第七章、概率图模型与推断

赵涵

2023 年 4 月 18 日

概率图模型使用图的方式表示概率分布。为了在图中添加各种概率，首先总结一下随机变量分布的一些规则：

- 求和规则：$p(x_1) = \int p(x_1, x_2)dx_2$
- 条件概率：$p(x_1, x_2) = p(x_1|x_2)p(x_2)$
- 链式规则：$p(x_1, x_2, \ldots, x_p) = \prod_{i=1}^{p} p(x_i|x_{i+1}, x_{i+2} \ldots x_p)$
- 贝叶斯规则：$p(x_1|x_2) = \frac{p(x_2|x_1)p(x_1)}{p(x_2)}$

上面所有的小写字母x都为随机变量，可以是标量也可以是向量。可以看到，在链式法则中，如果数据维度特别高，那么采样和计算非常困难，我们需要在一定程度上作出简化，在朴素贝叶斯中，作出了条件独立性假设。在Markov 假设中，给定数据的维度是以时间顺序出现的，给定当前时间的维度，那么下一个维度与之前的维度独立。在HMM（在后面的章节种后介绍该模型）中，采用了齐次Markov假设。在Markov假设之上，更一般的，加入条件独立性假设，对维度划分集合，使得$X_A \perp X_B|X_C$。

概率图模型采用图的特点表示上述的条件独立性假设，节点表示随机变量，边表示条件概率。概率图模型可以分为三大理论部分：

1. 表示：
   (a) 有向图（离散）：贝叶斯网络
   (b) 高斯图（连续）：高斯贝叶斯和高斯马尔可夫网路
   (c) 无向图（离散）：马尔科夫网络
2. 推断：
   (a) 精确推断
   (b) 近似推断
      i. 确定性近似——变分推断
      ii. 随机近似——马尔可夫蒙特卡罗方法
3. 学习
   (a) 参数学习
      i. 完备数据
      ii. 隐藏变量：EM算法
   (b) 结构学习

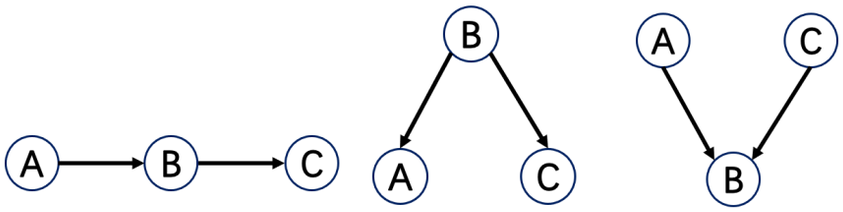
接下来就根据上述的分类，对这三大类进行详细的介绍。

1 有向图——贝叶斯网络

已知联合分布中，各个随机变量之间的依赖关系，那么可以通过拓扑排序（根据依赖关系）可以获得一个有向图。而如果已知一个图，也可以直接得到联合概率分布的因子分解：

$$p(x_1, x_2, \ldots, x_p) = \prod_{i=1}^{p} p(x_i | x_{parent(i)}) \tag{1}$$

那么实际的图中条件独立性是如何体现的呢？在局部任何三个节点，可以有三种结构： 我们先看图



(a) (b) (c)

Figure 1: (a)链式结构；(b)倒V结构；(b)正V 结构。

（a）这三个随机变量的联合概率分布：

$$p(A, B, C) = p(A)p(B|A)p(C|B) = p(A)p(B|A)p(C|B, A) \tag{2}$$

$$\Rightarrow p(C|B)p(C|B, A) \tag{3}$$

$$\Leftrightarrow p(C|B)p(A|B) = p(C|A, B)p(A|B) = p(C, A|B) \tag{4}$$

$$\Rightarrow C \perp A|B \tag{5}$$

我们可以看出，在随机变量B被观测到时，那么A 与C是相互独立的。接下来看图（b）的联合概率分布：

$$p(A, B, C) = p(A|B)p(B)p(C|B) = p(B)p(A|B)p(C|A, B) \tag{6}$$

$$\Rightarrow p(C|B) = p(C|B, A) \tag{7}$$

$$\Leftrightarrow p(C|B)p(A|B) = p(C|B, A)p(A|B) = p(C, A|B) \tag{8}$$

$$\Rightarrow C \perp A|B \tag{9}$$

根据联合概率分布，我们可以看到在随机变量B 被观测到时，那么A与C是相互独立的。接下来看图（c）的联合概率分布：

$$p(A, B, C) = p(A)p(C)p(B|C, A) = p(A)p(C|A)p(B|C, A) \tag{10}$$

$$\Rightarrow p(C) = p(C|A) \tag{11}$$

$$\Leftrightarrow C \perp A \tag{12}$$

我们可以看到，这种V型结构很有特点，在B 在观测后，A 与C不独立，但是在不观测B 时，A 与C 独立，换句话说，A，C不与B 条件独立。

从整体的图来看，可以引入**D划分**（D-Segmentation）的概念。对于类似上面图（a）和图（b）的关系，引入集合A，B，那么满足$A \perp B|C$的C集合中的点与A, B 中的点的关系都满足图（a）（b），满足图（c）关系的点都不在C 中。D划分应用在贝叶斯定理中：

$$p(x_i|x_{-i}) = \frac{p(x)}{\int p(x)dx_i} = \frac{\prod_{j=1}^{p} p(x_j|x_{parents(j)})}{\int \prod_{j=1}^{p} p(x_j|x_{parents(j)})dx_i} \tag{13}$$

可以发现，上下部分可以分为两部分，一部分是和x_i 相关的，另一部分是和x_i无关的，而这个无关的部分可以相互约掉。于是计算只涉及和x_i相关的部分。与x_i相关的部分可以写成：

$$p(x_i|x_{parents(i)})p(x_{child(i)}|x_i) \tag{14}$$

把x_i的相关随机变量可视化后，可以看出形状像一个毯子，称为Markov毯。实际应用的模型中，对这些条件独立性作出了假设，从单一到混合，从有限到无限（时间，空间）可以分为：

1. 朴素贝叶斯，单一的条件独立性假设$p(x|y) = \prod_{i=1}^{p} p(x_i|y)$，在D 划分后，所有条件依赖的集合就是单个元素。
2. 高斯混合模型：混合的条件独立。引入多类别的隐变量$z_1, z_2, \ldots, z_k$，$p(x|z) = N(\mu, \Sigma)$，条件依赖集合为多个元素。
3. 与时间相关的条件依赖：
   - Markov链
   - 高斯过程（无限维高斯分布）
4. 连续：高斯贝叶斯网络
5. 组合上面的分类：
   - 高斯混合模型与时序结合：动态模型
     - 隐马尔可夫模型（离散型随机变量）
     - 线性动态系统，卡曼滤波
     - 粒子滤波（非高斯，非线性）

2 无向图——马尔科夫随机场

上面我们讨论了有向图，接下来再看一下无向图。无向图没有了类似有向图的局部不同结构，在马尔可夫网络中，也存在D 划分的概念。直接将条件独立的集合$x_A \perp x_B|x_C$ 划分为三个集合。这个也叫全局Markov。对局部的节点，$x \perp (X - Neighbour(x)|Neighbour(x))$，这里$X$ 为所有的随机变量集合，这种条件独立性也叫局部Markov。对于成对的节点：$x_i \perp x_j|x_{-i-j}$，其中i, j不能相邻。称为成对Markov。事实上以上三个假设，全局局部成对是相互等价的。有了这个条件独立性的划分，还需要因子分解来实际计算。引入**团**（clique）的概念：团与最大团：给定图$G = (V, E)$。其中，$V = \{1, \ldots, n\}$ 是图G 的顶点集，E是图G的边集。图G 的团就是一个两两之间有边的顶点集合。简单地说，团是G的一个完全子图。如果一个团不被其他任一团所包含，即它不是其他任一团的真子集，则称该团为图G 的**极大团**（maximal clique）。顶点最多的极大团，称之为图G的**最大团**（maximum clique）。利用这个定义进行的x所有维度的联合概率分布的因子分解为，假设有K个团，Z就是对所有可能取值求和：

$$p(x) = \frac{1}{Z}\prod_{i=1}^{K}\phi(x_{ci}) \tag{15}$$

$$Z = \sum_{x_{ci}\in\mathcal{X}}\prod_{i=1}^{K}\phi(x_{ci}) \tag{16}$$

其中$\phi(x_{ci})$叫做势函数，它必须是一个正值[1]，可以记为：

$$\phi(x_{ci}) = \exp(-E(x_{ci})) \tag{17}$$

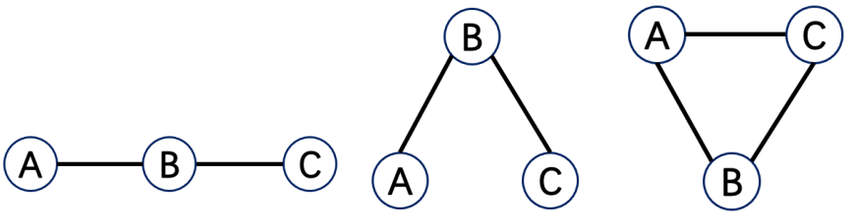
这种形式的分布，称为玻尔兹曼分布，Z被称为配分函数，那么$p(x)$为：

$$p(x) = \frac{1}{Z}\exp(-\sum_{i=1}^{K} E(x_{ci})) \tag{18}$$

这个分解和条件独立性等价（Hammesley-Clifford 定理），这个分布的形式也和指数族分布形式上相同，于是满足最大熵原理。

3 两种图的转化——道德图

我们常常想将有向图转为无向图，从而应用更一般的表达式，把这种方法称为**道德图**（moral graph）。 对于图（a），我们直接去掉箭头：



(a) (b) (c)

Figure 2: (a)链式结构的转化；(b)倒V结构的转化；(b) 正V结构的转化。

$$p(A,B,C) = P(A)p(B|A)p(C|B) = \phi(A,B)\phi(B,C) \tag{19}$$

对于图（b），也是直接去掉箭头：

$$p(A,B,C) = p(B)p(A|B)p(C|B) = \phi(A,B)\phi(B,C) \tag{20}$$

对于图（c），不仅需要去掉箭头，还要把父节点连接起来：

$$p(A,B,C) = p(A)p(C)p(B|A,C) = \phi(A,B,C) \tag{21}$$

总结三种情况可以概括为：

- 将每个节点的父节点两两相连；
- 将有向边替换为无向边。

对于一个有向图，可以通过引入环的方式，可以将其转换为无向图（Tree-like graph），这个图就叫做道德图。但是我们后面会介绍一种算法，称为信念传播算法，只对无环图有效，有环时，得到的结果是近似的，有一种方法，通过因子图可以把有环图变为无环图。

3.1 因子图

考一个无向有环图，如图（c），可以将其转化为： 其中$f = f(A,B,C)$。因子图不是唯一的，这是由于因式分解本身就对应一个特殊的因子图，将因式分解：$p(x) = \prod_s f_s(x_s)$可以进一步分解得到因子图。

[1]在这里我们不在给出严格的数学证明，从概率的定义可以看出，概率必须具有非负性，所以每个因子也都必须是非负性。

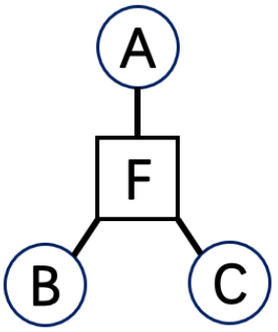


Figure 3: 因子图。

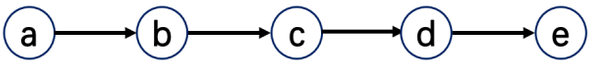


Figure 4: 有向图的链式结构，推断随机变量d 的边缘概率分布。

4 推断

推断的主要目的是求各种概率分布，包括边缘概率，条件概率，以及使用**最大后验概率**（Maximum A Posterior）来求得参数。通常推断可以分为：

1. 精确推断
   (a) **变量消除**（Variable Elimination）
   (b) **信念传播**（Belief Propagation），由VE发展而来
   (c) **交叉树**（Junction Tree），上面两种在树结构上应用，Junction Tree 在图结构上应用
2. 近似推断
   (a) **带环的信念传播**（Loop Belief Propagation）
   (b) **蒙特卡洛推断**（Mente Carlo Inference）：重要性采样，MCMC
   (c) **变分推断**（Variational Inference）

我们首先介绍精确推断。

4.1 精确推断

4.1.1 变量消除

变量消除的方法是在求解概率分布的时候，将相关的条件概率先行求和或积分，从而一步步地消除变量，例如在马尔可夫链中： 我们假设随机变量是离散型随机变量（连续型随机变量，只需把求和号改写为积分号即可），对联合概率分布求和，只留下随机变量d，得到随机变量d 的边缘概率分布：

$$p(e) = \sum_{a,b,c,d} p(a,b,c,d,e) = \sum_{d} p(e|d) \sum_{c} p(d|c) \sum_{b} p(c|b) \sum_{a} p(b|a)p(a) \tag{22}$$

变量消除的缺点很明显：

1. 计算步骤无法存储；

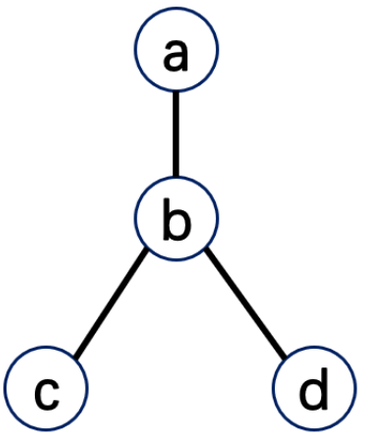


Figure 5: 无向图下的消息传递。

2. 消除的最优次序是一个NP-hard问题。

为了克服上面的困难，我们对上面的有向图进行观察，进而得到信念传播算法。

4.1.2 信念传播算法

我们写出中间随机变量的概率分布：

$$p(c) = \sum_{a,b,d,e} p(a,b,c,d,e) = (\sum_{b} p(c|b) \sum_{a} p(b|a)p(a))(\sum_{d} p(d|c) \sum_{e} p(e|d)) \tag{23}$$

可以看到，无论是对谁求边缘分布，都可以发现是逐步求和，像消息一样从头传递到目标节点，所以我们把这种形式称为消息传递，原则上对于没有圈的图都可以精确计算。我们对边缘概率里面的求和项，给出如下的定义：

$$\sum_{a} p(a)p(b|a) = m_{a\to b}(b) \tag{24}$$

$$\sum_{b} p(c|b)m_{a\to b}(b) = m_{b\to c}(c) \tag{25}$$

所以我们可以得到：

$$p(e) = \sum_{d} p(e|d)m_{d\to e}(e) \tag{26}$$

接下来我们看无向图： 这四个团（对于无向图是团，对于有向图就是概率为除了根的节点为1），有四个节点，三个边：

$$p(a,b,c,d) = \frac{1}{Z}\phi_a(a)\phi_b(b)\phi_c(c)\phi_d(d)\phi_{ab}(a,b)\phi_{bc}(b,c)\phi_{bd}(b,d) \tag{27}$$

我们套用上面关于有向图的观察，如果求解边缘概率$p(a)$，定义：

$$m_{c\to b}(b) = \sum_{c} \phi_c(c)\phi_{bc}(b,c) \tag{28}$$

$$m_{d\to b}(b) = \sum_{d} \phi_d(d)\phi_{bd}(b,d) \tag{29}$$

$$m_{b\to a}(a) = \sum_{b} \phi_b(b)\phi_{ba}(b,a)m_{c\to b}(b)m_{d\to b}(b) \tag{30}$$

这样概率就一步步地传播到了a：

$$p(a) = \phi_a(a) m_{b \to a}(a) \tag{31}$$

写成一般的形式，对于相邻节点i, j：

$$m_{j \to i}(i) = \sum_j \phi_j(j) \phi_{ij}(ij) \prod_{k \in Neighbour(j) - i} m_{k \to j}(j) \tag{32}$$

这个表达式，就可以保存计算过程了，只要对每条边的传播分别计算，对于一个无向树形图可以递归并行实现：

1. 任取一个节点a作为根节点；
2. 对这个根节点的邻居中的每一个节点，收集信息（计算入信息）
3. 对根节点的邻居，分发信息（计算出信息）

我们上述方法称为**信念传播算法**（Belief Propagation algorithm），又称为**消息传递**（message passing）算法。

4.1.3 Max-Product算法

在推断任务中，MAP也是常常需要的，MAP 的目的是寻找最佳参数：

$$(\hat{a}, \hat{b}, \hat{c}, \hat{d}) = \arg \max_{a,b,c,d} p(a, b, c, d|E) \tag{33}$$

这里的E表示剩余的随机变量。类似BP，我们采用信息传递的方式来求得最优参数，不同的是，我们在所有信息传递中，传递的是最大化参数的概率，而不是将所有可能求和：

$$m_{j \to i} = \max_j \phi_j \phi_{ij} \prod_{k \in Neighbour(j) \to i} m_{k \to j} \tag{34}$$

于是对于上面的图：

$$\max_a p(a, b, c, d) = \max_a \phi_a \phi_{ab} m_{c \to b} m_{d \to b} \tag{35}$$

这个算法是Sum-Product算法的改进，也是在HMM 中应用给的Viterbi算法的推广。

4.2 近似推断

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，需要计算：

$$p(\hat{x}|X) = \int_\theta p(\hat{x}, \theta|X) d\theta = \int_\theta p(\theta|X) p(\hat{x}|\theta, X) d\theta \tag{36}$$

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。换句话说，推断问题的中心是参数后验分布的求解，这种积分，精确推断往往无能为力，所以必须面临取近似的行为，对于取近似又分为两种：

- 确定性近似——变分推断
- 随机近似——MCMC，Gibbs

我们接下来讲解变分推断。

4.2.1 确定性近似——变分推断

作为变分推断，一般情况下近似方法就是平均场假设。作为概率图的延申，对于一个数据集合，这些数据可以认为是随机变量集合，我们认为这些随机变量背后会有一个小的随机变量集合，称为隐藏变量。我们记Z 为隐变量和参数的集合，Z_i 为第i 维的参数，把对数据X与隐变量Z之间的联合概率分布，利用贝叶斯规则，得到原始数据的概率分布：

$$p(X) = p(X, Z)/p(Z) \tag{37}$$

两边同时取对数：

$$\log p(X) = \log p(X, Z) - \log p(Z|X) = \log \frac{p(X, Z)}{q(Z)} - \log \frac{p(Z|X)}{q(Z)} \tag{38}$$

第二个等式加上一个$\log q(Z)$，减去一个$\log q(Z)$，不改变函数的值，$q(Z)$是隐变量的先验分布。左右两边同时乘$q(Z)$，然后分别对Z积分：

$$Left : \int_Z q(Z) \log p(X) dZ = \log p(X) \tag{39}$$

$$Right : \int_Z [\log \frac{p(X, Z)}{q(Z)} - \log \frac{p(Z|X)}{q(Z)}] q(Z) dZ = ELBO + KL(q, p) \tag{40}$$

右边的式子积分后分为两项，一项称为**证据下界**（Evidence Lower Bound），第二项是我们之前介绍特征选择时，已经见到过的，称为KL散度。我们把证据下界记为$L(q)$，因为左边是原始数据的概率分布，它不以模型为转移，换句话说，应该把$p(X)$认为是一个未知的常数，所以先验分布$q(Z)$越接近$p(Z|X)$，KL散度趋于0，这是一个路径，另一方面，我们可以通过让ELBO尽可能的大，趋近与$p(X)$，这两种路径，本质是等价的。我们这里从ELBO出发：

$$\hat{q}(Z) = \arg\max_{q(Z)} L(q) \tag{41}$$

假设$q(Z)$可以划分为M个组（平均场近似）：

$$q(Z) \approx \prod_{i=1}^{M} q_i(Z_i) \tag{42}$$

因此，在$L(q) = \int_Z q(Z) \log p(X, Z) dZ - \int_Z q(Z) \log q(Z)$中，把近似方程代入，第一项可以化为：

$$\begin{aligned} \int_Z q(Z) \log p(X, Z) dZ &= \int_Z \prod_{i=1}^{M} q_i(Z_i) \log p(X, Z) dZ & (43) \\ &= \int_{Z_j} q_j(Z_j) \int_{Z - Z_j} \prod_{i \neq j} q_i(Z_j) \log p(X, Z) dZ & (44) \\ &= \int_{Z_j} q_j(Z_j) E_{\prod_{i \neq j} q_i(Z_i)}[\log p(X, Z)] dZ_j & (45) \end{aligned}$$

第二项为：

$$\int_Z q(Z) \log q(Z) dZ = \int_Z \prod_{i=1}^{M} q_i(Z_i) \sum_{i=1}^{M} \log q_i(Z_i) dZ \tag{46}$$

对数把连乘转化为求和，所以观察这个求和的第一项为：

$$\int_Z \prod_{i=1}^{M} q_i(Z_i) \log q_1(Z_1) dZ = \int_{Z_i} q_1(Z_1) \log q_1(Z_1) dZ_1 \tag{47}$$

所以：

$$\int_Z q(Z)\log q(Z)dZ = \sum_{i=1}^{M}\int_{Z_i} q_i(Z_i)\log q_i(Z_i)dZ_i \tag{48}$$

$$= \int_{Z_j} q_j(Z_j)\log q_j(Z_j)dZ_j + Const \tag{49}$$

第二个等式，意味着我们只关心与Z_j有关的项，其余的默认为常数。把第一项中的$E_{\prod_{i\neq j} q_i(z_i)}[\log p(X,Z)] = \log \hat{p}(X,Z_j)$，那么把这两项代入到原方程，有：

$$\int_{Z_j} q_j(Z_j)\log \hat{p}(X,Z_j)dZ_j - \int_{Z_j} q_j(Z_j)\log q_j(Z_j)dZ_j = \int_{Z_j} q_j(Z_j)\log\frac{\hat{p}(X,Z_j)}{q_j(Z_j)} \tag{50}$$

$$= -KL(q_j(Z_j),\hat{p}(X,Z_j)) \leq 0 \tag{51}$$

$$\tag{52}$$

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。

下面我们采用梯度上升的方法，逐步逼近最大值。

4.2.2 随机梯度变分推断

从Z到X的过程叫做生成过程或译码，反过来的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：

$$\hat{q}(Z) = \arg\max_{q(Z)} L(q) \tag{53}$$

假定$q(Z) = q_\phi(Z)$，是和ϕ这个参数相连的概率分布。于是$\arg\max_{q(Z)} L(q) = \arg\max_\phi L(\phi)$，其中$L(\phi) = E_{q_\phi}[\log p_\phi(x^{(i)},Z) - \log q_\phi(Z)]$，这里$x^{(i)}$表示第$i$个样本。那么目标函数的梯度为：

$$\nabla L(\phi) = \nabla_\phi E_{q_\phi}[\log p_\phi(x^{(i)},Z) - \log q_\phi(Z)] \tag{54}$$

$$= \nabla_\phi \int q_\phi(Z)[\log p_\phi(x^{(i)},Z) - \log q_\phi(Z)]dZ \tag{55}$$

$$= \int \nabla_\phi q_\phi(Z)[\log p_\phi(x^{(i)},Z) - \log q_\phi(Z)]dZ + \int q_\phi(Z)\nabla_\phi[\log p_\phi(x^{(i)},Z) - \log q_\phi(Z)]dZ \tag{56}$$

$$= \int \nabla_\phi q_\phi(Z)[\log p_\phi(x^{(i)},Z) - \log q_\phi(Z)]dZ - \int q_\phi(Z)\nabla_\phi \log q_\phi(Z)dZ \tag{57}$$

$$= \int \nabla_\phi q_\phi(Z)[\log p_\phi(x^{(i)},Z) - \log q_\phi(Z)]dZ - \int \nabla_\phi q_\phi(Z)dZ \tag{58}$$

$$= \int \nabla_\phi q_\phi(Z)[\log p_\phi(x^{(i)},Z) - \log q_\phi(Z)]dZ \tag{59}$$

$$= \int q_\phi(Z)(\nabla_\phi \log q_\phi(Z))[\log p_\phi(x^{(i)},Z) - \log q_\phi(Z)]dZ \tag{60}$$

$$= E_{q(\phi)}[\nabla_\phi \log q_\phi(Z)\big(\log p_\phi(x^{(i)},Z) - \log q_\phi(Z)\big)] \tag{61}$$

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：

$$Z^l \sim q_\phi(Z) \tag{62}$$

$$E_{q(\phi)}[\nabla_\phi \log q_\phi(Z)\big(\log p_\phi(x^{(i)}, Z) - \log q_\phi(Z)\big)] \sim \frac{1}{L}\sum_{l=1}^{L} \nabla_\phi \log q_\phi(Z)\big(\log p_\phi(x^{(i)}, Z) - \log q_\phi(Z)\big) \tag{63}$$

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：

$$\nabla_\phi L(\phi) = \nabla_\phi E_{q(\phi)}[\log p_\phi(x^{(i)}, Z) - \log q_\phi(Z)] \tag{64}$$

上式来自于第六个等式。我们取：$Z = g_\phi(\varepsilon, x^{(i)}), \varepsilon \sim p(\varepsilon)$，这里$p(\varepsilon)$是人为给定的。所以后验概率为：$Z \sim q_\phi(Z|x^{(i)})$，通过重参数化，把随机性进行了转移，有$|q_\phi(Z|x^{(i)})dZ| = |p(\varepsilon)d\varepsilon|$。代入上面的梯度中：

$$\begin{aligned}
\nabla_\phi L(\phi) &= \nabla_\phi E_{q(\phi)}[\log p_\phi(x^{(i)}, Z) - \log q_\phi(Z)] && (65)\\
&= \nabla_\phi \int [\log p_\phi(x^{(i)}, Z) - \log q_\phi(Z)] q_\phi dZ && (66)\\
&= \nabla_\phi \int [\log p_\phi(x^{(i)}, Z) - \log q_\phi(Z)] p_\varepsilon d\varepsilon && (67)\\
&= E_{p(\varepsilon)}[\nabla_\phi [\log p_\phi(x^{(i)}, Z) - \log q_\phi(Z)]] && (68)\\
&= E_{p(\varepsilon)}[\nabla_Z [\log p_\phi(x^{(i)}, Z) - \log q_\phi(Z)] \nabla_\phi Z] && (69)\\
&= E_{p(\varepsilon)}[\nabla_Z [\log p_\phi(x^{(i)}, Z) - \log q_\phi(Z)] \nabla_\phi g_\phi(\varepsilon, x^{(i)})] && (70)
\end{aligned}$$

对上式进行蒙特卡洛采样，然后计算期望，得到梯度。接下来我们介绍蒙特卡罗方法。

4.2.3 随机近似——马尔可夫蒙特卡洛方法

MCMC 是一种随机的近似推断，其核心就是基于采样的随机近似方法蒙特卡洛方法。对于采样任务来说，有下面一些常用的场景：

1. 采样作为任务，用于生成新的样本；
2. 求和/求积分。

采样结束后，我们需要评价采样出来的样本点是不是好的样本集：

1. 样本趋向于高概率的区域；
2. 样本之间必须独立。

具体采样中，采样是一个困难的过程：

1. 无法采样得到归一化因子，即无法直接对概率$p(z) = \frac{1}{Z}\hat{p}(x)$ 采样，常常需要对CDF采样，但复杂的情况不行；
2. 如果归一化因子可以求得，但是对高维数据依然不能均匀采样（维度灾难），这是由于对p 维空间，总的状态空间是K^p 这么大，于是在这种情况下，直接采样也不行。

因此需要借助其他手段，如蒙特卡洛方法中的拒绝采样，重要性采样和MCMC。

蒙特卡洛方法旨在求得复杂概率分布下的期望值：

$$E_{z|x}[f(z)] = \int p(z|x)f(z)dz \simeq \frac{1}{N}\sum_{i=1}^{N} f(z_i) \tag{71}$$

也就是说，从概率分布中取N个点，从而近似计算这个积分。采样方法有：

1. 概率分布采样，首先求得概率密度的累积密度函数CDF，然后求得CDF的反函数，在0到1之间均匀采样，代入反函数，就得到了采样点。但是实际大部分概率分布不能得到CDF。

2. 拒绝采样（Rejection Sampling）：对于概率分布$p(z)$，引入简单的提议分布$q(z)$，使得$\forall z_i, Mq(z_i) \geq p(z_i)$，因为概率分布要求概率归一，所以一个概率分布不可能处处都比另一个大，所以我们用一个常数M，把先验分布$q(z)$进行放大，这样就满足了要求。我们先在$q(z)$ 中采样，定义接受率：$\alpha = \frac{p(x^{(i)})}{Mq(z^{(i)})} \leq 1$（可以看出，当两个分布非常接近时，接受率很大，反之很低）。算法描述为：

   (a) 取$z^{(i)} \sim q(z)$；

   (b) 在均匀分布$U(0,1)$中选取u；

   (c) 如果$u \leq \alpha$，则接受$z^{(i)}$，否则，拒绝这个值。

3. **重要性采样**（Importance Sampling）：直接对期望：$E_{p(z)}[f(z)]$进行采样。

$$E_{p(z)}[f(z)] = \int p(z)f(z)dz = \int \frac{p(z)}{q(z)}f(z)q(z)dz \simeq \frac{1}{N}\sum_{i=1}^{N} f(z_i)\frac{p(z_i)}{q(z_i)} \tag{72}$$

   于是采样在$q(z)$中采样，并通过权重计算和。重要值采样对于权重非常小的时候，效率非常低。重要性采样有一个变种Sampling-Importance-Resampling，这种方法，首先和上面一样进行采样，然后在采样出来的个样本中，重新采样，这个重新采样，使用每个样本点的权重作为概率分布进行采样。重要性采样能明显降低一般蒙特卡洛方法的采样次数，对于一个估算积分的数值解，如果想达到相同的标准差，需要的样本数量会从10^7降低到10^2的数量级。

接下来我们介绍一种采样方法，称为**马尔可夫链蒙特卡罗方法**（Markov Chain and Monte Carlo Method）。马尔可夫链是一个随机过程（研究过程是一个随机变量序列），它的时间和状态都是离散型随机变量，每次的观测值只依赖于上一次，与它相关的物理过程，最常见的是布朗运动（花粉在水面上无规则运动）。我们关注的主要是齐次的一阶马尔可夫链。马尔可夫链满足：

$$p(X_{t+1|X_1,X_2,\ldots,X_t}) = p(X_{t+1}|X_t) \tag{73}$$

这个公式可以写成状态转移矩阵的形式，矩阵元为：

$$p_{ij} = p(X_{t+1}|X_t = i) \tag{74}$$

状态转移矩阵表示了从一个状态，跳到另一个状态，发生的概率描述，我们有：

$$\pi_{t+1}(x^\star) = \int \pi_i(x)p_{x\to x^\star}dx \tag{75}$$

如果存在$\pi = (\pi(1), \pi(2), \cdots)$，$\sum_{i=1}^{+\infty}\pi(i) = 1$。这里的$\pi(i)$表示在$i$时刻时，随机变量对应的概率分布。那么我们称这样的随机序列为马尔可夫链X_t的平稳分布。平稳分布就是表示在某一个时刻后，分布不再改变。MCMC就是通过构建马尔可夫链概率序列，使其收敛到平稳分布$p(z)$。引入细致平衡：

$$\pi(x)p_{x\to x^\star} = \pi(x^\star)p_{x^\star\to x} \tag{76}$$

如果一个分布满足细致平衡，那么一定满足平稳分布（反之不成立）：

$$\int \pi(x)p_{x\to x^\star}dx = \int \pi(x^\star)p_{x^\star\to x}dx = \pi(x^\star) \tag{77}$$

细致平衡条件将平稳分布的序列和马尔可夫链的转移矩阵联系在一起了，通过转移矩阵可以不断生成样本点。假定随机取一个转移矩阵($Q = Q_{ij}$)，作为一个先验矩阵。我们有：

$$p(z)Q_{z\to z^\star}\alpha(z,z^\star) = p(z^\star)Q_{z^\star\to z}\alpha(z^\star,z) \tag{78}$$

取：

$$\alpha(z,z^\star) = \min\{1, \frac{p(z^\star Q_{z^{star}\to z})}{p(z)Q_{z\to z^\star}}\} \tag{79}$$

则：

$$p(z)Q_{z\to z^\star}\alpha(z,z^\star) = \min\{p(z)Q_{z\to z^\star}, p(z^\star)Q_{z^\star\to z}\} = p(z^\star)Q_{z^\star\to z}\alpha(z^\star,z) \tag{80}$$

于是，迭代就得到了序列，这个算法叫做Metropolis-Hastings算法：

1. 通过在0，1之间均匀分布取点u；
2. 生成$z^\star \sim Q(z^\star|z^{i-1})$；
3. 计算α值；
4. 如果$\alpha \geq u$，则$z^i = z^\star$，否则$z^i = z^{i-1}$。

这样取的样本就服从：

$$p(z) = \frac{\hat{p}(z)}{Z} \sim \hat{p}(z) \tag{81}$$

这里Z是归一化常数，一般称为配分函数。以上就是著名的MH采样算法。对于高维数据，即$p(Z) = p(z_1, z_2, ..., z_n)$，分布对应的数据点，维度很高，采样存在了很大的困难。我们对MH采样进行改进，提出**吉布斯采样方法**（Gibbs Sampling Method）。可以通过过固定被采样的维度其余的维度来简化采样过程$z_i \sim p(z_i|z_{i-1})$：

1. 给定初始值$z_1^0, z_2^0, \ldots$；
2. 在$t+1$时刻，采样$z_i^{t+1} \sim p(z_i|z_{i-1})$，从第一个维度依次采样。

可以计算Gibbs采样的接受率：

$$\frac{p(z^\star)Q_{z^\star\to z}}{p(z)Q_{z\to z^\star}} = \frac{p(z_i^\star|z_{-i}^\star)p(z_{-i}^\star)p(z_i|z_{-i}^\star)}{p(z_i|z_{-i})p(z_{-i})p(z^\star|z_{-i})} \tag{82}$$

对于每个Gibbs采样步骤都有$z_{-i} = z_{-i}^\star$，这是由于每个维度i采样的时候，其余的参量保持不变。这里我们直接认为Q就是p，把任意选择的Q就直接选为当前的概率分布。所以上式为1。于是Gibbs 采样过程中，相当于找到了一个步骤，使得所有的接受率为1。

最后我们在讨论一下平稳分布，定义**随机矩阵**（Random Matrix）：

$$Q = \begin{bmatrix} Q_{11} & Q_{12} & \cdots & Q_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{K1} & Q_{K2} & \cdots & Q_{KK} \end{bmatrix} \tag{83}$$

要求这个矩阵每一行或者每一列的和都是1。随机矩阵的特征值都小于等于1。我们说，这个Q是我们构造出来的，不唯一。我们通过这个Q还有已知分布$q(x)$，通过采样逼近$p(x)$。假设只有一个特征值为$\lambda_i = 1$。于是在马尔可夫过程中：

$$q^{t+1}(x = j) = \sum_{i=1}^{K} q^t(x = i)Q_{ij} \tag{84}$$

$$\Rightarrow q^{(t+1)} = q^t \cdot Q = q^1 \cdot Q^t \tag{85}$$

根据求和的写法，这里的q是一个行向量。于是，对Q 做对角化，有：

$$q^{t+1} = q^1 A \Lambda^t A^{-1} \tag{86}$$

如果m足够大，那么$\Lambda^m = diag(0, 0, \ldots, 1, \ldots, 0)$，则：$q^{m+1} = q^m$，则趋于平稳分布了。马尔可夫链可能具有平稳分布的性质，所以我们可以构建马尔可夫链使其平稳分布收敛于需要的概率分布（设计转移矩阵）。

在采样过程中，需要经历一定的时间（燃烧期/混合时间）才能达到平稳分布。但是MCMC方法有一些问题：

1. 无法判断是否已经收敛；
2. 燃烧期过长（维度太大，并且维度之间有关，可能无法采样到某些维度），例如在GMM中，可能无法采样到某些峰。于是在一些模型中，需要对隐变量之间的关系作出约束，如RBM假设隐变量之间无关。
3. 样本之间一定是有相关性的，如果每个时刻都取一个点，那么每个样本一定和前一个相关，这可以通过间隔一段时间采样。

说完了缺点，最后谈一下采样的初衷与评价采样得到的样本好坏。采样的初衷，第一点，采样本身就是常见的一种任务；第二点，就是求和或者求积分（期望）。那么什么是好的样本呢？

- 样本趋向于高概率区域；
- 样本之间独立。

但是我们说，采样本身是困难的，其一是配分函数的关系，其二是随机变量的维度太高。

References