_\phi \mathbb{E}_{q_\phi}[\log p_\theta(x^i,z)-\log q_\phi(z)] \\
&= \nabla_\phi \int q_\phi(z)[\log p_\theta(x^i,z)-\log q_\phi(z)]dz \\
&= \int \nabla_\phi q_\phi(z)[\log p_\theta(x^i,z)-\log q_\phi(z)]dz + \int q_\phi(z)\nabla_\phi[\log p_\theta(x^i,z)-\log q_\phi(z)]dz \\
&= \int \nabla_\phi q_\phi(z)[\log p_\theta(x^i,z)-\log q_\phi(z)]dz - \int q_\phi(z)\nabla_\phi \log q_\phi(z)dz \\
&= \int \nabla_\phi q_\phi(z)[\log p_\theta(x^i,z)-\log q_\phi(z)]dz - \int \nabla_\phi q_\phi(z)dz \\
&= \int \nabla_\phi q_\phi(z)[\log p_\theta(x^i,z)-\log q_\phi(z)]dz \\
&= \int q_\phi(\nabla_\phi \log q_\phi)(\log p_\theta(x^i,z)-\log q_\phi(z))dz \\
&= \mathbb{E}_{q_\phi}[(\nabla_\phi \log q_\phi)(\log p_\theta(x^i,z)-\log q_\phi(z))]
\end{aligned} \tag{13}
$$

这个期望可以通过蒙特卡洛采样来近似，从而得到梯度，然后利用梯度上升的方法来得到参数：

$$
z^l \sim q_\phi(z) \tag{14}
$$

$$
\mathbb{E}_{q_\phi}[(\nabla_\phi \log q_\phi)(\log p_\theta(x^i,z)-\log q_\phi(z))] \sim \frac{1}{L}\sum_{l=1}^{L}(\nabla_\phi \log q_\phi)(\log p_\theta(x^i,z)-\log q_\phi(z))
$$

但是由于求和符号中存在一个对数项，于是直接采样的方差很大，需要采样的样本非常多。为了解决方差太大的问题，我们采用 Reparameterization 的技巧。

考虑：

$$
\nabla_\phi L(\phi) = \nabla_\phi \mathbb{E}_{q_\phi}[\log p_\theta(x^i,z)-\log q_\phi(z)] \tag{15}
$$

我们取：$z=g_\phi(\varepsilon,x^i),\varepsilon\sim p(\varepsilon)$，于是对后验：$z\sim q_\phi(z|x^i)$，有 $|q_\phi(z|x^i)dz|=|p(\varepsilon)d\varepsilon|$。代入上面的梯度中：

$$
\begin{aligned}
\nabla_\phi L(\phi) &= \nabla_\phi \mathbb{E}_{q_\phi}[\log p_\theta(x^i,z)-\log q_\phi(z)] \\
&= \nabla_\phi L(\phi) = \nabla_\phi \int [\log p_\theta(x^i,z)-\log q_\phi(z)]q_\phi dz \\
&= \nabla_\phi \int [\log p_\theta(x^i,z)-\log q_\phi(z)]p_\varepsilon d\varepsilon \\
&= \mathbb{E}_{p(\varepsilon)}[\nabla_\phi[\log p_\theta(x^i,z)-\log q_\phi(z)]] \\
&= \mathbb{E}_{p(\varepsilon)}[\nabla_z[\log p_\theta(x^i,z)-\log q_\phi(z)]\nabla_\phi z] \\
&= \mathbb{E}_{p(\varepsilon)}[\nabla_z[\log p_\theta(x^i,z)-\log q_\phi(z)]\nabla_\phi g_\phi(\varepsilon,x^i)]
\end{aligned} \tag{16}
$$

对这个式子进行蒙特卡洛采样，然后计算期望，得到梯度。