**DMM的EM**

E步（Expectation步）：

1. **初始化参数：** 从模型参数的初始猜测开始，包括混合分量的混合系数以及各个组成部分的均值和协方差。
2. **计算后验概率（责任）：** 对于每个数据点，使用当前参数估计计算属于混合中每个组分的后验概率（责任）。通常使用贝叶斯定理进行计算。

$$P(Z_{ik}|X_i, \theta^{(t)}) = \frac{\pi_k^{(t)} \cdot f(X_i|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \cdot f(X_i|\mu_j^{(t)}, \Sigma_j^{(t)})} \tag{1}$$

其中 $Z_{ik}$ 是表示数据点 $i$ 是否属于组分 $k$ 的指示变量，$X_i$ 是观察到的数据点，$\theta^{(t)}$ 表示当前参数估计。

M步（Maximization步）：

1. **更新混合系数（$\pi_k$）：** 根据计算得到的责任更新混合系数。

$$\pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} P(Z_{ik}|X_i, \theta^{(t)}) \tag{2}$$

2. **更新组分参数（$\mu_k, \Sigma_k$）：** 根据计算得到的责任更新每个组分的均值（$\mu_k$）和协方差矩阵（$\Sigma_k$）。

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{N} P(Z_{ik}|X_i, \theta^{(t)}) \cdot X_i}{\sum_{i=1}^{N} P(Z_{ik}|X_i, \theta^{(t)})} \tag{3}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{N} P(Z_{ik}|X_i, \theta^{(t)}) \cdot (X_i - \mu_k^{(t+1)}) \cdot (X_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{N} P(Z_{ik}|X_i, \theta^{(t)})} \tag{4}$$

3. **计算对数似然：** 计算给定当前参数估计的数据的对数似然。

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \ln \left( \sum_{k=1}^{K} \pi_k \cdot f(X_i|\mu_k, \Sigma_k) \right) \tag{5}$$

4. **收敛检查：** 检查收敛标准，如对数似然的变化或参数值低于阈值。如果满足标准，终止算法；否则，返回到E步。

# EM for Bernoulli Mixture Models

**E-Step:**

$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \frac{\sum_{z_{nk}} z_{nk} \left[\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)\right]^{z_{nk}}}{\sum_{z_{nj}} \left[\pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)\right]^{z_{nj}}} = \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)}.$$

**M-Step:**

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \qquad \pi_k = \frac{N_k}{N}$$

$$\overline{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \qquad \boldsymbol{\mu}_k = \overline{\mathbf{x}}_k$$

**HMM的EM**

连续：

# EM Learning of HMMs (I)

$$
\begin{aligned}
\gamma(\mathbf{z}_n) &= p(\mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \\
\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) &= p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \\
\gamma(z_{nk}) &= \mathbb{E}[z_{nk}] = \sum \gamma(\mathbf{z}) z_{nk} \\
\xi(z_{n-1,j}, z_{nk}) &= \mathbb{E}[z_{n-1,j} z_{nk}] = \sum_{\mathbf{z}} \gamma(\mathbf{z}) z_{n-1,j} z_{nk}
\end{aligned}
$$

$$
p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}
$$

$$
\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} \qquad \boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})}
$$

$$
A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})} \qquad \boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_{n=1}^{N} \gamma(z_{nk})}
$$

离散：

# EM Learning of HMMs (II)

$$
\begin{aligned}
\gamma(\mathbf{z}_n) &= p(\mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \\
\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) &= p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \\
\gamma(z_{nk}) &= \mathbb{E}[z_{nk}] = \sum \gamma(\mathbf{z}) z_{nk} \\
\xi(z_{n-1,j}, z_{nk}) &= \mathbb{E}[z_{n-1,j} z_{nk}] = \sum_{\mathbf{z}} \gamma(\mathbf{z}) z_{n-1,j} z_{nk}
\end{aligned}
$$

$$
p(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^{D} \prod_{k=1}^{K} \mu_{ik}^{x_i z_k}
$$

$$
\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})}
$$

$$
A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})} \qquad \mu_{ik} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^{N} \gamma(z_{nk})}
$$

**What is the EM procedure? When do we need the EM procedure for machine learning?**

The EM algorithm is a general technique for finding maximum likelihood solutions for probabilistic models having latent variables. We first choose some initial values for the means, covariances, and mixing coefficients. Then we alternate between the following two updates that we shall call the E step and the M step. In the E step, we use the current values for the parameters to evaluate the posterior probabilities, or responsibilities. We then use these probabilities in the M step, to re-estimate the means, covariances, and mixing coefficients.

The EM procedure helps estimate parameters in situations where direct optimization or analytical solutions are infeasible due to the presence of latent variables or incomplete information. It is a powerful tool for finding maximum likelihood estimates in a wide range of statistical models.

**What is the EM procedure in terms of the Q function?**

The Q function $Q(\theta, \theta^{(t)})$ is defined as follows:

$$Q(\theta, \theta^{(t)}) = \mathbb{E}_{Z|X,\theta^{(t)}}[\log P(X, Z|\theta)]$$

1. **E-step (Expectation step):** Compute the expected value of the log-likelihood function.
   $Q(\theta, \theta^{(t)}) = \sum_Z P(Z|X, \theta^{(t)}) \cdot \log P(X, Z|\theta)$
2. **M-step (Maximization step):** Update the parameter estimates $\theta$ by maximizing the Q function:
   $\theta^{(t+1)} = \arg\max_\theta Q(\theta, \theta^{(t)})$

**What is the EM procedure in terms of likelihood and KL divergence?**

The likelihood function $L(\theta)$ for observed data $X$ given model parameters $\theta$ is expressed as:

$$L(\theta) = P(X|\theta)$$

1. **E-step (Expectation step):**
   - Evaluate the posterior distribution of latent variables $Z$ given the observed data and the current parameter estimates:
     $P(Z|X, \theta^{(t)}) = \frac{P(X,Z|\theta^{(t)})}{P(X|\theta^{(t)})}$
   - Compute the expected log-likelihood function:
     $Q(\theta, \theta^{(t)}) = \sum_Z P(Z|X, \theta^{(t)}) \cdot \log P(X, Z|\theta)$
2. **M-step (Maximization step):**
   - Update the parameter estimates $\theta$ by maximizing the expected log-likelihood:
     $\theta^{(t+1)} = \arg\max_\theta Q(\theta, \theta^{(t)})$

$KL(P(Z|X, \theta^{(t)})||P(Z|X, \theta))$ Minimizing this KL divergence is equivalent to maximizing the expected log-likelihood in the E-step.

**What is the EM procedure in terms of optimization of non-convex function?**

- The E-step computes the expected log-likelihood, involving an expectation over the latent variable distribution. This step may not have a closed-form solution and may require numerical methods.

- The M-step optimizes the parameters to maximize the expected log-likelihood. Since the expected log-likelihood is a non-convex function with respect to the parameters, traditional optimization techniques for non-convex problems are employed. Methods such as gradient-based optimization, quasi-Newton methods, or even metaheuristic optimization algorithms can be used in this step.