10 近似推断

在概率模型的应用中，一个中心任务是在给定观测（可见）数据变量X的条件下，计算潜在变量Z的后验概率分布$p(Z|X)$，以及计算关于这个概率分布的期望。
对于实际应用中的许多模型来说，计算后验概率分布或者计算关于这个后验概率分布的期望是不可行的。这可能是由于潜在空间的维度太高，以至于无法直接计算，或者由于后验概率分布的形式特别复杂，从而期望无法解析地计算。在连续变量的情形中，需要求解的积分可能没有解析解，而空间的维度和被积函数的复杂度可能使得数值积分变得不可行。对于离散变量，求边缘概率的过程涉及到对隐含变量的所有可能的配置进行求和。这个过程虽然原则上总是可以计算的，但是我们在实际应用中经常发现，隐含状态的数量可能有指数多个，从而精确的计算所需的代价过高。
根据近似方法依赖于随机近似还是确定近似，方法大体分为两大类。

10.1 变分推断

变分方法本质上没有任何近似的东西，但是它们通常会被用于寻找近似解。寻找近似解的过程可以这样完成：限制需要最优化算法搜索的函数的范围，例如只考虑二次函数，或者考虑由固定的基函数线性组合而成的函数，其中只有线性组合的系数可以发生变化。
假设我们有一个纯粹的贝叶斯模型，其中每个参数都有一个先验概率分布。这个模型也可以有潜在变量以及参数，我们会把所有潜在变量和参数组成的集合记作Z。类似地，我们会把所有观测变量的集合记作X.对于N个独立同分布的数据，其中$X = \{x_1, \cdots, x_N\}, Z = \{z_1, \cdots, z_N\}$.概率模型确定了联合概率分布$p(X, Z)$，我们的目标是找到对后验概率分布$p(Z|X)$以及模型证据$p(X)$的近似。
可以将对数边缘概率分解，得到：

$$\ln p(x) = \mathcal{L}(q) + \mathbf{KL}(q||p)$$

$$\mathcal{L}(q) = \int q(Z) \ln \left\{ \frac{p(X,Z)}{q(Z)} \right\} \mathrm{d}Z$$

$$\mathbf{KL}(q||p) = -\int q(Z) \ln \left\{ \frac{p(Z|X)}{q(Z)} \right\} \mathrm{d}Z$$

与EM算法相比，参数向量θ不再出现，被整合到Z当中。与之前一样，可以通过关于$q(Z)$优化下界$\mathcal{L}(q)$使之达到最大值，这等价于最小化$\mathbf{KL}$散度。若可任意选择$q(Z)$，则下界最大值出现在$\mathbf{KL}$散度等于0的时候，此时$q(Z)$等于后验概率分布$p(Z|X)$.所以对于$q(Z)$的受限类型，在该范围内找到使得$\mathbf{KL}$散度最小的概率分布。所以要限制$q(Z)$使得它易于处理，也要使得这个范围足够大，充分灵活，使得它可以对真实后验概率分布足够好的近似。
需要强调的是，施加限制条件的唯一目的是为了计算方便，并且在这个限制条件下，我们应该使用尽可能丰富的近似概率分布。特别地，对于高度灵活的概率分布来说，没有“过拟合”现象。使用灵活的近似仅仅使得我们更好地近似真实的后验概率分布。

10.1.1 分解概率分布

我们限制概率分布$q(Z)$的范围。假设我们将Z的元素划分成若干个互不相交的组，记作$Z_i, i = 1, \cdots, M$,我们假定q分布关于这些分组可以分解：

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

我们希望对于$\mathcal{L}(q)$关于所有概率分布$q_i(Z_i)$进行自由形式的（变分）最优化，将$q_i(Z_i)$记作q_i：

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(X,Z) - \sum_i \ln q_i \right\} \mathrm{d}Z \\
&= \int q_j \left\{ \int \ln p(X,Z) \prod_{i \neq j} q_i \mathrm{d}Z_i \right\} \mathrm{d}Z_j - \int q_j \ln q_j \mathrm{d}Z_j + \text{const} \\
&= \int q_j \ln \tilde{p}(X, Z_j) \mathrm{d}Z_j - \int q_j \ln q_j \mathrm{d}Z_j + \text{const} \\
&= \mathbf{KL}(q_j || \tilde{p}(X, Z_j)) + \text{const}
\end{aligned}$$

其中：

$$\ln \tilde{p}(X, Z_j) = \mathbb{E}_{i \neq j}[\ln p(X, Z)] + \text{const}$$
$$\mathbb{E}_{i \neq j}[\ln p(X, Z)] = \int \ln p(X, Z) \prod_{i \neq j} q_j \mathrm{d}Z_i$$

可以得到最优解$q_j^*(Z_j)$的一般表达式为：

$$\ln q_j^*(Z_j) = \mathbb{E}_{i \neq j}[\ln p(X, Z)] + \text{const}$$

这个解表明，为了得到因子q_j 的最优解的对数，我们只需考虑所有隐含变量和可见变量上的联合概率分布的对数，然后关于所有其他的因子$\{q_i\}$取期望即可，其中$i \neq j$.
其中可加性常数可以通过对概率分布 $q_j^*(Z_j)$进行归一化的方式得到：

$$q_j^*(Z_j) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(X, Z)])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(X, Z)]) \mathrm{d}Z_j}$$

这些方程并没有给出一个显式的解，因为最优化$q_j^*(Z_j)$的公式（10.9）的右侧表达式依赖于关于其他的因子$q_i(Z_i), i \neq j$计算的期望。于是，我们会用下面的方式寻找出一个相容的解：首先，恰当地初始化所有的因子$q_i(Z_i)$,然后在各个因子上进行循环，每一轮用一个修正后的估计来替换当前因子。

10.1.2 分解近似的性质

变分推断的方法基于的是真实后验概率分布的分解近似，当我们想要使用分解的高斯分布近似一个高斯分布的问题得时候，考虑两个相关变量$z = (z_1, z_2)$上的高斯分布$p(z) = \mathcal{N}(z|\mu, \Lambda^{-1})$,其中均值和精度：

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

由于精度矩阵的对称性，可以得到：$\Lambda_{12} = \Lambda_{21}$,我们希望使用分解的高斯分布$q(z) = q_1(z_1)q_2(z_2)$,可以得到：

$$\begin{aligned} \ln q_1^*(z_1) &= \mathbb{E}_{z_2}[\ln p(z)] + \text{const} \\ &= \mathbb{E}_{z_2}[-\frac{1}{2}(z_1 - \mu_1)^2 \Lambda_{11} - (z_1 - \mu_1)\Lambda_{12}(z_2 - \mu_2)] + \text{const} \\ &= -\frac{1}{2} z_1^2 \Lambda_{11} + z_1 \mu_1 \Lambda_{11} - z_1 \Lambda_{12}(\mathbb{E}[z_2] - \mu_2) + \text{const} \end{aligned}$$

所以可以看到$q_1^*(z_1)$是一个高斯分布，这里我们并未假设$q(z_i)$是高斯分布，但是通过对所有可能分布的KL散度变分最优化推导得到了该结论。
使用配方法，可以得到：

$$q_q^*(z_1) = \mathcal{N}(z_1|m_1, \Lambda_{11}^{-1})$$

其中的:

$$m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2] - \mu_2)$$

所以$q_2^*(z_2)$也是一个高斯分布:

$$q_2^*(z_2) = \mathcal{N}(z_2|m_2|\Lambda_{22}^{-1})$$

其中:

$$m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbb{E}[z_1] - \mu_1)$$

所以这些解是相互耦合的，所以可将变分解看作是重估计方程，然后在变量之间循环，更新这些解，直到满足收敛准则为止。
该问题可以找到解析解，如果$\mathbb{E}[z_1] = \mu_1, \mathbb{E}[z_2] = \mu_2$，只要概率分布非奇异，则该解是唯一解，结果如10.2（a），均值被严重低估:分解变分近似对于后验概率分布的近似倾向于过于紧凑。

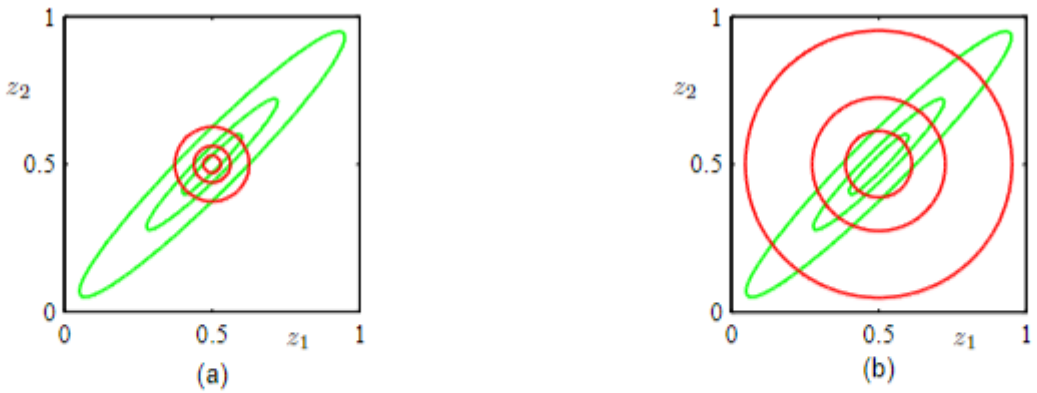

图 10.2: 两种形式的KL散度的对比。绿色轮廓线对应于两个变量z_1和z_2上的相关高斯分布$p(\boldsymbol{z})$的1、2、3个标准差的位置，红色轮廓线表示相同变量上的近似分布$q(\boldsymbol{z})$的同样位置。近似分布$q(\boldsymbol{z})$由两个独立的一元高斯分布的乘积给出。(a)图中，参数通过最小化Kullback-Leibler散度$\mathrm{KL}(q \parallel p)$的方式获得。(b)图中，参数通过最小化相反的Kullback-Leibler散度$\mathrm{KL}(p \parallel q)$的方式获得。

当我们最小化相反的KL散度$\mathbf{KL}(p||q)$,KL散度可以写作:

$$\mathbf{KL}(p||q) = -\int p(Z)\left[\sum_{i=1}^{M} \ln q_i(Z_i)\right] \mathrm{d}Z + \text{const}$$

使用拉格朗日乘数法，可以得到:

$$q_j^*(Z_j) = \int p(Z) \prod_{i \neq j} \mathrm{d}Z_i = p(Z_j)$$

在这种情况下，我们看到$q_j(Z_j)$的最优解等于对应的边缘概率分布$p(Z)$。注意，这是一个解析解，不需要迭代。该结果对均值的近似是正确的，但是它把相当多的概率质量放到了实际上具有很低的概率的变量空间区域中。
如果我们考虑用一个单峰分布近似多峰分布的问题，基于最小化KL(q ‖ p)的变分方法倾向于找到这些峰值中的一个。相反，如果我们最小化KL(p ‖ q)，那么得到的近似会在所有的均值上取平均。
这两种KL散度都是散度的α家族的成员，定义为：

$$D_\alpha(p||q) = \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{\frac{1+\alpha}{2}} q(x)^{\frac{1-\alpha}{2}} \mathrm{d}x\right)$$

$\mathbf{KL}(p||q)$对应于$\alpha \to 1$,$\mathbf{KL}(q||p)$对应于$\alpha \to -1$.对于所有的α，都有$D_\alpha(p||q) \geq 0$,当且仅当$p(x) = q(x)$时等号成立。
假设$p(x)$是固定的分布，关于某个概率分布$q(x)$的集合最小化$D_\alpha(p||q)$,则对于$\alpha \leq -1$时，散度是零强制的，即对于任意使得$p(x) = 0$的x值，都有$q(x) = 0$,通常q(x)会低估p(x)的支持，因此倾向于寻找具有最大质量的峰值。而对于$\alpha \geq 1$的情况，散度是零避免的，即对于任何使得$p(x) > 0$的x值，都有$q(x) > 0$.

10.1.4 模型比较

除了在隐含变量Z上进行推断之外，我们可能还希望对比一组候选模型。索引为m的模型的先验概率分布为$p(m)$。这样，我们的目标是近似后验概率分布$p(m|X)$，其中X是观测数据。
因为不同的模型可能具有不同的结构，并且隐含变量Z的维度实际上可能不同。因此我们不能简单地考虑考虑分解近似$q(Z)q(m)$，而是必须意识到Z的后验概率分布必须以m为条件，所以我们必须考虑$q(Z,m) = q(Z|m)q(m)$。我们已经可以验证下面的基于变分概率分布的分解方式:

$$\ln p(X) = \mathcal{L} - \sum_m \sum_Z q(Z|m)q(m) \ln \frac{p(Z,m|X)}{q(Z|m)q(m)}$$

其中$\mathcal{L}$是$\ln p(X)$的下界，形式为：

$$\mathcal{L} = \sum_m \sum_Z q(Z|m)q(m) \ln \frac{p(Z,X,m)}{q(Z|m)q(m)}$$

我们可以使用拉格朗日乘数法关于概率分布$q(m)$最大化$\mathcal{L}$，结果为:

$$q(m) \propto p(m) \exp\{\mathcal{L}_m\}$$

其中：

$$\mathcal{L}_m = \sum_Z q(Z|m) \ln \frac{p(Z,X|m)}{q(Z|m)}$$

我们关于$q(Z|m)$最大化$\mathcal{L}_m$，可以发现对于不同的m，解是耦合的。

10.4 指数族分布

对于许多模型来说，完整数据是服从指数族分布的，但是边缘概率分布不是服从指数族分布的。
我们可以将潜在变量与参数区分开。潜在变量Z是分散的，数量随数据集规模的增大而增大。参数θ是分散的，它的数量是固定的，与数据集的规模无关。
现在假设观测变量和隐含变量的联合概率分布为指数族分布的成员，参数为自然参数η，即:

$$p(X,Z|\eta)=\prod_{n=1}^{N}h(x_n,z_n)g(\eta)\exp(\eta^Tu(x_n,z_n))$$

对于η的共轭先验:

$$p(\eta|\nu_0,\chi_0)=f(\nu_0,\chi_0)g(\eta)^{\nu_0}\exp(\nu_0\eta^T\chi_0)$$

可以分解$q(Z,\eta)=q(Z)q(\eta)$,得到:

$$\begin{aligned}\ln q^*(Z)&=\mathbb{E}_\eta[\ln p(X,Z|\eta)]+\text{const}\\&=\sum_{n=1}^{N}\{\ln h(x_n,z_n)+\mathbb{E}[\eta^T]u(x_n,z_n)\}+\text{const}\end{aligned}$$

因此我们看到它可以分解为一组相互独立的项的加和，每个n都对应于一项，因此$q^*(Z)$的解可以在n上进行分解，即$q^*(Z)=\prod_n q^*(z_n)$。这是诱导分解的一个例子。两侧取指数，我们有:

$$q^*(z_n)=h(x_n,z_n)g(\mathbb{E}[\eta])\exp\{\mathbb{E}[\eta^T]u(x_n,z_n)\}$$

对于参数上的变分分布，有:

$$\begin{aligned}\ln q^*(\eta)&=\ln p(\eta|\nu_0,\chi_0)+\mathbb{E}_Z[\ln p(X,Z|\eta)]+\text{const}\\&=\nu_0\ln g(\eta)+\nu_0\eta^T\chi_0+\sum_{n=1}^{N}\{\ln g(\eta)+\eta^T\mathbb{E}_{z_n}[u(x_n,z_N)]\}+\text{const}\end{aligned}$$

两侧取指数，可以得到:

$$q^*(\eta)=f(\nu_N,\chi_N)g(\eta)^{\nu_N}\exp(\nu_N\eta^T\chi_N)$$

其中:

$$\nu_N = \nu_0 + N$$

$$\nu_N \chi_N = \nu_0 \chi_0 + \sum_{n=1}^{N} \mathbb{E}_{z_n}[u(x_n, z_n)]$$

$q^*(\eta)$与$q^*(z_n)$得解是相互耦合的，所以可以使用二阶段的迭代方法求解。在变分E步骤中，我们使用潜在变量上的当前后验概率分布$q(z_n)$计算充分统计量的期望$\mathbb{E}[u(x_n, z_n)]$，并且使用这个结果计算参数上的修正的后验概率分布$q(\eta)$。然后，在接下来的变分M步骤中，我们使用修正后的参数后验概率分布寻找自然参数的期望$\mathbb{E}[\eta^T]$，它给出了潜在变量上的修正后的变分分布。

10.4.1 变分信息传递

对应于有向图的联合概率分布可以写成下面的分解形式:

$$p(x) = \prod_i p(x_i|\mathrm{pa}_i)$$

其中x_i表示与节点i关联的变量，pa_i表示与节点i相对应的父节点集合。这里x_i可以是一个潜在变量，也可以是属于观测变量集合。现在对于变分近似，假定概率分布$q(z)$可以对x_i分解，即:

$$q(x) = \prod_i q_i(x_i)$$

于是根据前面的结果可以得到:

$$\ln q_j^*(x_j) = \mathbb{E}_{i \neq j}\left[\sum_i \ln p(x_i|\mathrm{pa}_i)\right] + \mathrm{const}$$

唯一依赖于x_j的项是由$p(x_j|\mathrm{pa}_j)$给出的x_j的条件概率分布以及任何在条件集合中有x_j的条件概率分布。所以这些条件概率分布对应于节点j的子节点，所以也依赖于子节点的同父节点。所以$q_j^*(x_j)$所以来的所有节点组成的集合对应于节点x_i的马尔科夫毯。因此，在变分后验概率分布中的更新因子表示图上的一个局部计算。

如果我们现在确定模型的形式，其中所有的条件概率分布都有一个共轭-指数族的结构，那么变分推断的过程可以被转化为局部信息传递算法。特别地，对于一个特定的结点来说，一旦它接收到了来自所有的父结点和所有的子结点的信息，那么与这个结点相关联的概率分布就可以被更新。这反过来需要子结点从它们的同父结点已经接收完毕信息。下界的计算也可以得到简化，因为许多必要的值已经作为信息传递框架的一部分计算完毕。分布的信息传递形式有很好的缩放性质，对于大的网络很合适。

10.5 局部变分方法

另一种“局部”的方法涉及到寻找模型中的单独的变量或者变量组上定义的函数的界限。例如，我们可能寻找条件概率分布$p(y|x)$的界限，这个条件概率本身仅仅是一个由有向图模型描述的更大的概率模型中的一个因子。引入界限的目的显然是简化最终得到的概率分布。这个局部近似可以应用于多个变量，直到得到一个可以处理的近似。对数函数的凸函数性质在求解全局变分方法的下界时起到关键作用，我们将凸函数定义为每条弦都在函数上方的函数，同样对于凸函数，只需要将最大化转为最小化，下界转为上界即可。

比如对于函数$f(x) = \exp(-x)$是x的一个凸函数，我们使用x的线性函数逼近：如果这个线性函数对应于一条切线，则它是$f(x)$的下界。我们将得到具体的x处$y(x)$的切线，例如在$x = \xi$处，使用一阶泰勒展开：

$$y(x) = f(\xi) + f'(\xi)(x - \xi)$$

所以$y(x) \leq f(x)$等号只在$x = \xi$处成立。所以得到：

$$y(x) = \exp(-\xi) - \exp(-\xi)(x - \xi)$$

这是一个以ξ为参数的线性函数，我们定义$\eta = -\exp(-\xi)$,所以：

$$y(x, \eta) = \eta x - \eta + \eta \ln(-\eta)$$

不同的η对应于不同的切线，所有的切线都是函数的下界，所以$f(x) \geq y(x, \eta)$所以得到：

$$f(x) = \max_{\eta} \{\eta x - \eta + \eta \ln(-\eta)\}$$

我们可以使用凸对偶（convex duality）的框架更加一般地形式化描述这种方法：

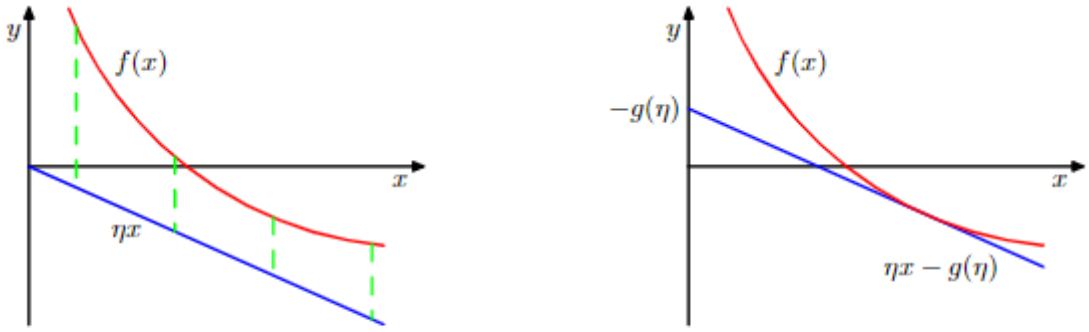

图 10.11: 在左图中，红色曲线给出了一个凸函数$f(x)$，蓝色曲线表示线性函数ηx，它是$f(x)$的一个下界，因为对于所有的x都有$f(x) > \eta x$。对于给定的斜率η的值，具有相同斜率的切线的接触点可以通过关于x最小化差距$f(x) - \eta x$的方式得到，差距用绿色虚线表示。这定义了对偶函数$g(\eta)$，它对应于具有斜率η的切线的截距（的负值）。

$$\begin{aligned} g(\eta) &= -\min_{x} \{f(x) - \eta x\} \\ &= \max_{x} \{\eta x - f(x)\} \end{aligned}$$

同样地，可以得到：

$$f(x) = \max_{\eta}\{\eta x - g(\eta)\}$$

所以对于$f(x) = \exp(-x)$,可以得到$g(\eta) = \eta - \eta \ln -\eta$.
对于凹函数，可以得到上界：

$$f(x) = \min_{\eta}\{\eta x - g(\eta)\}$$
$$g(\eta) = \min_{x}\{\eta x - f(x)\}$$

如果感兴趣的函数不是凸函数（或者凹函数），那么我们不能直接应用这种方法得到上述界限。然而，我们可以首先寻找函数或者参数的一个可逆变换，这个变换将函数或者参数变换为一个凸函数的形式。然后，我们计算共轭函数，之后变换回原始的变量。
对于logistic sigmod函数：

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

我们将其取对数后就是一个凹函数，其共轭函数为：

$$g(\eta) = -\eta \ln(-\eta) - (1-\eta)\ln(1-\eta)$$

他是一个二元变量的熵，变量取值为1的概率为η，可以得到对数sigmod函数的上界

$$\ln \sigma(x) \leq \eta x - g(\eta)$$

然后取指数，可以得到logistic sigmod函数的一个上界：

$$\sigma(x) \leq \exp(\eta x - g(\eta))$$

我们也可以得到sigmoid函数的下界，下界的函数形式是高斯形式。对输入变量和函数本身都进行变换。首先，我们取logistic函数的对数，然后将其分解，即

$$\begin{aligned}\ln \sigma(x) &= -\ln(1+e^{-x}) \\ &= -\ln e^{-\frac{x}{2}}(e^{(\frac{x}{2})+e^{-\frac{x}{2}}}) \\ &= \frac{x}{2} - \ln(e^{(\frac{x}{2})+e^{-\frac{x}{2}}})\end{aligned}$$

我们现在注意到，函数$f(x) = -\ln\left(e^{\frac{x}{2}} + e^{-\frac{x}{2}}\right)$是变量$x^2$的一个凸函数，这一点可以通过取二阶导数的方式证明。这产生了$f(x)$的下界，它是x^2的一个线性函数，它的共轭函数为

$$g(\eta) = \max_{x^2}\left\{\eta x^2 - f\left(\sqrt{x^2}\right)\right\} \tag{10.139}$$

根据驻点的条件可得

$$0 = \eta - \frac{\mathrm{d}x}{\mathrm{d}x^2}\frac{\mathrm{d}}{\mathrm{d}x}f(x) = \eta + \frac{1}{4x}\tanh\left(\frac{x}{2}\right) \tag{10.140}$$

如果我们将这个值记作x，对应于在这个特定的η值下，函数与切线的接触点，记作η，那么我们有

$$\eta = -\frac{1}{4\xi}\tanh\left(\frac{\xi}{2}\right) = -\frac{1}{2\xi}\left[\sigma(\xi) - \frac{1}{2}\right] = -\lambda(\xi) \tag{10.141}$$

其中，我们定义了$\lambda = -\eta$，保持与Jaakkola and Jordan（2000）的相容性。我们不把λ看成变分参数，相反，我们可以令ξ为变分参数，因为这会产生共轭函数的更简单的表达式，它的形式为

$$g(\lambda(\xi)) = -\lambda(\xi)\xi^2 - f(\xi) = -\lambda(\xi)\xi^2 + \ln\left(e^{\frac{\xi}{2}} + e^{-\frac{\xi}{2}}\right) \tag{10.142}$$

这里，$f(x)$的界限可以写成

$$f(x) \geqslant -\lambda(\xi)x^2 - g(\lambda(\xi)) = -\lambda(\xi)x^2 - \lambda(\xi)\xi^2 - \ln\left(e^{\frac{\xi}{2}} + e^{-\frac{\xi}{2}}\right) \tag{10.143}$$

sigmoid函数的界限就变成了

$$\sigma(x) \geqslant \sigma(\xi)\exp\left\{\frac{x-\xi}{2} - \lambda(\xi)(x^2 - \xi^2)\right\} \tag{10.144}$$

其中$\lambda(\xi)$的定义为（10.141）。这个界限如图10.12的右图所示。我们看到，界限的函数形式是x的二次函数的指数形式。当我们寻找通过logistic sigmoid函数定义的后验概率分布的高斯表示时，这个界限的形式很有用。

logistic sigmoid函数在二值变量上的概率模型中经常出现，因为它是将log odds函数转换为后验概率分布的函数。对于多类分布，对应的变换由softmax函数给出。不幸的是，这里推导出logistic sigmoid函数的下界不能直接扩展到softmax函数。Gibbs（1997）提出了一种构建高斯分布的方法，这个高斯分布被猜想为是一个界限（虽然没有给出严格的证明），这可以用于将局部变分方法应用到多分类问题。

我们会在10.6.1节看到局部变分界限的一个例子。然而，现阶段从一般的角度考虑这些界限如何被使用是很有意义的。假设我们想计算一个形式如下的积分

$$I = \int \sigma(a)p(a)\,\mathrm{d}a \tag{10.145}$$

其中$\sigma(a)$是一个logistic sigmoid函数，$p(a)$是一个高斯概率密度。当我们项计算贝叶斯模型中的预测分布时，这种积分会经常出现，此时$p(a)$表示一个后验参数分布。由于积分是无法直接计算的，因此我们使用变分界限（10.144），我们将它写成$\sigma(a) \geq f(a,\xi)$，其中ξ是一个变分参数。积分现在变成了两个指数-二次函数的乘积，因此可以解析地求出积分，给出I的界限

$$I \geq \int f(a,\xi)p(a)\,\mathrm{d}a = F(\xi) \tag{10.146}$$

我们可以自由地选择变分参数ξ，这里我们选择最大化函数$F(\xi)$的值ξ^*。得到的值$F(\xi^*)$表示在所有的界限中最紧致的界限，可以用来近似I。然而，这个最优化的界通常不是精确的。虽然logistic sigmoid函数的界限$\sigma(a) \geq f(a,\xi)$可以被精确地最优化，但是ξ的最优选择依赖于a的值，从而界限只对一个a的值是精确的。由于$F(\xi)$可以通过对a的所有值上进行积分的方式得到，因此ξ^*的值表示一个折中，权值为概率分布$p(a)$。

10.7 期望传播

先考虑关于$q(\boldsymbol{z})$最小化$\mathrm{KL}(p \,\|\, q)$的问题，其中$p(\boldsymbol{z})$是一个固定的概率分布，$q(\boldsymbol{z})$是指数族分布的一个成员，因此根据公式（2.194），可以写成

$$q(\boldsymbol{z}) = h(\boldsymbol{z})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{z})\} \tag{10.184}$$

作为$\boldsymbol{\eta}$的一个函数，Kullback-Leibler散度变成了

$$\mathrm{KL}(p \,\|\, q) = -\ln g(\boldsymbol{\eta}) - \boldsymbol{\eta}^T\mathbb{E}_{p(\boldsymbol{z})}[\boldsymbol{u}(\boldsymbol{z})] + \text{常数} \tag{10.185}$$

其中常数项与自然参数$\boldsymbol{\eta}$无关。我们可以通过令关于$\boldsymbol{\eta}$的梯度等于零的方式，在这个概率分布族中最小化$\mathrm{KL}(p \,\|\, q)$，结果为

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}_{p(\boldsymbol{z})}[\boldsymbol{u}(\boldsymbol{z})] \tag{10.186}$$

然而，我们已经看到，在公式（2.226）中，$\ln g(\boldsymbol{\eta})$的负梯度有概率分布$q(\boldsymbol{z})$下$\boldsymbol{u}(\boldsymbol{z})$的期望给定。令这两个结果相等，我们有

$$\mathbb{E}_{q(\boldsymbol{z})}[\boldsymbol{u}(\boldsymbol{z})] = \mathbb{E}_{p(\boldsymbol{z})}[\boldsymbol{u}(\boldsymbol{z})] \tag{10.187}$$

我们看到，最优解仅仅对应于将充分统计量的期望进行匹配。因此，例如，如果$q(\boldsymbol{z})$是一个高斯分布$\mathcal{N}(\boldsymbol{z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$，那么我们通过令$q(\boldsymbol{z})$的均值$\boldsymbol{\mu}$等于分布$p(\boldsymbol{z})$的均值并且令协方差$\boldsymbol{\Sigma}$等于$p(\boldsymbol{z})$的协方差，即可最小化Kullback-Leibler散度。这有时被称为矩匹配（moment matching）。图10.3(a)给出了这个的一个例子。

现在，让我们利用这个结果，得到近似推断的一个实用的算法。对于许多概率模型来说，数据$\mathcal{D}$和隐含变量（包括参数）$\boldsymbol{\theta}$的联合概率分布由一组因子的乘积组成，形式为

$$p(\mathcal{D},\boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta}) \tag{10.188}$$

这个结果可能由独立同分布的数据的模型产生，其中对于每个数据点$\boldsymbol{x}_n$，都有一个因子$f_n(\boldsymbol{\theta}) = p(\boldsymbol{x}_n \mid \boldsymbol{\theta})$，且因子$f_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$对应于先验概率分布。更一般地，它也适用于任何由有向图定义的模型，其中每个因子是一个条件概率分布，对应于一个结点。也适用于无向图，其中每个因子是一个团块势函数。我们感兴趣的是计算后验概率分布$p(\boldsymbol{\theta} \mid \mathcal{D})$用于进行预测，以及计算模型证据$p(\mathcal{D})$用于进行模型比较。根据公式（10.188），后验概率分布为

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{1}{p(\mathcal{D})} \prod_i f_i(\boldsymbol{\theta}) \tag{10.189}$$

模型证据为

$$p(\mathcal{D}) = \int \prod_i f_i(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \tag{10.190}$$

期望传播基于后验概率分布的近似，这个近似也由一组因子的乘积给出，即

$$q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_i \tilde{f}_i(\boldsymbol{\theta}) \tag{10.191}$$

其中，近似中的每个因子$\tilde{f}_i(\boldsymbol{\theta})$对应于真实后验概率分布（10.189）中的一个因子$f_i(\boldsymbol{\theta})$，因子$\frac{1}{Z}$是归一化常数，用来确保公式（10.191）的左侧的积分等于1。为了得到一个实用的算法，我们需要对因子$\tilde{f}_i(\boldsymbol{\theta})$进行一定的限制，特别地，我们会假定因子来自指数族分布。于是，因子的乘积也是指数族分布，因此可以用充分统计量的有限集合来描述。例如，如果每个$\tilde{f}_i(\boldsymbol{\theta})$是一个高斯分布，那么整体的近似$q(\boldsymbol{\theta})$也是高斯分布。

理想情况下，我们通过最小化真实后验概率分布与近似分布之间的Kullback-Leibler散度的方式来确定$\tilde{f}_i(\boldsymbol{\theta})$，这个散度为

$$\mathrm{KL}(p \,\|\, q) = \mathrm{KL}\left(\frac{1}{p(\mathcal{D})} \prod_i f_i(\boldsymbol{\theta}) \,\middle\|\, \frac{1}{Z} \prod_i \tilde{f}_i(\boldsymbol{\theta}) \right) \tag{10.192}$$

期望传播通过在所有剩余因子的环境中对每个因子进行优化，从而取得了一个效果好得多的近似。首先，这种方法初始化因子$\tilde{f}_i(\boldsymbol{\theta})$，然后在因子之间进行循环，每次优化一个因子。这种方法的思想类似于之前讨论的变分贝叶斯框架的因子更近过程。假设我们希望优化因子$\tilde{f}_j(\boldsymbol{\theta})$。首先，我们将这个因子从乘积中移除，得到$\prod_{i \neq j} \tilde{f}_i(\boldsymbol{\theta})$。从概念上讲，我们要确定因子$\tilde{f}_j(\boldsymbol{\theta})$的一个修正形式，使得乘积

$$q^{\text{新}}(\boldsymbol{\theta}) \propto \tilde{f}_j(\boldsymbol{\theta}) \prod_{i \neq j} \tilde{f}_i(\boldsymbol{\theta}) \tag{10.193}$$

尽可能地接近

$$f_j(\boldsymbol{\theta}) \prod_{i \neq j} \tilde{f}_i(\boldsymbol{\theta}) \tag{10.194}$$

其中我们保持所有$i \neq j$的因子$\tilde{f}_i(\boldsymbol{\theta})$固定。这保证了近似在由剩余的因子定义的后验概率较高的区域最精确。后面，当我们将EP应用于“聚类问题”的时候，我们会看到这种效果的一个例子。为了完成这一点，我们首先从当前的对后验概率的近似中移除因子$\tilde{f}_j(\boldsymbol{\theta})$，方法是定义下面的未归一化的分布

$$q^{\setminus j}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\tilde{f}_j(\boldsymbol{\theta})} \tag{10.195}$$

其中我们保持所有$i \neq j$的因子$\tilde{f}_i(\boldsymbol{\theta})$固定。这保证了近似在由剩余的因子定义的后验概率较高的区域最精确。后面，当我们将EP应用于"聚类问题"的时候，我们会看到这种效果的一个例子。为了完成这一点，我们首先从当前的对后验概率的近似中移除因子$\tilde{f}_j(\boldsymbol{\theta})$，方法是定义下面的未归一化的分布

$$q^{\backslash j}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\tilde{f}_j(\boldsymbol{\theta})} \tag{10.195}$$

注意，我们反过来从$i \neq j$的因子的乘积中求出$q^{\backslash j}(\boldsymbol{\theta})$，虽然在实际应用中，除法通常更容易。它现在与因子$f_j(\boldsymbol{\theta})$结合，得到概率分布

$$\frac{1}{Z_j} f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \tag{10.196}$$

其中Z_j是归一化常数，形式为

$$Z_j = \int f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \tag{10.197}$$

我们现在通过最小化Kullback-Leibler散度

$$\mathrm{KL}\left(\frac{f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta})}{Z_j} \middle\| q^{\text{新}}(\boldsymbol{\theta}) \right) \tag{10.198}$$

来确定一个修正的因子$\tilde{f}_j(\boldsymbol{\theta})$。这很容易求解，因为近似分布$q^{\text{新}}(\boldsymbol{\theta})$来自指数族分布，因此我们可以使用结果（10.187），这个公式告诉我们，参数$q^{\text{新}}(\boldsymbol{\theta})$可以通过匹配公式（10.196）的对应矩的充分统计量的期望的方式获得。我们会假设这是一个可以计算的操作。例如，如果我们将$q(\boldsymbol{\theta})$选择为高斯概率分布$\mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$，那么$\boldsymbol{\mu}$被设置为（未归一化的）分布$f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta})$的均值，$\boldsymbol{\Sigma}$被设置为它的方差。更一般地，得到指数族分布的任意成员的所需的分布是很容易的，只要它能够被归一化即可，因为充分统计量的期望可以与归一化系数的导数相关联，正如公式（2.226）所述。图10.14说明了EP近似的过程。

根据公式（10.193），我们看到修正的因子$\tilde{f}_j(\boldsymbol{\theta})$可以按照下面的方法得到：取$q^{\text{新}}(\boldsymbol{\theta})$，然后除以剩余的因子，即

$$\tilde{f}_j(\boldsymbol{\theta}) = K \frac{q^{\text{新}}(\boldsymbol{\theta})}{q^{\backslash j}(\boldsymbol{\theta})} \tag{10.199}$$

其中我们使用了公式（10.195）。系数K通过下面的方式确定：将等式（10.199）的两侧乘以$q^{\backslash j}(\boldsymbol{\theta})$，然后积分，可得

$$K = \int \tilde{f}_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \tag{10.200}$$

其中我们已经使用了$q^{\text{新}}(\boldsymbol{\theta})$已经归一化这一事实。于是，$K$的值可以通过匹配零阶矩的方式得到

$$\int \tilde{f}_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} = \int f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \tag{10.201}$$

将这个式子与公式（10.197）结合，我们看到$K = Z_j$，因此可以通过计算公式（10.197）中的积分的方式得到。

在实际应用中，在因子集合中会进行多次迭代，每次都修正所有的因子。之后，使用公式（10.191）可以得到后验概率分布$p(\boldsymbol{\theta} \mid \mathcal{D})$的近似，模型证据$p(\mathcal{D})$可以使用公式（10.190）来近似，其中因子$f_i(\boldsymbol{\theta})$被替换为它们的近似$\tilde{f}_i(\boldsymbol{\theta})$。

我们给定观测数据集$\mathcal{D}$和随机变量$\boldsymbol{\theta}$上的联合概率分布，用因子的乘积的形式表示

$$p(\mathcal{D}, \boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta}) \tag{10.202}$$

我们希望使用下面形式的分布

$$q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_i \tilde{f}_i(\boldsymbol{\theta}) \tag{10.203}$$

来近似后验概率分布$p(\boldsymbol{\theta} \mid \mathcal{D})$。我们也希望近似模型证据$p(\mathcal{D})$。

- 初始化所有的近似因子$\tilde{f}_i(\theta)$
- 通过设置$q(\theta) \propto \prod_i \tilde{f}_i * \theta$初始化后验近似
- 直到收敛：
  - 选择一个因子$\tilde{f}_i(\theta)$进行优化
  - 通过$q^{\setminus j}(\theta) = \frac{q(\theta)}{\tilde{f}_j(\theta)}$ 从后验概率中移除$\tilde{f}_j(\theta)$
  - 算新的后验概率分布，方法为：令$q^{\text{new}}(\theta)$的充分统计量（矩）等于$q^{\setminus j}(\theta) f_j(\theta)$的充分统计量（矩），包括计算归一化系数

$$Z_j = \int q^{\setminus j}(\theta) f_j(\theta) \mathrm{d}theta$$

  - 计算和存储新的因子

$$\tilde{f}_j(\theta) = Z_j \frac{q^{\text{new}}(\theta)}{q^{\setminus j}(\theta)}$$

- 计算模型证据的近似

$$p(D) \simeq \int \prod_i \tilde{f}_i(\theta) \mathrm{d}\theta$$

期望传播的一个缺点是，它不保证迭代会收敛。然而，对于指数族分布的近似$q(\boldsymbol{\theta})$，如果迭代确实收敛，那么求得的解是特定的势函数的驻点（Minka, 2001a），虽然每轮EP迭代未必减小势函数的值。这与变分贝叶斯相反。变分贝叶斯中，每轮迭代保证不会减小界限。直接优化EP的代价函数是可能的，这种情况下，它保证收敛，虽然会导致算法更慢，实现起来更复杂。

变分贝叶斯和EP的另一个区别是来自于两个算法所最小化的KL散度的形式，因为前者最小化$\mathrm{KL}(q \| p)$，而后者最小化$\mathrm{KL}(p \| q)$。正如我们在图10.3中看到的那样，对于多峰的概率分布$p(\boldsymbol{\theta})$，最小化$\mathrm{KL}(p \| q)$会产生较差的近似。特别地，如果将EP应用于混合概率分布，那么得到的结果没有意义，因为得到的近似试图覆盖后验概率分布的所有峰值。相反，在logistic类型的模型中，EP通常要比局部变分方法和拉普拉斯近似方法的表现更好（Kuss and Rasmussen, 2006）。