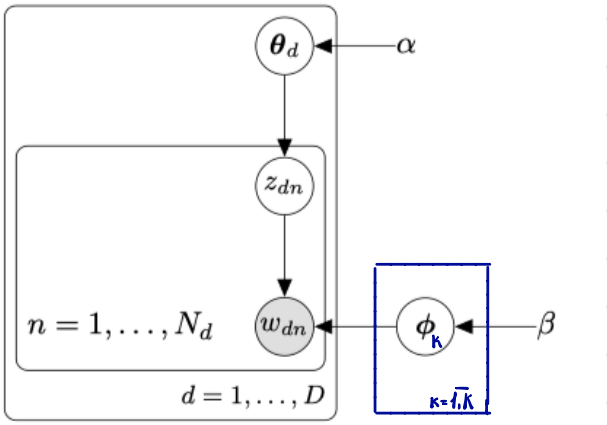
ML2 Home Assignment 6

id: 12179078

Volodymyr Medentsiy

Problem 3



Topics $\varphi_k \in \mathbb{R}^{|V|} \sim Dir(\beta, \ldots, \beta)$ $\forall k = \overline{1,K}$

Documents $w_d \in D$:

a) Topic distribution $\theta_d \sim Dir(\alpha, \ldots, \alpha)$

b) Words w_{dn} in document d:

i) $z_{dn} \sim Mult(\theta_d)$

ii) $w_{dn} | z_{dn}, \varphi_{z_{dn}} \sim Mult(\varphi_{z_{dn}})$

$$A_{dk} = \sum_{n=1}^{N_d} \delta(z_{dn} = k)$$

$$B_{kw} = \sum_{d=1}^{D} \sum_{i=n}^{N_d} \delta(w_{dn} = w)\, \delta(z_{dn} = k)$$

$$M_k = \sum_{w} B_{kw}$$

1. The joint over observed data and latent:

$$P_{joint} = p\left(\{\{w_{dn}\}_{n=1}^{N_d}\}_{d=1}^{D}, \{\theta_d\}_{d=1}^{D}, \{\varphi_k\}_{k=1}^{K}, \{\{z_{dn}\}_{n=1}^{N_d}\}_{d=1}^{D}\right) =$$

$$= \prod_{k=1}^{K} p(\varphi_k|\beta) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p(w_{dn}|z_{dn}, \varphi_{z_{dn}}) \cdot p(z_{dn}|\theta_d)\, p(\theta_d|\alpha) =$$

$$= \prod_{k=1}^{K} p(\varphi_k|\beta) \cdot \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(w_{dn}|z_{dn}, \varphi_{z_{dn}}) \cdot p(z_{dn}|\theta_d) \quad (=)$$

$$p(\varphi_k|\beta) = \frac{1}{B(\vec{\beta})} \cdot \prod_{w \in V} \varphi_{kw}^{\beta-1}$$

$$p(\theta_d|\alpha) = \frac{1}{B(\vec{\alpha})} \prod_{k=1}^{K} \theta_{dk}^{\alpha-1}$$

$$p(z_{dn}|\theta_d) = \prod_{k=1}^{K} \theta_{dk}^{[z_{dn}=k]}$$

$$p(w_{dn}|z_{dn}, \varphi_{z_{dn}}) = \prod_{w \in V} \varphi_{z_{dn}w}^{[w_{dn}=w]}$$

So $\prod_{d=1}^{D} \prod_{n=1}^{N_d} p(w_{dn}|z_{dn}, \varphi_{z_{dn}}) = \prod_{w \in V} \varphi_{kw}^{B_{kw}}$, where φ_{kw} is the w-th component of the topic k

$\prod_{d=1}^{D} \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) = \prod_{d=1}^{D} \theta_{dk}^{A_{dk}}$, where θ_{dk} is the k-th component in a distribution of topic d.

$$(=) \prod_{k=1}^{K} \frac{1}{B(\vec{\beta})} \cdot \prod_{w \in V} \varphi_{kw}^{\beta-1} \cdot \prod_{d=1}^{D} \frac{1}{B(\vec{\alpha})} \cdot \prod_{k=1}^{K} \theta_{dk}^{\alpha-1} \cdot \prod_{w \in V} \varphi_{kw}^{B_{kw}}\, \theta_{dk}^{A_{dk}} =$$

$$= \frac{1}{B^K(\vec{\beta}) \cdot B^D(\vec{\alpha})} \cdot \prod_{k=1}^{K} \prod_{w \in V} \varphi_{kw}^{B_{kw}+\beta-1} \cdot \prod_{d=1}^{D} \theta_{dk}^{A_{dk}+\alpha-1}$$

2. $p\left(\{\{w_{dn}\}_{n=1}^{N_d}\}_{d=1}^{D}, \{\{z_{dn}\}_{n=1}^{N_d}\}_{d=1}^{D}\right) = \int_0^1\int_0^1 p_{joint}\, d\theta_{dk}\, d\varphi_{kw} =$

$$\int_0^1\int_0^1 \frac{1}{B^K(\vec{\beta})\cdot B^D(\vec{\alpha})} \cdot \prod_{k=1}^{K}\prod_{w\in V} \varphi_{kw}^{B_{kw}+\beta-1} \cdot \prod_{d=1}^{D} \theta_{dk}^{A_{dk}+\alpha-1}\, d\theta_{dk}\, d\varphi_{kw} =$$

$$= \frac{1}{B^K(\vec{\beta})\, B^D(\vec{\alpha})} \cdot \prod_{k=1}^{K}\prod_{w\in V} \int_0^1 \varphi_{kw}^{B_{kw}+\beta-1}\, d\varphi_{kw} \prod_{d=1}^{D}\int_0^1 \theta_{dk}^{A_{dk}+\alpha-1}\, d\theta_{dk} \quad (=)$$

$$\int_0^1 \varphi_{kw}^{B_{kw}+\beta-1}\, d\varphi_{kw} = \frac{\varphi_{kw}^{B_{kw}+\beta}}{B_{kw}+\beta}\Big|_0^1 = \frac{1}{B_{kw}+\beta}$$

$$\int_0^1 \theta_{dk}^{A_{dk}+\alpha-1}\, d\theta_{dk} = \frac{\theta_{dk}^{A_{dk}+\alpha}}{A_{dk}+\alpha}\Big|_0^1 = \frac{1}{A_{dk}+\alpha}$$

$$(=) \quad \frac{1}{B^K(\vec{\beta})\, B^D(\vec{\alpha}) \prod_{k=1}^{K}\prod_{d=1}^{D}(A_{dk}+\alpha) \cdot \prod_{w\in V}(B_{kw}+\beta)}.$$

3. $p\left(z_{di} \mid \{\{w_{dn}\}_{n=1}^{N_d}\}_{d=1}^{D}, \{\{z_{dn}\}_{n=1/i}^{N_d}\}_{d=1}^{D}\right) =$

$$= \frac{p\left(\{\{w_{dn}\}_{n=1}^{N_d}\}_{d=1}^{D}, \{\{z_{dn}\}_{n=1}^{N_d}\}_{d=1}^{D}\right)}{p\left(\{\{w_{dn}\}_{n=1}^{N_d}\}_{d=1}^{D}, \{\{z_{dn}\}_{n=1/i}^{N_d}\}_{d=1}^{D}\right)} = \left[\begin{array}{l} p\left(\{\{w_{dn}\}_{n=1}^{N_d}\}_{d=1}^{D}, \{\{z_{dn}\}_{n=1/i}^{N_d}\}_{d=1}^{D}\right) = \\ = p(w_{di}) \cdot p\left(\{\{w_{dn}\}_{n=1/i}^{N_d}\}_{d=1}^{D}, \{\{z_{dn}\}_{n=1/i}^{N_d}\}_{d=1}^{D}\right) \end{array}\right] =$$

$$= \frac{B^K(\beta)\, B^D(\vec{\alpha}) \prod_{k=1}^{K}\prod_{d=1}^{D}(A_{dk}+\alpha) \prod_{w\in V/w_i}(B_{kw}+\beta)}{B^K(\vec{\beta})\, B^D(\vec{\alpha}) \prod_{k=1}^{K}\prod_{d=1}^{D}(A_{dk}+\alpha) \prod_{w\in V}(B_{kw}+\beta)} \cdot p(w_{di})$$

$$p(w_{di}) = \sum_{z_{di}} p(w_{di} \mid z_{di}, \varphi_{di}) \cdot p(z_{di}) = \sum_{z_i} Mult(w_{di} \mid \varphi_{di}) \cdot Mult(z_{di} \mid \theta_d)$$

Problem 1

a)

$\tilde{q}(x) = c\,q(x)$

$\tilde{q}(x) \geq \tilde{p}(x) \quad \forall x$

Pseudocode of Rejection Sampling:

1. Set number of samples N
2. $i = 0$
3. do while $i < N$
3.1. $z \sim q(x)$
3.2. $u \sim U[0; \tilde{q}(z)]$
3.3. if $u > \tilde{p}(z)$:
reject z
else:
accept z as a sample from $p(x)$
$i \mathrel{+}= 1$

b) Yes, samples are independent.

c) $p(x) = c_p \tilde{p}(x)$

$q(x) = c_q \tilde{q}(x)$

Using Bishop's 11.23:

$$w_e = \frac{\tilde{r}_e}{\sum_m \tilde{r}_m} = \frac{\tilde{p}(z^{(e)})/q(z^{(e)})}{\sum_m \tilde{p}(z^{(m)})/q(z^{(m)})} = \frac{p(z^{(e)})}{c_p \cdot q(z^{(e)}) \cdot \sum_m \frac{p(z^{(m)})}{c_p \cdot q(z^{(m)})}} =$$

$$= \frac{p(z^{(l)}) / q(z^{(l)})}{\sum_m p(z^{(m)}) / q(z^{(m)})}.$$

d) Independence Sampler with $q(x_{t+1}|x_t) = q(x_t)$

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

$$= \left[q(x_t|x_{t+1}) = \frac{q(x_{t+1}|x_t)\, q(x_t)}{q(x_{t+1})} = \frac{q(x_{t+1})\, q(x_t)}{q(x_{t+1})} = q(x_t) \right] =$$

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

e) Samples are drawn from a probability distribution, which is independent from the previous sample ($x_{t+1} \sim q(x_{t+1})$) but the acceptance probability depends on x_t, thus samples are not independent.

f) We will obtain such a sequence:

$$x_1, x_1, x_3, x_4, x_4.$$

In practical applications we will most probably retain only: x_1, x_3, x_4, so no duplicates.

g) Rejection and Importance sampler work very bad in high dimensional spaces. I think there are two reasons:

1. it's hard to find approximating distribution $q(x)$ such that its high-density regions will coincide with such regions in $p(x)$.

2. the approximation error, volume between $q(x)$ and $p(x)$, increases as the dimensionality of x increases (curse of dimensionality)

MCMC works better in high-dimensional spaces than rejection and importance sampling, but still its performance becomes worse with the increase in dimensionality. For example, let's consider Metropolis-Hastings algorithm, with symmetric proposal and assume that $p(x) = p(x^{(1)}) \cdot \ldots \cdot p(x^{(D)})$, $x \in \mathbb{R}^D$ then we can show how rapidly decreases acceptance probability with increase of D:

if $\frac{p(x_{t+1}^{(i)})}{p(x_t^{(i)})} = 0{,}9$ => acceptance probability for each component of x is $0{,}9$ but acceptance probability for the whole x is $0{,}9^D$. which could be $<= 1$ as D increases. This example also illustrates curse of dimensionality.

Problem 2

$x \sim N(x \mid \mu, \tau^{-1})$

$\mu \sim N(\mu \mid \mu_0, s_0)$

$\tau \sim \text{Gamma}(\tau \mid a, b)$

Derive Gibbs sampling distribution for the posterior $p(\mu, \tau \mid x)$.

1. The sampling procedure:

for $t = 1, T$

$\mu^{(t)} \sim p(\mu \mid x, \tau^{(t-1)})$

$\tau^{(t)} \sim p(\tau \mid x, \mu^{(t)})$

Let's find $p(\mu \mid x, \tau^{(t-1)})$ and $p(\tau \mid x, \mu^{(t)})$ using 2.139–2.143 and 2.149–2.151 in Bishop.

$p(\mu \mid x, \tau^{(t-1)}) = N(\mu \mid \tilde{\mu}, \tilde{s}^2)$, assume that we observe one x and $\tau^{(t-1)}$ is known

with $\tilde{\mu} = \frac{(\tau^{(t-1)})^{-1}}{s_0 + (\tau^{(t-1)})^{-1}} \mu_0 + \frac{s_0}{s_0 + (\tau^{(t-1)})^{-1}} \cdot x$

$\tilde{s}^2 = \left(\frac{1}{s_0} + \tau^{(t-1)}\right)^{-1} = \frac{s_0}{1 + s_0 \cdot \tau^{(t-1)}}$

$p(\tau \mid x, \mu^{(t)}) = \text{Gam}(\tau \mid \tilde{a}, \tilde{b})$, assume that we observe one x and $\mu^{(t)}$ is known.

with $\tilde{a} = a_0 + 1/2$

$\tilde{b} = b_0 + 1/2 \cdot (x - \mu^{(t)})$

Problem 4

Given: $p(x|\mu) = \prod_{i=1}^{D} \mu_i^{x_i} (1-\mu_i)^{1-x_i}$

$x \in [0;1]^D$, $\mu \in [0;1]^D$

a) $E\, p(x|\mu) = \mu$.

b) $Cov\, p(x|\mu) = E(x-\mu)(x-\mu)^T$:

$$E(x_i - \mu_i)(x_j - \mu_j) = E(x_i x_j - \mu_i x_j - x_i \mu_j + \mu_i \mu_j) \overset{(=)}{}$$

if $i \neq j$ $\overset{1}{(=)}$ $Ex_i Ex_j - \mu_i Ex_j - Ex_i \cdot \mu_j + \mu_i \mu_j = 0$.

if $i = j$ $\overset{2}{(=)}$ $\mu_i (1-\mu_i)$

So $Cov\, x = diag(\mu(1-\mu))$

Given: $p(x|\mu, \pi) = \sum_{k=1}^{K} \pi_k p(x|\mu_k)$ with $\pi = (\pi_1, \dots, \pi_K)$, $\mu = (\mu_1, \dots, \mu_K)$ and $\forall i: \mu_i \in [0;1]^D$. And $p(x|\mu_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i} (1-\mu_{ki})^{1-x_i}$

c) $Ex = E \sum_{k=1}^{K} \pi_k p(x|\mu_k) = \sum_{k=1}^{K} \pi_k E\, p(x|\mu_k) = \sum_{k=1}^{K} \pi_k \cdot \mu_k$.

Given: data set $X = (x_1, \dots, x_N)$

d) $$\text{Log likelihood} = \log p(X|\mu, \pi) = \log \prod_{n=1}^{N} p(x_i|\mu, \pi) = \sum_{i=1}^{N} \log p(x_i|\mu, \pi) =$$

$$= \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k p(x_i|\mu_k) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \left(\pi_k \cdot \prod_{d=1}^{D} \mu_{kd}^{x_i^d} (1-\mu_{kd})^{1-x_i^d} \right),$$

with x_i^d – d-th component of observed data point x_i.

e) We use loglikelihood because it decomposes $\log \prod$ into $\sum \log$ and

thus we can then easily derive closed-form max-likelihood solutions.
However in this mixture model we have $\sum \log \sum$ so we can't further decompose $\log \sum$ to easily derive analytical solution and will need numerical methods to optimize the log-likelihood.

Given: we introduce latent variables z_n - one-hot encoded vector

$$p(x_n, z_n | \mu, \pi) = p(z_n|\pi)\, p(x_n|z_n, \mu) = \prod_{k=1}^{K} \pi_k^{z_{nk}}\, p(x_n|\mu_k)^{z_{nk}}$$

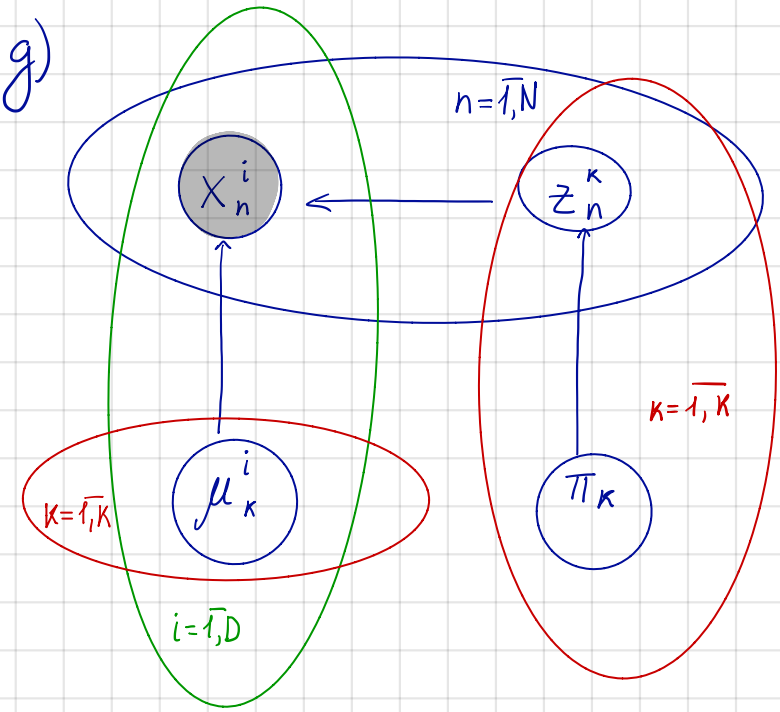
f) $$\log p(X, Z | \mu, \pi) = \log \prod_{n=1}^{N} p(x_n, z_n | \mu, \pi) =$$

$$= \sum_{n=1}^{N} \log p(x_n, z_n | \mu, \pi) = \sum_{n=1}^{N} \log p(z_n|\pi)\, p(x_n|z_n, \mu) =$$

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

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \left(z_{nk} \log \pi_k + z_{nk} \sum_{i=1}^{D} \left(x_n^i \log \mu_{ki} + (1-x_n^i) \log (1-\mu_{ki}) \right) \right)$$

g)



h) VEM objective function $B(\{q_n(z_n)\}, \mu, \pi)$

$$B(\{q_n(z_n)\}, \mu, \pi) = \sum_n \sum_{z_n} q_n(z_n) \log p(x_n, z_n | \mu, \pi) - \sum_n \sum_{z_n} q_n(z_n) \log q_n(z_n) =$$

$$= \sum_n \sum_{z_n} q_n(z_n) \Big[\sum_{k=1}^{K} \left(z_{nk} \log \pi_k + z_{nk} \sum_{i=1}^{D} \left(x_n^i \log \mu_{ki} + (1 - x_n^i) \log (1 - \mu_{ki}) \right) \right)$$

$$- \log q_n(z_n) \Big]$$

i) Constraints: $\sum_k \pi_k = 1$ and $\forall n \sum_{z_n} q(z_n) = 1$.

Then

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

j) $q_n(z_n)$ approximates the posterior $p(z_n | x_n, \mu, \pi, \tilde{\lambda}, \{\lambda_n\})$ and can be viewed as the responsibility which z_n takes in explaining x_n.

$$\frac{\partial \tilde{B}(\{q_n(z_n)\}, \mu, \pi, \tilde{\lambda}, \{\lambda_n\})}{\partial q_n(z_n)} =$$

$$= \sum_{k=1}^{K} \left(z_{nk} \log \pi_k + z_{nk} \sum_{i=1}^{D} \left(x_n^i \log \mu_{ki} + (1 - x_n^i) \log (1 - \mu_{ki}) \right) \right) -$$

$$- \log q_n(z_n) - 1 + \lambda_n = 0 \Rightarrow$$

$$\Rightarrow \log q_n(z_n) = \log \prod_{k=1}^{K} \pi_k^{z_{nk}} \cdot \left(\prod_{i=1}^{D} \mu_{ki}^{x_n^i} \cdot (1 - \mu_{ki})^{1 - x_n^i} \right)^{z_{nk}} - 1 + \lambda_n \Rightarrow$$

$$\Rightarrow q_n(z_n) = \exp(\lambda_n - 1) \cdot \prod_{k=1}^{K} \pi_k^{z_{nk}} \cdot \left(\prod_{i=1}^{D} \mu_{ki}^{x_n^i} \cdot (1 - \mu_{ki})^{1 - x_n^i} \right)^{z_{nk}}$$

k) M - step:

$$\frac{\partial \tilde{B}(\{q_n(z_n)\}, \mu, \pi, \tilde{\lambda}, \{\lambda_n\})}{\partial \pi_k} = \sum_n \sum_{z_n} q_n(z_n)\, z_{nk} \cdot \frac{1}{\pi_k} + \tilde{\lambda} = 0$$

$$\Rightarrow \quad \pi_k = \frac{-\sum_n \sum_{z_n} q_n(z_n)\, z_{nk}}{\tilde{\lambda}}$$

Let's find $\tilde{\lambda}$:

$$\tilde{\lambda} \underbrace{\sum_k \pi_k}_{\| \; 1} = -\sum_k \sum_n \sum_{z_n} q_n(z_n)\, z_{nk} \quad \Rightarrow \quad \tilde{\lambda} = -N$$

$$\Rightarrow \quad \pi_k = \frac{\sum_n \sum_{z_n} q_n(z_n)\, z_{nk}}{N}$$