Machine Learning 2 - Homework 6

Pascal M. Esser

May 13, 2018

Collaborators: Sindy Löwe, Gabriele Cesa

1

In this question we are interested in generating samples from a probability density $p(x)$ with $x \in R^d$. We are given an approximation $q(x)$ of $p(x)$. We will denote unnormalized densities as $\tilde{p}$ and $\tilde{q}$.

a) *Assume that you have a constant c such that $\tilde{q}(x) = cq(x)$ and $\tilde{q}(x)$‘$p(x), \forall x$. Describe with pseudocode the "Rejection Sampler" algorithm.* See Algorithm 1

b) *Are the samples you generate independent from each other?* Yes, samples from the current state do not depend on the last one only on the uniform it is sampled from which is independent

c) *An "Importance Sampler" accepts all samples but weights them using weights w_n. Provide the expression for w_n in terms of $p(x_n)$ and $q(x_n)$.*

$$w_n = \frac{\tilde{p}(x_n)}{\tilde{q}(x_n)} \tag{1}$$

d) *An "Independence Sampler" uses a proposal distribution of the form $q(x_{t+1}|x_t) = q(x_{t+1})$(i.e. the proposed new state is independent of the previous state) and subsequently accepts or rejects this proposed state as the next state of the Markov chain. Provide the expression for the Metropolis Hastings accept probability $\alpha(x_{t+1}, x_t)$ in terms of p and q for the Independence Sampler.*

$$\alpha(x_{t+1}, x_t) = \min\left(1, \frac{q(x_t)\tilde{p}(x_{t+1})}{p(x_t)q(x_{t+1})}\right) \tag{2}$$

e) *Are two subsequent samples from the Independence Sampler independent or dependent in general? Explain your answer.* x_{t+1} and x_t are dependent as the acceptance probability depends both on the current and previous state.

f) *Imagine we run the Independence sampler for 5 steps and during these 5 steps we propose the states x_1, x_2, x_3, x_4, x_5 (think of these represent as numeric values, e.g. 0.34, 3.5, 2.67, 0.82, 1.60). The*

Algorithm 1 Rejection Sampler

```
while len(samples)<max_len_samples do
    sample $x_i \sim \tilde{q}$
    sample $u_i \sim u(0, \tilde{q}(x_i))$
    accept $x_i$ iff $u_i < p(x_i)$ else reject
```

MCMC procedure rejects the proposals x_2 and x_5. Which sequence of states will the Independence sampler generate after 5 steps? x_1, x_1, x_3, x_4, x_4

g) *Will any of the three samplers discussed above work in high-dimensional settings (e.g., $d > 20$)? Explain your answer by discussing how this "curse of dimensionality" will affect each of the three samplers discussed above.*
rejection sampling → accepting rate decreases exponentially. Volume between surfaces in high dimensions gets exponentially bigger and therefore we cant use the approach as it relies on a tight approximation of one distribution by another one, this also holds for importance sampling
importance sampling → hard to find tight bounding distribution in higher dimensions
Independence sampler → works as it uses MC for decomposition.

2

we need expressions for $p(\mu|x,\tau)$ and $p(\tau|x,\mu)$ for Gibbs sampling from the posterior.

$$\begin{aligned}
p(\mu|x,\tau) =& \frac{p(x,\mu,\tau)}{\int p(\tau,x,\mu)d\mu} && (3)\\
=& \frac{p(x|\mu,\tau)p(\mu),p(\tau)}{p(\tau)\int p(x|\tau,\mu)p(\mu)d\mu} && (4)\\
\propto& \mathcal{N}(x|\mu,\tau^{-1})\mathcal{N}(\mu|\mu_0 s_0) && (5)\\
=& \frac{1}{\sqrt{2\pi\tau^{-1}}}\exp\left(-\frac{(x-\mu)^2}{2\tau^{-1}}\right)\frac{1}{\sqrt{2\pi s_0}}\exp\left(-\frac{(\mu-\mu_0)^2}{2s_0}\right) && (6)\\
\propto& \exp\left(\frac{1}{2}\tau(x^2-\mu^2-2x\mu)-\frac{1}{2s_0}(\mu_0^2-\mu^2-2\mu_0\mu)\right) && (7)\\
=& \exp\left(-\left(\frac{1}{2}\tau+\frac{1}{2s_0}\right)\left(\mu-\frac{\tau x-\mu_0 s_0^{-1}}{\tau+s_0^{-1}}\right)+const\right) && (8)\\
\propto& \mathcal{N}\left(\mu\left|\frac{\tau x-\mu_0 s_0^{-1}}{\tau+s_0^{-1}},\tau^{-1}+s_0\right.\right) && (9)
\end{aligned}$$

$$\begin{aligned}
p(\tau|x,\mu) =& \frac{p(x,\mu,\tau)}{\int p(\tau,x,\mu)d\tau} && (10)\\
=& \frac{p(x|\mu,\tau)p(\mu)p(\tau)}{p(\mu)\int p(x|\mu,\tau)p(\tau)d\tau} && (11)\\
\propto& \mathcal{N}(x|\mu,\tau^{-1})\text{Gamma}(\tau|a,b) && (12)\\
=& \frac{1}{\sqrt{2\pi\tau^{-1}}}\exp\left(-\frac{(x-\mu)^2}{2\tau^{-1}}\right)\frac{b^a}{\Gamma(a)}\tau^{a-1}e^{-b\tau} && (13)\\
=& \frac{1}{\sqrt{2\pi}}\tau^{\frac{1}{2}}\frac{b^a}{\Gamma(a)}\tau^{a-1}\exp\left(-\frac{1}{2}\tau((x-\mu)^2+2b)\right) && (14)\\
\propto& \text{Gamma}\left(\tau\left|a+\frac{1}{2},\frac{1}{2}((x-\mu)^2+2b)\right.\right) && (15)
\end{aligned}$$

3

3.1

Write down the joint probability over the observed data and latent variables.

$$p(\boldsymbol{w}, \boldsymbol{z}, \theta, \phi|\alpha, \beta) = p(\phi|\beta)p(\theta|\alpha)p(\boldsymbol{z}|\theta)p(\boldsymbol{w}|\phi_z) \tag{16}$$

3.2

Hint: You can use

$$p(z_{dn} = k|\boldsymbol{\theta}_d) = \theta_{dk} \tag{17}$$

$$p(w_{dn} = w|z_{dn} = k, \boldsymbol{\phi}) = \phi_{kw} \tag{18}$$

$$\tag{19}$$

Which gives

$$\prod_{n=1}^{N_d} p(z_{dn}|\boldsymbol{\theta}_d) = \prod_{k=1}^{K} \theta_{dk}^{A_{dk}} \tag{20}$$

$$and \tag{21}$$

$$\prod_{d=1}^{D}\prod_{n=1}^{N_d} p(w_{dn}|z_{dn}, \boldsymbol{\phi}) = \prod_{k=1}^{K}\prod_{w} \phi_{kw}^{B_{kw}} \tag{22}$$

$$\tag{23}$$

Integrate out the parameters $\boldsymbol{\theta}_d$'s and $\boldsymbol{\phi}_k$'s from the joint probability. Express this result in terms of the counts N_d, M_k, A_{dk}, and B_{kw}.

If we want to integrate out the parameters $\boldsymbol{\theta}_d$'s and $\boldsymbol{\phi}_k$'s we can write:

$$p(\boldsymbol{w}, \boldsymbol{z}|\alpha, \beta) = \iint p(\boldsymbol{w}, \boldsymbol{z}, \theta, \phi|\alpha, \beta)d\theta d\phi \tag{24}$$

$$= \iint p(\phi|\beta)p(\theta|\alpha)p(\boldsymbol{z}|\theta)p(\boldsymbol{w}|\phi_z)d\theta d\phi \tag{25}$$

$$= \int p(\boldsymbol{z}|\theta)p(\theta|\alpha)d\theta \int p(\boldsymbol{w}|\phi_z)p(\phi|\beta)d\phi \tag{26}$$

Now solving the two integrals for θ and ϕ separately:

$$\begin{aligned}
\int p(\boldsymbol{z}|\theta)p(\theta|\alpha)d\theta &= \int \prod_{d}^{D}\prod_{n}^{N} p(z_{dn}|\boldsymbol{\theta}_d)\prod_{i} p(\boldsymbol{\theta}_i|\alpha)d\theta_d \qquad (27)\\
&= \int \prod_{d}^{D}\prod_{k}^{K} \theta_{dk}^{A_{dk}}\frac{1}{B(\alpha)}\prod_{k}\theta_{dk}^{\alpha_k-1}d\theta_d \qquad (28)\\
&\text{with } B(a) = \frac{\prod_K \Gamma(a_k)}{\Gamma(\sum_k a_k)} \qquad (29)\\
&= \int \prod_{d}^{D}\frac{1}{B(\alpha)}\prod_{k}^{K}\theta_{dk}^{A_{dk}}\prod_{k}\theta_{dk}^{\alpha_k-1}d\theta_d \qquad (30)\\
&= \int \prod_{d}^{D}\frac{1}{B(\alpha)}\prod_{k}^{K}\theta_{dk}^{A_{dk}+\alpha_k-1}d\theta_d \qquad (31)\\
&= \prod_{d}^{D}\frac{B(\alpha+\boldsymbol{A}_d)}{B(\alpha)}\int\frac{1}{B(\alpha)+\boldsymbol{A}_d}\prod_{k}^{K}\theta_{dk}^{A_{dk}+\alpha_k-1}d\theta_d \qquad (32)\\
&= \prod_{d}^{D}\frac{B(\alpha+\boldsymbol{A}_d)}{B(\alpha)}\int Dir(\boldsymbol{\theta}_d|\alpha+A_{d1},....,\alpha+A_{dK})d\theta_d \qquad (33)\\
&= \prod_{d}^{D}\frac{B(\alpha+\boldsymbol{A}_d)}{B(\alpha)} \qquad (34)
\end{aligned}$$

similar to the approach above we rewrite:

$$\begin{aligned}
\int p(\boldsymbol{w}|\phi_z)p(\phi|\beta)d\phi &= \int\prod_{k}^{K}\left(\frac{1}{B(\beta)}\prod_{v\in V}\phi_{kv}^{\beta-1}\right)\left(\prod_{v\in V}\phi_{kv}^{B_{kv}}\right)d\phi \qquad (35)\\
&= \int\prod_{k}^{K}\left(\frac{1}{B(\beta)}\prod_{v\in V}\phi_{kv}^{B_{kv}+\beta-1}\right)d\phi \qquad (36)\\
&= \prod_{k}^{K}\frac{B(\beta+B_k)}{B(\beta)}\int\frac{1}{B(\beta+B_k)}\prod_{v\in V}\phi_{kv}^{B_{kv}+\beta-1}d\phi \qquad (37)\\
&= \prod_{k}^{K}\frac{B(\beta+B_k)}{B(\beta)}\int Dir(\phi_k|\beta+B_{Kv_1},...,\beta+N_{kv_{|V|}})d\phi \qquad (38)\\
&= \prod_{k}^{K}\frac{B(\beta+B_k)}{B(\beta)} \qquad (39)
\end{aligned}$$

Putting the parts together again:

$$\begin{aligned}
p(\boldsymbol{w},\boldsymbol{z}|\alpha,\beta) &= \prod_{d}^{D}\frac{B(\alpha+\boldsymbol{A}_d)}{B(\alpha)}\prod_{k}^{K}\frac{B(\beta+B_k)}{B(\beta)} \qquad (40)\\
&= \frac{1}{B(\alpha)^D}\frac{1}{B(\beta)^K}\prod_{d}^{D}B(\alpha+\boldsymbol{A}_d)\prod_{k}^{K}B(\beta+B_k) \qquad (41)
\end{aligned}$$

3.3

Derive the Gibbs sampling updates for z_{di} *with all parameters integrated out.*

$$\begin{aligned}
p(z_{di}|\boldsymbol{z}_{-di},\boldsymbol{w}) &= \frac{p(\boldsymbol{w},\boldsymbol{z}|\alpha,\beta)}{p(\boldsymbol{w},\boldsymbol{z}_{-di}|\alpha,\beta)} && (42)\\
&= \frac{\frac{B(\alpha+\boldsymbol{A}_d)}{B(\alpha)}\prod_k^K \frac{B(\beta+B_k)}{B(\beta)}}{\frac{B(\alpha+\boldsymbol{A}_{-id})}{B(\alpha)}\prod_k^K \frac{B(\beta+B_{-ik})}{B(\beta)}} && (43)\\
&= \frac{B(\alpha+\boldsymbol{A}_d)\prod_k^K B(\beta+B_k)}{B(\alpha+\boldsymbol{A}_{-id})\prod_k^K B(\beta+B_{-ik})} && (44)
\end{aligned}$$

4

a)

$$\mathbf{E}[x_i] = \sum_{x_i \in [0,1]} x_i \mu_i^{x_i}(1-\mu_i)^{1-x_i} = \mu_i \tag{45}$$

$$\mathbf{E}[\boldsymbol{x}] = \boldsymbol{\mu} \tag{46}$$

b) if $i \neq j, x_i \perp\!\!\!\perp x_j \Rightarrow \Sigma_{ij} = 0, \Sigma_{ii} = \mu_i(1-\mu_i)$

c) using a) and linearity of expectation

$$\mathbf{E}[x_i] = \sum_i x_i \sum_k \pi_k p(x_i|\mu_{ki}) = \sum_k \pi_k \mu_{ki} \tag{47}$$

d)

$$\ln p(\boldsymbol{X}|\mu,\pi) = \sum_n \ln p(\boldsymbol{x}_n|\boldsymbol{\mu},\boldsymbol{\pi}) = \ln \prod_n p(\boldsymbol{x}_n|\boldsymbol{\mu},\boldsymbol{\pi}) \tag{48}$$

$$= \sum_n \ln \sum_k \pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k) \tag{49}$$

$$= \sum_n \ln \sum_k \pi_k \prod_i \mu_{ki}^{x_{ni}} \left((1-\mu_{ki})^{1-x_{ni}}\right) \tag{50}$$

e) because of the sum in the log we can not find a closed form solution

f)

$$\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\mu},\boldsymbol{\pi}) = \ln \prod_n \prod_k \pi_k^{z_{nk}} p(\boldsymbol{x}|\boldsymbol{\mu}_k)^{z_{nk}} \tag{51}$$

$$= \sum_n \sum_k z_{nk} \left(\ln \pi_k + \sum_i x_{ni} \ln \mu_{ki} + (1-x_{ni}) \ln(1-\mu_{ki}) \right) \tag{52}$$

g) See Figure 1

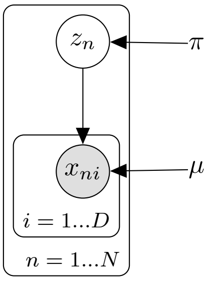

Figure 1: corresponding graphical model using plate notation

h)

$$\begin{aligned}
\mathcal{B} &= \sum_n \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \ln p(\boldsymbol{x}_n, \boldsymbol{z}_n | \boldsymbol{\mu}, \boldsymbol{\pi}) - \sum_n \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \ln q_n(\boldsymbol{z}_n) && (53)\\
&= \sum_n \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \sum_k z_{nk} \left(\ln \pi_k + \sum_i x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki}) \right) && (54)\\
&\quad - \sum_n \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \ln q_n(\boldsymbol{z}_n) && (55)
\end{aligned}$$

i)

$$\tilde{\mathcal{B}} = \mathcal{B} + \lambda \left(\sum_k \pi_k - 1 \right) + \sum_n \lambda_n \left(\sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) - 1 \right) \tag{56}$$

j)

$$\begin{aligned}
\frac{\partial \tilde{\mathcal{B}}}{\partial q(\boldsymbol{z}_{nk})} &= \ln \pi_k + \sum_i x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki}) + - \ln q_n(\boldsymbol{z}_{nk}) + \lambda_n - 1 = 0 && (57)\\
q_n(\boldsymbol{z}_{nk}) &= \pi_k \prod_i \left(\mu_{ki}^{x_{ni}} \left(1 - \mu_{ki}^{1 - x_{ni}}\right) exp(\lambda_n - 1) \right) && (58)\\
\frac{\partial \tilde{\mathcal{B}}}{\partial \lambda_n} &= \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) = 0 && (59)\\
&= -1 + \sum_k exp(\lambda_n - 1) \pi_k \prod_i \mu_{ki}^{x_{ni}} \left(1 - \mu_{ki}^{1 - x_{ni}}\right) && (60)\\
\lambda_n &= 1 - \ln \sum_k \pi_k \prod_i \mu_{ki}^{x_{ni}} \left(1 - \mu_{ki}^{1 - x_{ni}}\right) && (61)\\
q_n(\boldsymbol{z}_{nk}) &= \frac{\pi_k \prod_i \mu_{ki}^{x_{ni}} \left(1 - \mu_{ki}^{1 - x_{ni}}\right)}{\sum_j \pi_j \prod_i \mu_{ji}^{x_{ni}} \left(1 - \mu_{ji}^{1 - x_{ni}}\right)} && (62)\\
&= \frac{\pi_k p(\boldsymbol{x}_n | \mu_k)}{\sum_j \pi_j p(\boldsymbol{x}_n | \mu_j)} && (63)\\
&= \frac{\prod_k \pi_k^{z_{ik}} p(\boldsymbol{x}_n | \mu_k)^{z_{ik}}}{\sum_j \pi_j p(\boldsymbol{x}_n | \mu_j)} && (64)\\
&= \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n | \boldsymbol{\mu}_k, \boldsymbol{\pi})}{\sum_j \pi_j p(\boldsymbol{x}_n | \boldsymbol{\mu}_j)} && (65)\\
&= \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n | \boldsymbol{\mu}, \boldsymbol{\pi})}{p(\boldsymbol{x}_n | \boldsymbol{\mu}, \boldsymbol{\pi})} && (66)\\
&= p(\boldsymbol{z}_n | \boldsymbol{x}_n, \boldsymbol{\mu}, \boldsymbol{\pi}) && (67)
\end{aligned}$$

we can therefore interpret the E-Step as minimizing the KL-Divergence between q and p.

k)

$$\frac{\partial \tilde{\mathcal{B}}}{\partial q(\pi_k} = \sum_n \sum_{z_n} \frac{q(z_{nk})}{\pi_k} + \lambda = 0 \tag{68}$$

$$\pi_k = -\frac{\sum_n \sum_{z_n} q(z_{nk})}{\lambda} \tag{69}$$

$$\sum_k \pi_k \sum_k - \frac{\sum_i \sum_{z_n} q(z_{ni})}{\lambda} \tag{70}$$

$$\lambda = -\sum_k \sum_i \sum_{z_n} q(z_{ni}) \tag{71}$$

$$\pi_k = \frac{\sum_n \sum_{z_n} q(z_{nk})}{\sum_k \sum_i \sum_{z_n} q(z_{ni})} \tag{72}$$

$$\tag{73}$$

5

From the task we get directly: $\mathbf{E}[M_t] = 0, \mathbf{E}[M_t^2] = 0.5$. we set $Z^{(r)} = \sum_t M_t$ and: $p(M_t = -1) = 0.25, p(M_t = 0) = 0.5, p(M_t = 1) = 0.25$.

$$\mathbf{E}\left[\left(Z^{(r)}\right)^2\right] = \mathbf{E}\left[\left(\sum_t^r M_t\right)^2\right] = \sum_t \sum_s \mathbf{E}\left[M_t M_s\right] = \sum_t \mathbf{E}\left[M_t^2\right] = \sum_t \frac{1}{2} = \frac{r}{2} \tag{74}$$