

Collaborators :

- Victor Zuanazzi

ML - 2

Homework Assignment - 6

Shantanu Chandra
(12048461)

Q1.

a) Rejection Sampling algo :

1. $n \leftarrow 0$, samples $\leftarrow []$
2. while $n < N$:
 3. $x_i \sim q(x)$
 4. range $= c \cdot \tilde{q}(x_i)$
 5. refer $= \tilde{p}(x_i)$
 6. $u_i \sim U[0, \text{range}]$
 7. if $u_i > \text{refer}$:
 8. samples $\leftarrow \text{samples} \cup x_i$
 9. $n \leftarrow n + 1$

b) The samples are independent as they are independently generated from the known distribution $q(z)$.

c) $w_n = \frac{\tilde{p}(x_n)}{\tilde{q}(x_n)} = \frac{z_p p(x_n)}{z_q q(x_n)}$ $\underline{\underline{\text{Ans.}}}$

d) $\alpha(x_{t+1}, x_t) = \min \left(1, \frac{\tilde{p}(x_{t+1}) \tilde{q}(x_t | x_{t+1})}{\tilde{p}(x_t) \tilde{q}(x_{t+1} | x_t)} \right)$

$$= \min \left(1, \frac{\tilde{p}(x_{t+1}) \tilde{q}(x_t | x_{t+1})}{\tilde{p}(x_t) \tilde{q}(x_{t+1})} \right) \quad (\text{using } \tilde{q}(x_{t+1}|x_t) = q(x_{t+1}))$$

$$= \min \left(1, \frac{p(x_{t+1}) z_p}{p(x_t) z_p} \frac{\tilde{q}(x_t, x_{t+1})}{\tilde{q}(x_{t+1})^2} \right) \quad (\text{using } \tilde{p}(z) = z_p p(z) \text{ & using joint of } \tilde{q}(x_t, x_{t+1}))$$

$$= \min \left(1, \frac{p(x_{t+1})}{p(x_t)} \cdot \frac{g(x_{t+1}|x_t)}{g(x_{t+1})^2} \cdot \frac{q(x_t)}{q(x_{t+1})} \right) \quad (\text{using } q(x_{t+1}|x_t) = q(x_{t+1}) \text{ & using joint of } \tilde{q}(x_t, x_{t+1}))$$

$$= \min \left(1, \frac{p(x_{t+1}) q(x_t)}{p(x_t) q(x_{t+1})} \right)$$

$\underline{\underline{\text{Ans.}}}$

e) From the above eq. we can see that the acceptance probability of new proposed sample (x_{t+1}) depends on previous sample (x_t). Thus subsequent samples are generally not independent. However, as defined, the sampling is independent $x_{t+1} \sim q(x_{t+1}|x_t) = q(x_{t+1})$.

$\underline{\underline{\text{Ans.}}}$

f) Independence Sampler produces : $[x_1, x_2, x_3, x_4, x_5] = [0.34, 0.34, 2.67, 0.82, 0.82]$
 $\underline{\underline{\text{Ans.}}}$

g) ① Rejection Sampling : if $p(x)$ is k dimensional then we need to find $q(x)$ such that $kq(x) \geq p(x)$. Thus, first limitation comes from choosing a suitable $q(x)$ in high dimensions.

Secondly, The acceptance rate is ratio of volumes under $p(x)$ & $kq(x)$. Since $p(x)$ & $kq(x)$ are normalized, it is $\frac{1}{k}$. This diminishes rapidly as the dimensions increase.

(as illustrated in Bishop pg 531). Thus we end up rejecting most samples.

- ② Importance Sampling: just like above, we need to come up with $q(x)$ but without the earlier upper bounding constraint. However since $w_n = \frac{p(x_n)}{q(x_n)}$, it is difficult to come up with $q(x)$ such that both $p(x) & q(x)$ are large together to reduce variance & hence rejection rate (Eq: $p(x) > 0, q(x) < 0 \Rightarrow$ high variance). In higher dimensions making sure of $p(x) > 0 & q(x) > 0$ is even more difficult.
- ③ Independence Sampling: it is a MCMC sampling method & hence in theory its performance is not limited by choice of a good proposal distribution. It performs better than the above 2 in higher dimensions however can still suffer from curse of dimensionality as the no. of transition steps can exponentially increase with higher dimensions due to higher rejection rate

Q 2.

In Gibbs sampling we can model the posterior from the conditionals as it can be possible to easily sample from it (eg: Markov blanket).

Here, the posterior can be broken into conditionals as $p(\mu, \tau | x) \rightarrow p(\mu | \tau, x), p(\tau | \mu, x)$

$$\begin{aligned} \text{For } \tau : \quad p(\tau | \mu, x) &= \frac{p(\mu, x | \tau) p(\tau)}{p(x, \mu)} = \frac{p(x | \mu, \tau) p(\mu | \tau)}{p(x | \mu) p(\mu)} p(\tau) \quad [\because \mu \& \tau \text{ are decoupled given } x] \\ &= \frac{p(x | \mu, \tau) p(\tau)}{p(x | \mu)} \propto p(\tau | \mu, x) p(\tau) \end{aligned}$$

$$\begin{aligned} \text{Here, } p(x | \mu, \tau) &= N(x | \mu, \tau^{-1}) \\ p(\tau) &= p(\tau | a, b) = \text{gamma}(\tau | a, b) \end{aligned}$$

Since they both belong to the exponential family, things become easy

$$\begin{aligned} p(\tau | \mu, x) &\propto \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\tau}{2}(x-\mu)^2\right)}_{\text{Normal}} \cdot \underbrace{\frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau)}_{\text{Gamma}} \\ &= \frac{b^a}{\Gamma(a)} \tau^{a+\frac{1}{2}-1} \exp\left(-\tau\left[\frac{(x-\mu)^2}{2} + b\right]\right) \\ &= \text{gamma}\left(\tau \mid a + \frac{1}{2}, \frac{(x-\mu)^2}{2} + b\right) \\ &\equiv \text{Ans.} \end{aligned}$$

For μ :

Similarly for μ , we have multiplication of 2 Gaussians:

$$\begin{aligned} p(\mu | x, \tau) &= p(x | \mu, \tau) \cdot p(\mu | \mu_0, s_0) \\ &= N(x | \mu, \tau^{-1}) \cdot N(\mu | \mu_0, s_0) \\ &= N\left(\mu \mid \left(\tau + \frac{1}{s_0}\right)^{-1} \left(x\tau + \frac{\mu_0}{s_0}\right), \tau + \frac{1}{s_0}\right) \quad (\text{using 2.141 \& 2.142 of Bishop}) \\ &\equiv \text{Ans.} \end{aligned}$$

Q3.

D) The joint probability is given by :

$$p(w, z, \phi, \theta | \alpha, \beta) = p(\theta | \alpha) p(z | \theta) p(\phi | \beta) p(w | \phi, z) \quad \left[\begin{array}{l} \alpha = (\alpha_1, \dots, \alpha_D) \\ \beta = (\beta_1, \dots, \beta_K) \end{array} \right]$$

$$= \prod_{d=1}^D \text{Dir}(\theta_d | \alpha_d) \prod_{k=1}^K \text{Dir}(\phi_k | \beta_k) \prod_{n=1}^{N_d} \text{Mult}(z_{nd} | \theta_n) \text{Mult}(w_{dn} | \phi_{dn}, z_{dn}) \\ = f_{\theta, \phi, z, w}.$$

2)

$$p(w | z, \phi) = \prod_{d=1}^D \prod_{n=1}^{N_d} p(w_{dn} | z_{dn}, \phi_{dn})$$

$$= \prod_{d=1}^D \prod_{n=1}^{N_d} \prod_{k=1}^K \prod_w p(w_{dn} | z_{dn}, \phi_{dn}) \underbrace{\delta(z_{nd} = k)}_{\text{Mult}(\phi_{dn})} \underbrace{\delta(w_{dn} = w)}_{B_{dnw}}$$

$$= \prod_{k=1}^K \prod_w p(w_{dn} = w | z_{dn} = k, \phi_{dn}) \underbrace{\sum_{n=1}^{N_d} \delta(z_{nd} = k)}_{A_{dnk}} \underbrace{\delta(w_{dn} = w)}_{B_{dnw}}$$

$$= \prod_{k=1}^K \prod_w \phi_{dnk}^{B_{dnw}}$$

$$p(z | \theta) = \prod_{n=1}^{N_d} \prod_{d=1}^D p(z_{nd} | \theta) \\ = \prod_{n=1}^{N_d} \prod_{d=1}^D \prod_{k=1}^K p(z_{nd} = k | \theta_{dk}) \underbrace{\delta(z_{nd} = k)}_{\text{Mult}(\theta_{dk})} \\ = \prod_{k=1}^K \prod_{d=1}^D p(z_{nd} = k | \theta_{dk}) \underbrace{\sum_{n=1}^{N_d} \delta(z_{nd} = k)}_{A_{dk}} \\ = \prod_{k=1}^K \prod_{d=1}^D \theta_{dk}^{A_{dk}}$$

Using the above 2 results, we evaluate the integrals :

$$\iint_{\theta, \phi} p(w, z, \theta, \phi | \alpha, \beta) d\theta d\phi = \int_{\theta} p(\theta | \alpha) p(z | \theta) d\theta \int_{\phi} p(\phi | \beta) p(w | \phi, z) d\phi$$

$$\int_{\theta} p(\theta | \alpha) p(z | \theta) d\theta = \prod_{d=1}^D \int_{\theta_d} p(\theta_d | \alpha_d) \prod_{n=1}^{N_d} p(z_{nd} | \theta_n) d\theta \\ = \prod_{d=1}^D \int_{\theta_d} \frac{1}{B(\alpha_d)} \theta_d^{x-1} \cdot \prod_{k=1}^K \theta_{dk}^{A_{dk}} d\theta_d \\ = \prod_{d=1}^D \int_{\theta_d} \frac{1}{B(\alpha_d)} \prod_{k=1}^K \theta_{dk}^{A_{dk} + \alpha_{dk} - 1} d\theta_d \\ = \prod_{d=1}^D \frac{B(A_d + \alpha_d)}{B(\alpha_d)} \int_{\theta_d} \frac{1}{B(A_d + \alpha_d)} \theta_{dk}^{A_{dk} + \alpha_{dk} - 1} d\theta_d \\ = \prod_{d=1}^D \frac{B(A_d + \alpha_d)}{B(\alpha_d)} \int_{\theta_d} \text{Dir}(\theta_d | A_{dk} + \alpha_d) = 1 \quad (\text{Area of a distribution})$$

$$= \prod_{d=1}^D \frac{B(A_d + \alpha)}{B(\alpha)} \quad \text{--- (1)}$$

$$\begin{aligned}
\int_{\Phi} p(\Phi | \beta) p(w | \Phi, z) d\Phi &= \prod_{k=1}^K \int_{\Phi_k} p(\Phi_k | \beta) \prod_{n=1}^{N_d} \prod_{a=1}^D p(w_{nd} | z_{nd}, \phi_{aw}) d\phi_k \\
&= \prod_{k=1}^K \int_{\Phi_k} \frac{1}{\Phi_k^{p-1}} \cdot \prod_{aw} \phi_{aw}^{B_{aw}} d\phi_k \\
&= \prod_{k=1}^K \frac{B(B_k + p)}{B(p)} \int_{\Phi_k} \frac{1}{\Phi_k^{B_k + p}} \prod_{aw} \phi_{aw}^{p + B_{aw}-1} d\phi_k \\
&= \prod_{k=1}^K \frac{B(B_k + p)}{B(p)} \int_{\Phi_k} \text{Dir}(\Phi_k | B_k + p) d\Phi_k \\
&\qquad\qquad\qquad = 1 \\
&= \prod_{k=1}^K \frac{B(B_k + p)}{B(p)} \quad \text{--- (2)}
\end{aligned}$$

Finally, combining (1) & (2) :

$$p(w, z | \alpha, \beta) = \prod_{d=1}^D \frac{B(A_d + \alpha)}{B(\alpha)} \cdot \prod_{k=1}^K \frac{B(B_k + p)}{B(p)}$$

Ans.

- 3) In Gibbs sampling, to sample from $p(z) = p(z_1, \dots, z_m)$ we iteratively sample each individual dimension conditioned on others. Thus in our case :

$$\begin{aligned}
p(z_i | z_{-i}, w) &= \frac{p(w, z)}{p(w, z_{-i})} \quad (\text{where } z_{-i} \text{ represents all } z \text{ except } z_i) \\
&= \frac{p(w, z)}{p(w_{-i}, z_{-i})} \underbrace{p(w_i)}_{=\text{constant}}
\end{aligned}$$

Using results from previous question :

$$\begin{aligned}
&= \prod_{d=1}^D \frac{B(A_d + \alpha)}{B(A_d^{(-i)} + \alpha)} \prod_{k=1}^K \frac{B(B_k + p)}{B(B_k^{(-i)} + p)} \cdot \frac{1}{p(w_i)} \\
&\qquad\qquad\qquad \text{[Ans.]}
\end{aligned}$$

(Here, $i = \text{dimens. of } z = \text{nd}$)

[Here, $^{(-i)}$ denotes the same distribution but without z_i]

Q4.

a) $\mathbb{E}[x] = \sum_{x_i} x_i p(x_i) = 1 \cdot \mu_i + 0 \cdot (1 - \mu_i) = \stackrel{\Rightarrow}{=} \mu_i$ ans.

b) $\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$
 $= \mathbb{E}[x_i^2] - \mu_i^2$
 $= \mu_i(1 - \mu_i)$

Since the dimensions of x are independent of each other,
 $\text{cov}[x] = \text{diag.}[\mu_i(1 - \mu_i)]$
 $\stackrel{\Rightarrow}{=} \text{Ans.}$

c) $\mathbb{E}[x] = \sum_{x_i} p(x_i | \mu, \pi) x_i$
 $= \sum_{x_i} \sum_{k=1}^K \pi_k p(x_i | \mu_k) x_i$
 $= \sum_{k=1}^K \pi_k \sum_{x_i} p(x_i | \mu_k) x_i$
 $= \sum_{k=1}^K \pi_k \cdot \mu_k \quad (\text{using result from previous part})$
 $\stackrel{\Rightarrow}{=} \text{Ans.}$

d) Assuming each sample to be i.i.d :

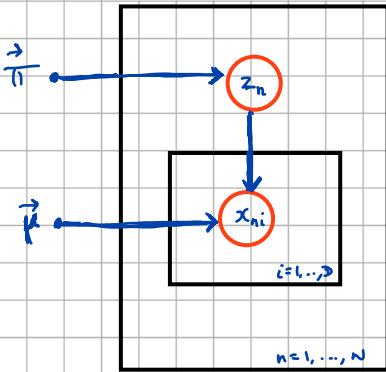
$$\begin{aligned} \log p(x_n | \mu, \pi) &= \log \prod_{k=1}^K (\pi_k p(x_n | \mu_k)) \\ &= \log \left[\prod_{k=1}^K \left(\pi_k \prod_{i=1}^D [\mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1-x_{ni}}] \right) \right] \\ \therefore \log p(x | \mu, \pi) &= \log \left[\prod_{n=1}^N \left[\prod_{k=1}^K \left(\pi_k \prod_{i=1}^D [\mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1-x_{ni}}] \right) \right] \right] \\ &= \sum_{n=1}^N \log \left[\prod_{k=1}^K \left(\pi_k \prod_{i=1}^D [\mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1-x_{ni}}] \right) \right] \\ &\stackrel{\Rightarrow}{=} \text{Ans.} \end{aligned}$$

e) Due to the summation inside the log, analytically finding a closed form solution is difficult.
 Thus we can optimize this using standard likelihood methods.

f) complete log data likelihood :

$$\begin{aligned} \log p(x, z | \mu, \pi) &= \log \left[\prod_{n=1}^N p(z_n | \pi) p(x_n | z_n, \mu) \right] \\ &= \sum_{n=1}^N \left[\log \left(\prod_{k=1}^K \pi_k^{z_{nk}} \cdot \left(\prod_{i=1}^D \mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1-x_{ni}} \right)^{z_{nk}} \right) \right] \\ &= \sum_{n=1}^N \sum_{k=1}^K \left[z_{nk} \log \pi_k + z_{nk} \sum_{i=1}^D (x_{ni} \log \mu_{ki} + (1 - x_{ni}) \log (1 - \mu_{ki})) \right] \\ &\stackrel{\Rightarrow}{=} \text{Ans.} \end{aligned}$$

g) Data likelihood : $p(x_n, z_n | \mu, \pi) = \underbrace{p(z_n | \pi)}_{\downarrow} p(x_n | z_n, \mu)$



z_n = latent/unobserved

x_n = observed

μ, π = parameters

π is one-of- K vector.

μ is $K \times D$ matrix.

h) To optimize the complete data likelihood, we introduce $q(z)$ as a distribution over 'z' :

$$\begin{aligned} \ln p(x|\theta) &= \ln q_z(\theta) + KL(q||p) \\ &\leq L(q, \theta) \end{aligned} \quad \text{where, } L(q, \theta) = \sum_z q_z(z) \ln \left[\frac{p(x, z|\theta)}{q_z(z)} \right] \quad (\text{Bishop 9.70 - 9.72})$$

$$KL(q||p) = -\sum_z q_z(z) \ln \left[\frac{p(z|x, \theta)}{q_z(z)} \right]$$

Thus, we can optimize the complete data likelihood by optimizing the ELBO which gives us the VEM optimization objective :

$$\begin{aligned} \mathcal{B}(q_n(z_n), \mu, \pi) &= \sum_{n=1}^N \sum_z q_n(z_n) \ln \frac{p(x_n, z_n | \mu, \pi)}{q_n(z_n)} \\ &= \sum_{n=1}^N \sum_z \left[q_n(z_n) \log p(x_n, z_n | \mu, \pi) - q_n(z_n) \log q_n(z_n) \right] \\ &= \sum_{n=1}^N \sum_{z_n=1}^K \left[q_n(z_n) \sum_{k=1}^K \left[z_{nk} \log \pi_{zk} + z_{nk} \sum_{i=1}^D (x_{ni} \log \mu_{ki} + (1-x_{ni}) \log (1-\mu_{ki})) \right] \right] - \sum_{n=1}^N \sum_{z_n=1}^K q_n(z_n) \log q_n(z_n) \\ &\stackrel{\text{Ans}}{=} \text{Ans}. \end{aligned}$$

i) There are 2 constraints here, $\sum_k \pi_k = 1$ & $\sum_{z_n=1}^K q_n(z_n) = 1 \quad \forall n \in \{1, \dots, N\}$

$$\therefore \tilde{\mathcal{B}}(q_n(z_n), \mu, \pi) = \mathcal{B}(q_n(z_n), \mu, \pi) + \lambda \sum_k (\pi_k = 1) + \sum_n \lambda_n \left(\sum_{z_n=1}^K q_n(z_n) - 1 \right)$$

~~Ans~~.

j) In the E-step we derive $\tilde{\mathcal{B}}(q_n(z_n), \mu, \pi)$ w.r.t $q_n(z_{ni})$:

$$\frac{\partial \tilde{\mathcal{B}}(q_n(z_n), \mu, \pi)}{\partial q_n(z_{ni})} = \sum_{z_n=1}^K \left(\ln \pi_k + \sum_{i=1}^D [x_{ni} \ln \mu_{ki} + (1-x_{ni}) \ln (1-\mu_{ki})] \right) - \ln q_n(z_n) - 1 + \lambda_n$$

(each word of a doc. belongs to just 1 topic 'k')

$$\log g_n(z_n) = \lambda_{n-1} + \log \pi_k + \sum_{i=1}^D (x_{ni} \log \mu_{ki} + (1-x_{ni}) \log (1-\mu_{ki}))$$

$$g_n(z_n) = \exp[\lambda_{n-1}] \cdot \pi_k \prod_{i=1}^D \frac{x_{ni}}{\mu_{ki}} \frac{(1-x_{ni})}{(1-\mu_{ki})} \quad \text{--- (1)}$$

Now, we know that $\sum_{k=1}^K g_n(z_n) = 1$. We use this to solve for $e^{\lambda_{n-1}}$:

$$1 = e^{\lambda_{n-1}} \sum_{k=1}^K \pi_k \prod_{i=1}^D \frac{x_{ni}}{\mu_{ki}} \frac{(1-x_{ni})}{(1-\mu_{ki})}$$

$$\therefore e^{\lambda_{n-1}} = \frac{1}{\sum_{k=1}^K \pi_k \prod_{i=1}^D \frac{x_{ni}}{\mu_{ki}} \frac{(1-x_{ni})}{(1-\mu_{ki})}}$$

Substituting this back in (1), we get :

$$g_n(z_n) = \frac{\prod_{i=1}^D \frac{x_{ni}}{\mu_{ki}} \frac{(1-x_{ni})}{(1-\mu_{ki})}}{\sum_{k=1}^K \pi_k \prod_{i=1}^D \frac{x_{ni}}{\mu_{ki}} \frac{(1-x_{ni})}{(1-\mu_{ki})}}$$

\equiv
Ans.

This is the posterior distribution :
 $p(z_n | x_n, \mu, \pi)$

\equiv
Ans.

k) In the M-step, we maximize μ & π while keeping $g_n(z_n)$ constant :

For μ :

$$\frac{\partial \tilde{B}(g_n(z_n), \mu, \pi)}{\partial \mu_{ki}} = \sum_{z_n=1}^K g_n(z_n) \underbrace{\left(\frac{x_{ni}}{\mu_{ki}} + \frac{1-x_{ni}}{1-\mu_{ki}} \right)}_{=1} + 0$$

$$0 = \sum_{z_n=1}^K g_n(z_n) (x_{ni} - \mu_{ki})$$

$$\mu_{ki} = \frac{\sum_{z_n=1}^K g_n(z_n) x_{ni}}{\sum_{z_n=1}^K g_n(z_n)}$$

\equiv
Ans.

For π :

$$\frac{\partial \tilde{B}(g_n(z_n), \mu, \pi)}{\partial \pi_k} = \sum_{z_n=1}^K g_n(z_n) \frac{\bar{z}_{nk}^{-1}}{\pi_k} + \lambda$$

$$-\lambda \pi_k = \sum_{z_n=1}^K g_n(z_n)$$

$$-\lambda \sum_{k=1}^K \pi_k = \sum_{k=1}^K \sum_{z_n=1}^K g_n(z_n)$$

$$\boxed{\lambda = -N}$$

$$\therefore \pi_k = \frac{1}{N} \sum_{z_n=1}^K g_n(z_n)$$

\equiv
Ans.