Maximum likelihood estimation of Markov Chains

We use MLE to estimate the transition matrix A from data $\mathcal{D} = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$.

Likelihood of any particular sentence $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_{T_i}^{(i)})$ of length T_i, given parameters $\theta = (A_{ij}, \pi_j)_{i,j=1}^K$

$$p(\mathbf{x}^{(i)}|\theta) = \prod_{j=1}^{K} \pi_j^{1[x_1^{(i)}=j]} \prod_{t=2}^{T_i} \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{1[x_t^{(i)}=k, x_{t-1}^{(i)}=j]}$$

Log-likelihood of $\mathcal{D}$ (all sentences treated as independent)

$$\log p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}|\theta) = \sum_{j=1}^{K} N_j^1 \log \pi_j + \sum_{j=1}^{K} \sum_{k=1}^{K} N_{jk} \log A_{jk} \qquad (1)$$

where we define the counts

$$N_j^1 = \sum_{i=1}^{N} 1[x_1^{(i)} = j], \qquad N_{jk} = \sum_{i=1}^{N} \sum_{t=2}^{T_i} 1[x_t^{(i)} = k, x_{t-1}^{(i)} = j].$$

To calculate the MLE of A and π, we need to find parameters $\hat{\theta} = \left(\hat{A}_{ij}, \hat{\pi}_j\right)_{i,j=1}^K$ such that $\log p(\mathcal{D}|\hat{\theta})$ is maximized.

For $\hat{\pi}$, we only focus on the first term in R.H.S. of (1)

$$\begin{aligned}
\hat{\pi} &= \arg\max_{\pi} \sum_j N_j^1 \log \pi_j \\
&= \arg\max_{\pi} \sum_j \frac{N_j^1}{\sum_{j'} N_{j'}^1} \log \pi_j \\
&= \arg\max_{\pi} \sum_j q_j \log \pi_j \qquad \text{(define probabilities } q_j = \frac{N_j^1}{\sum_{j'} N_{j'}^1}\text{)} \\
&= \arg\max_{\pi} -H(q) - D_{KL}(q\|\pi) \\
&= \arg\min_{\pi} D_{KL}(q\|\pi) = q.
\end{aligned}$$

Hence,

$$\hat{\pi}_j = q_j = \frac{N_j^1}{\sum_{j'} N_{j'}^1},$$

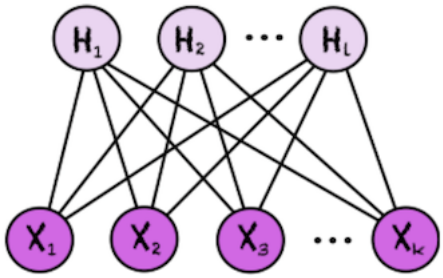
and similarly,

$$\hat{A}_{jk} = \frac{N_{jk}}{\sum_{k'} N_{jk'}}.$$

Gibbs sampler for Restricted Boltzmann Machines (RBMs)

Model for $(X_1, \ldots, X_k, H_1, \ldots, H_l) \in \{-1, 1\}^{k+l}$ (c.f. *Tutorial 3*)

$$p(x_1, \ldots, x_k, h_1, \ldots, h_l) \;\propto\; \exp\{\sum_i \alpha_i x_i + \sum_i \beta_i h_i + \sum_{i=1}^{k}\sum_{j=1}^{l} J_{ij} x_i h_j\}. \quad (2)$$



We can easily generate new samples from the learned distribution, as the visible units are conditionally independent given the hidden units, and vice versa.

$$p(\mathbf{x}|\mathbf{h}) = \prod_{i=1}^{k} p(x_i|\mathbf{h}), \qquad p(\mathbf{h}|\mathbf{x}) = \prod_{j=1}^{l} p(h_j|\mathbf{x})$$

Given (2), we can see that

$$p(x_i, \mathbf{h}) = p(x_i, h_1, h_2, \ldots, h_l) \;\propto\; \exp\{\alpha_i x_i + \sum_j \beta_j h_j + \sum_{j=1}^{l} J_{ij} x_i h_j\}.$$

Hence, for any value of x_i and $\mathbf{h}$, we have

$$p(x_i, \mathbf{h}) = \frac{1}{Z} \exp\{\alpha_i x_i + \sum_j \beta_j h_j + \sum_{j=1}^{l} J_{ij} x_i h_j\},$$

where the normalizing factor Z does not depend on $x_i, \mathbf{h}$. Now, we are ready to calculate $p(x_i|\mathbf{h})$:

$$
\begin{aligned}
p(x_i = 1|\mathbf{h}) &= \frac{p(x_i = 1, \mathbf{h})}{p(\mathbf{h})} \\
&= \frac{p(x_i = 1, \mathbf{h})}{p(x_i = 1, \mathbf{h}) + p(x_i = -1, \mathbf{h})} \\
&= \frac{\frac{1}{Z}\exp\{\alpha_i + \sum_j \beta_j h_j + \sum_{j=1}^{l} J_{ij} h_j\}}{\frac{1}{Z}\exp\{\alpha_i + \sum_j \beta_j h_j + \sum_{j=1}^{l} J_{ij} h_j\} + \frac{1}{Z}\exp\{-\alpha_i + \sum_j \beta_j h_j - \sum_{j=1}^{l} J_{ij} h_j\}} \\
&= \frac{1}{1 + \exp\{-2(\alpha_i + \sum_{j=1}^{l} J_{ij} h_j)\}} \\
&= \sigma(2(\alpha_i + \sum_{j=1}^{l} J_{ij} h_j)),
\end{aligned}
$$

with $\sigma(y) = 1/(1 + e^{-y})$ called the **sigmoid function**. We can similarly show that

$$
p(h_j = 1|\mathbf{x}) = \sigma(2(\beta_j + \sum_{i=1}^{k} J_{ij} x_i))
$$

Gibbs sampling for the Ising model

The previous example generalizes to any Ising model. So suppose that

$$
p(\mathbf{x}) \;\propto\; \exp\{\sum_i b_i x_i + \sum_{i \sim j} J_{ij} x_i x_j\} \qquad \text{for all } \mathbf{x} \in \{-1, 1\}^m.
$$

Fix $k \in \{1, \ldots, m\}$. The corresponding full conditional is

$$
\begin{aligned}
&p(x_k = 1|\mathbf{x}_{\setminus k}) \\
&= \frac{p(x_k = 1, \mathbf{x}_{\setminus k})}{p(x_k = -1, \mathbf{x}_{\setminus k}) + p(x_k = 1, \mathbf{x}_{\setminus k})} \\
&= \frac{\exp\{\sum_{i \neq k} b_i x_i + b_k + \sum_{l \sim k} J_{kl} x_l + \sum_{i \sim j; i \neq k, j \neq k} J_{ij} x_i x_j\}}{\exp\{\sum_{i \neq k} b_i x_i - (b_k + \sum_{l \sim k} J_{kl} x_l) + \sum_{i \sim j; i \neq k, j \neq k} J_{ij} x_i x_j\} + \exp\{\sum_{i \neq k} b_i x_i + b_k + \sum_{l \sim k} J_{kl} x_l + \sum_{i \sim j; i \neq k, j \neq k} J_{ij} x_i x_j\}} \\
&= \frac{1}{1 + \exp\{-2(b_k + \sum_{l \sim k} J_{kl} x_l)\}} \\
&= \sigma(2(b_k + \sum_{l \sim k} J_{kl} x_l)),
\end{aligned}
$$

This suggest that the Gibbs sampler can be trivially implemented for the Ising model. The adventage of the the previous example is that for bipartite graphs, we can update the whole group of variables at once.