Collaborators :

- Victor Zuanazzi

ML - 2

Homework Assignment - 6

Shantanu Chandra
(12048461)

Q1.

a) Rejection Sampling algo :

```
1.  n ← 0 , samples ← [ ]
2.  while n < N :
3.      x_i ~ q(x)
4.      range = c · q̃(x_i)
5.      refer = p̃(x_i)
6.      u_i ~ U[0, range]
7.      if u_i > refer :
8.          samples ← samples ∪ x_i
9.      n ← n+1
```

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

c) $w_n = \frac{\tilde{p}(x_n)}{\tilde{q}(x_n)} = \frac{Z_p\, p(x_n)}{Z_q\, q(x_n)}$ // Ans.

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

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

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

$$= \min\left(1, \frac{p(x_{t+1})}{p(x_t)} \cdot \frac{q(x_{t+1}|x_t) \cdot q(x_t)}{q(x_{t+1})^{\not 2}}\right) \quad (\text{using } q(x_{t+1}|x_t) = q(x_{t+1}) \;\&\; \text{using joint of } \tilde{q}(x_t, x_{t+1}))$$

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

// Ans.

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

// Ans.

f) Independence Sampler produces : $[x_1, x_1, x_3, x_4, x_4] = [0.34, 0.34, 2.67, 0.82, 0.82]$

// Ans.

g) ① Rejection Sampling : if $p(x)$ is hi dimensional then we need to find $q(x)$ such that $kq(z) \geq p(z)$ $\forall z$. Thus, first limitation comes from choosing a suitable $q(z)$ in high dimensions. Secondly, the acceptance rate is ratio of volumes under $p(z)$ & $kq(x)$. Since $p(x)$ & $kq(x)$ are normalized, it is $1/k$. 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 (Eg: $p(x) \gg 0$, $q(x) \approx 0$ = 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

Q2.

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 $p(\mu, \tau | x) \rightarrow p(\mu | \tau, x)$, $p(\tau | \mu, x)$

For τ:

$$p(\tau | \mu, x) = \frac{p(\mu, x | \tau)\, p(\tau)}{p(x, \mu)} = \frac{p(x | \mu, \tau)\, p(\mu | \tau)\, p(\tau)}{p(x | \mu)\, p(\mu)} \qquad \left[\because \mu \text{ \& } \tau \text{ are d-separated given } x\right]$$

$$= \frac{p(x | \mu, \tau)\, p(\tau)}{p(x | \mu)} \propto p(x | \mu, \tau)\, p(\tau)$$

Here, $p(x | \mu, \tau) = \mathcal{N}(x | \mu, \tau^{-1})$

$p(\tau) = p(\tau | a, b) = \text{Gamma}(\tau | a, b)$

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

$$p(\tau | \mu, x) \propto \underbrace{\sqrt{\frac{\tau}{2\pi}} \exp\left(\frac{-\tau}{2}(x-\mu)^2\right)} \cdot \underbrace{\frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau)}$$

$$= \frac{b^a}{\Gamma(a)} \tau^{a-1+\frac{1}{2}} \exp\left(-\tau\left[\frac{(x-\mu)^2}{2} + b\right]\right)$$

$$= \text{Gamma}\left(\tau \mid a + \tfrac{1}{2}, \frac{(x-\mu^2)}{2} + b\right)$$

= Ans.

For μ:

Similarly for μ, we have multiplication of 2 Gaussians:

$$p(\mu | x, \tau) = p(x | \mu, \tau) \cdot p(\mu | \mu_0, s_0)$$
$$= \mathcal{N}(x | \mu, \tau^{-1}) \cdot \mathcal{N}(\mu | \mu_0, s_0)$$
$$= \mathcal{N}\left(\mu \mid \left(\tau + \frac{1}{s_0}\right)^{-1}\left(x\tau + \frac{\mu_0}{s_0}\right), \tau + \frac{1}{s_0}\right) \qquad \text{(using 2.141 \& 2.142 of Bishop)}$$

= Ans.

Q3.

1) The joint probability is given by:

$$p(w, z, \phi, \theta \mid \alpha, \beta) = p(\theta \mid \alpha)\, p(z \mid \theta)\, p(\phi \mid \beta)\, p(w \mid \phi, z) \qquad \left[\begin{array}{l} \alpha = (\alpha, \dots \alpha) \\ \beta = (\beta, \dots \beta) \end{array}\right]$$

$$= \prod_{d=1}^{D} Dir(\theta \mid \alpha) \prod_{k=1}^{K} Dir(\phi_k \mid \beta) \prod_{n=1}^{N_d} Mult(z_{dn} \mid \theta_d)\, Mult(w_{dn} \mid \phi_{dn}, z_{dn})$$

= ans.

2)

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

$$= \prod_{d=1}^{D} \prod_{n=1}^{N_d} \prod_{k=1}^{K} \prod_{w} p(w_{dn} \mid z_{dn}, \phi_{kw})^{\delta(z_{nd}=k)\,\delta(w_{nd}=w)}$$

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

$$= \prod_{k=1}^{K} \prod_{w} \phi_{kw}^{B_{kw}}$$

$$p(z \mid \theta) = \prod_{n=1}^{N_d} \prod_{d=1}^{D} p(z_{nd} \mid \theta)$$

$$= \prod_{n=1}^{N_d} \prod_{d=1}^{D} \prod_{k=1}^{K} p(z_{nd}=k \mid \theta_{dk})^{\delta(z_{nd}=k)}$$

$$= \prod_{k=1}^{K} \prod_{d=1}^{D} \underbrace{p(z_{nd}=k \mid \theta_{dk})}_{Mul(\theta_{dk})}^{\underbrace{\sum_{n=1}^{N_d} \delta(z_{nd}=k)}_{A_{dk}}}$$

$$= \prod_{k=1}^{K} \prod_{d=1}^{D} \theta_{dk}^{A_{dk}}$$

Using the above 2 results, we evaluate the integrals:

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

$$\int_{\theta} p(\theta \mid \alpha) p(z \mid \theta)\, d\theta = \prod_{d=1}^{D} \int_{\theta_d} \underbrace{p(\theta_d \mid \alpha)}_{} \prod_{n=1}^{N_d} \underbrace{p(z_{nd} \mid \theta_d)}_{}\, d\theta$$

$$= \prod_{d=1}^{D} \int_{\theta_d} \underbrace{\frac{1}{B(\alpha)} \theta_{dk}^{\alpha-1}}_{} \cdot \prod_{k=1}^{K} \theta_{dk}^{A_{dk}}\, d\theta_d$$

$$= \prod_{d=1}^{D} \int_{\theta_d} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{dk}^{\alpha + A_{dk} - 1}\, d\theta_d$$

$$= \prod_{d=1}^{D} \frac{B(A_d + \alpha)}{B(\alpha)} \int_{\theta_d} \frac{1}{B(A_d + \alpha)} \theta^{A_{dk} + \alpha - 1}\, d\theta_d$$

$$= \prod_{d=1}^{D} \frac{B(A_d + \alpha)}{B(\alpha)} \int_{\theta_d} \underbrace{Dir(\theta_d \mid A_{dk} + \alpha)}_{=1 \text{ (area of a distribution)}}$$

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

$$\int_{\phi} p(\phi|\beta)\, p(w|\phi, z)\, d\phi = \prod_{k=1}^{K} \int_{\phi_k} p(\phi_k|\beta) \prod_{n=1}^{N_d} \prod_{d=1}^{D} p(w_{nd}|z_{nd}, \phi_{nd})\, d\phi_k$$

$$= \prod_{k=1}^{K} \int_{\phi_k} \frac{1}{B(\beta)} \phi_{kw}^{\beta-1} \cdot \prod_{w} \phi_{kw}^{B_{kw}}\, d\phi_k$$

$$= \prod_{k=1}^{K} \frac{B(B_k + \beta)}{B(\beta)} \int_{\phi_k} \frac{1}{B(B_k + \beta)} \prod_{w} \phi_{kw}^{\beta + B_{kw} - 1}\, d\phi_k$$

$$= \prod_{k=1}^{K} \frac{B(B_k + \beta)}{B(\beta)} \underbrace{\int_{\phi_k} \mathrm{Dir}(\phi_k | B_k + \beta)\, d\phi_k}_{=1}$$

$$= \prod_{k=1}^{K} \frac{B(B_k + \beta)}{B(\beta)} \quad \text{— (2)}$$

Finally, combining (1) & (2):

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

And.

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

$$p(z_i | z_{\setminus i}, w) = \frac{p(w, z)}{p(w, z_{\setminus i})} \qquad (\text{where } \setminus i \text{ represents all } z \text{ except } z_i)$$

$$= \frac{p(w, z)}{p(w_{\setminus i}, z_{\setminus i})\, \underbrace{p(w_{\setminus i})}_{\text{= constant}}}$$

Using results from previous question:

$$= \prod_{d=1}^{D} \frac{B(A_d + \alpha)}{B(A_d^{(-i)} + \alpha)} \prod_{k=1}^{K} \frac{B(B_k + \beta)}{B(B_k^{(-i)} + \beta)} \cdot \frac{1}{p(w_{\setminus i})}$$

(Here, i = dimens. of z = nd)

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

And.

Q4.

a) $\mathbb{E}[x] = \sum_{x_i \in \{0,1\}} x_i \, p(x_i) = 1 \cdot \mu_i + 0 \cdot (1-\mu_i) = \vec{\mu}$ Ans.

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

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

c) $\mathbb{E}[x] = \sum_{x_i} p(x_i|\mu,\pi)\, x_i$

$= \sum_{x_i} \sum_{k=1}^{K} \pi_k \, p(x_i|\mu_k)\, x_i$

$= \sum_{k=1}^{K} \pi_k \sum_{x_i} p(x_i|\mu_k)\, x_i$

$= \sum_{k=1}^{K} \pi_k \cdot \mu_k$ (using result from previous part)
Ans.

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

$\log p(x_n|\mu,\pi) = \log \sum_{k=1}^{K} \left(\pi_k \, p(x_n|\mu_k)\right)$

$= \log\left[\sum_{k=1}^{K} \left(\pi_k \prod_{i=1}^{D} \left[\mu_{ki}^{x_{ni}} (1-\mu_{ki})^{1-x_i}\right]\right)\right]$

$\therefore \log p(X|\mu,\pi) = \log\left[\prod_{n=1}^{N}\left[\sum_{k=1}^{K}\left(\pi_k \prod_{i=1}^{D}\left[\mu_{ki}^{x_{ni}}(1-\mu_{ki})^{1-x_{ni}}\right]\right)\right]\right]$

$= \sum_{n=1}^{N} \log\left[\sum_{k=1}^{K}\left(\pi_k \prod_{i=1}^{D}\left[\mu_{ki}^{x_{ni}}(1-\mu_{ki})^{1-x_{ni}}\right]\right)\right]$
Ans.

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.

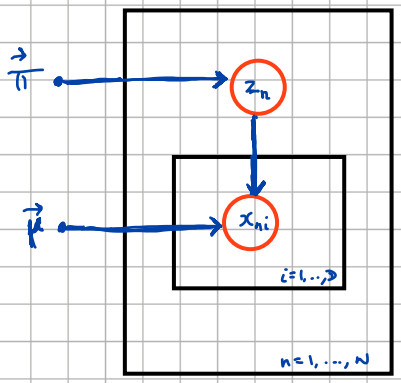
f) complete log data likelihood :

$\log p(X,Z|\mu,\pi) = \log\left[\prod_{n=1}^{N} p(z_n|\pi)\, p(x_n|z_n,\mu)\right]$

$= \sum_{n=1}^{N}\left[\log\left(\prod_{k=1}^{K}\left[\pi_k^{z_{nk}} \cdot \left(\prod_{i=1}^{D} \mu_{ki}^{x_{ni}}(1-\mu_{ki})^{1-x_{ni}}\right)^{z_{nk}}\right]\right)\right]$

$= \sum_{n=1}^{N}\sum_{k=1}^{K}\left[z_{nk}\log\pi_k + z_{nk}\sum_{i=1}^{D}\left(x_{ni}\log\mu_{ki} + (1-x_{ni})\log\mu_{ki}\right)\right]$
Ans.

g) Data likelihood : $p(x_n, z_n | \mu, \pi) = p(z_n|\pi)\, p(x_n|z_n, \mu)$



z_n = latent / unobserved
x_n = observed
μ, π = parameters
$\vec{\pi}$ is one-of-K vector.
$\vec{\mu}$ is K×D matrix.

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

$$\ln p(x|\theta) = L(q,\theta) + KL(q\|p) \geq L(q,\theta)$$

$L(q,\theta)$: ELBO

where, $L(q,\theta) = \sum_z q(z) \ln\left[\frac{p(x,z|\theta)}{q(z)}\right]$ (Bishop 9.70 - 9.72)

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

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

$$\mathcal{B}(q_n(z_n), \mu, \pi) = \sum_{n=1}^{N} \sum_z q_n(z_n) \ln \frac{p(x_n, z_n|\mu,\pi)}{q_n(z_n)}$$

$$= \sum_{n=1}^{N} \sum_z \left[q_n(z_n) \log p(x_n, z_n|\mu,\pi) - q_n(z_n) \log_z q_n(z_n) \right]$$

$$= \sum_{n=1}^{N} \sum_{z_n=1}^{K} \left[q_n(z_n) \sum_{k=1}^{K} \left[z_{nk} \log \pi_{z_n} + z_{nk} \sum_{i=1}^{D} \left(x_{ni} \log \mu_{ki} + (1-x_{ni}) \log \mu_{ki} \right) \right] \right] - \sum_{n=1}^{N} \sum_{z_n=1}^{K} q_n(z_n) \log q_n(z_n)$$

Ans.

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

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

Ans.

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

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

($z_{nk} = 1$)

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

$$\log q_n(z_n) = \lambda_n - 1 + \log \pi_k + \sum_{i=1}^{D} \left(x_{ni} \log \mu_{ki} + (1-x_{ni}) \log (1-\mu_{ki}) \right)$$

$$q_n(z_n) = \exp[\lambda_n - 1] \cdot \pi_k \prod_{i=1}^{D} \mu_{ki}^{x_{ni}} (1-\mu_{ki})^{(1-x_{ni})} \quad \text{— (1)}$$

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

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

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

Substituting this back in (1), we get:

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

Ans.

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

Ans.

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

For μ:

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

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

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

Ans.

For π:

$$\frac{\partial \tilde{B}(q_n(z_n), \mu, \pi)}{\partial \pi_k} = \sum_{z_n=1}^{K} q_n(z_n) \frac{\overbrace{z_{nk}}^{=1}}{\pi_k} + \lambda$$

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

$$-\lambda \underbrace{\sum_{k=1}^{K} \pi_k}_{=1} = \underbrace{\sum_{k=1}^{K} \sum_{z_n=1}^{K} q_n(z_n)}_{=N}$$

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

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

Ans.