Chapter 8

Model Inference and Averaging

EM algorithm is summarized as $\theta^{n+1} = \arg\max_\theta \left\{ \mathbb{E}\left[\log \mathbb{P}\left(Z, Z^m; \theta\right)\right]_{Z^m|Z;\theta^n} \right\}$. The art of choosing Z lies in the aim to simplify the maximizing of $\log \mathbb{P}\left(Z, Z^m; \theta\right) = \log \mathbb{P}\left(Z|Z^m; \theta\right) + \log \mathbb{P}\left(Z^m; \theta\right)$ over θ.

Problem 8.1

Gibbs Inequality

$$
\begin{aligned}
\mathbb{E}\left[\log \frac{r(Y)}{q(Y)}\right]_q &= \int_y q(y) \log \frac{r(y)}{q(y)} dy \\
\text{(Jensen's inequality)} &\geq \log\left(\int_y q(y) \frac{r(y)}{q(y)} dy\right) = 0
\end{aligned}
$$

$$
\begin{aligned}
& R(\theta', \theta) = \mathbb{E}\left[\log \mathbb{P}\left(Z^m|Z;\theta\right)\right]_{Z^m|Z;\theta'} \\
&= \int \mathbb{P}\left(Z^m|Z;\theta\right) \log \mathbb{P}\left(Z^m|Z;\theta'\right) dZ^m \\
&= \int \mathbb{P}\left(Z^m|Z;\theta\right) \log \frac{\mathbb{P}\left(Z^m|Z;\theta'\right)}{\mathbb{P}\left(Z^m|Z;\theta\right)} dZ^m - \int \mathbb{P}\left(Z^m|Z;\theta\right) \log \frac{1}{\mathbb{P}\left(Z^m|Z;\theta\right)} dZ^m \\
&\geq -\int \mathbb{P}\left(Z^m|Z;\theta\right) \log \frac{1}{\mathbb{P}\left(Z^m|Z;\theta\right)} dZ^m \text{ with the equality holds when } \theta' = \theta.
\end{aligned}
$$

This suggest that deviating θ' from θ can only increase $-R(\theta', \theta)$, meaning that maximizing $Q(\theta', \theta)$ over θ' will always leads to the increase in the log likelihood.

Problem 8.2

$$
\begin{aligned}
F(\theta, \tilde{P}) =& \mathbb{E}\left[l_0(Z, Z^m; \theta)\right]_{\tilde{P}} - \mathbb{E}\left[\log \tilde{P}(Z^m)\right]_{\tilde{P}} \\
=& \sum_{Z^m} \tilde{P}(Z^m) \log \mathbb{P}\left(Z, Z^m; \theta\right) - \sum_{Z^m} \tilde{P}(Z^m) \log \mathbb{P}\left(Z^m\right) \\
=& \sum_{Z^m} \tilde{P}(Z^m) \log \left(\mathbb{P}\left(Z^m | Z; \theta\right) \mathbb{P}\left(Z; \theta\right)\right) - \sum_{Z^m} \tilde{P}(Z^m) \log \mathbb{P}\left(Z^m\right) \\
=& \sum_{Z^m} \tilde{P}(Z^m) \log \frac{\mathbb{P}\left(Z^m | Z; \theta\right)}{\mathbb{P}\left(Z^m\right)} + \mathbb{P}\left(Z; \theta\right) \\
\text{(Jensen's inequality)} \leq& \log \left(\sum_{Z^m} \tilde{P}(Z^m) \frac{\mathbb{P}\left(Z^m | Z; \theta\right)}{\mathbb{P}\left(Z^m\right)}\right) + \mathbb{P}\left(Z; \theta\right) = \mathbb{P}\left(Z; \theta\right),
\end{aligned}
$$

where the last inequality becomes equality only when $\tilde{P}(Z^m) = \mathbb{P}\left(Z^m | Z; \theta\right)$.

Problem 8.3

One sentence proof of Gibbs sampling: The transition probabilities induced by Gibbs sampling satisfies detailed balance equation when we set the stationary distribution to be the joint distribution, and thus it leads to a reversible Markov chain with the equilibrium distribution being the joint distribution.

$$
\begin{aligned}
& \frac{1}{M-m+1} \sum_{t=m}^{M} \mathbb{P}\left(u_k \middle| U_l^{(t)}, l \neq k\right) \\
=& \frac{1}{M-m+1} \sum_{t=m}^{M} \sum_{U_{\sim k} \in \boldsymbol{U}_{\sim k}} \mathbb{P}\left(u_k | U_{\sim k}\right) \mathbf{1}\left(U_{\sim k}^{(t)} = U_{\sim k}\right) \\
=& \sum_{U_{\sim k} \in \boldsymbol{U}_{\sim k}} \mathbb{P}\left(u_k | U_{\sim k}\right) \sum_{t=m}^{M} \frac{\mathbf{1}\left(U_{\sim k}^{(t)} = U_{\sim k}\right)}{M-m+1} \\
\to & \sum_{U_{\sim k} \in \boldsymbol{U}_{\sim k}} \mathbb{P}\left(u_k | U_{\sim k}\right) \mathbb{P}\left(U_{\sim k}\right) \text{ as } M \text{ approaches infinity} \\
=& \mathbb{P}\left(u_k\right).
\end{aligned}
$$

Problem 8.7

From Equation (8.46), it should be clear that $\log \mathbb{P}\left(Z; \theta\right) = Q(\theta, \theta) - R(\theta, \theta)$. Then we have

$$
\begin{aligned}
g(\theta', \theta) \triangleq& Q(\theta', \theta) + \log \mathbb{P}\left(Z; \theta\right) - Q(\theta, \theta) \\
=& Q(\theta', \theta) - R(\theta, \theta).
\end{aligned}
$$

From Problem 8.1 we know that $-R(\theta', \theta) \geq -R(\theta, \theta)$ (KL divergence is always non-negative), which yield

$$
Q(\theta', \theta) - R(\theta, \theta) \leq Q(\theta', \theta) - R(\theta', \theta) = \log \mathbb{P}\left(Z; \theta'\right).
$$

Then, by denoting $f(\theta') = \log \mathbb{P}(Z; \theta')$, we have $g(\theta', \theta)$ minorizes function $f(\theta')$.

The EM algorithm, which is to maximize $Q(\theta', \theta)$ over θ', effectively maximizes $g(\theta', \theta)$ as well given that they differ by a constant term $R(\theta, \theta)$.