Deep Learning 2023040163 Zhao Han Hong

Homework 3

1 True or False Questions

Problem 1

False.

Problem 2

True.

2 Q & A

Problem 3

We have the variance being

$$\begin{aligned}\mathrm{Var}\left[\frac{1}{N}\sum_{x\sim q}\frac{p(x)}{q(x)}f(x)\right] &= \frac{1}{N^2}\cdot N\cdot \mathrm{Var}_{x\sim q}\left[\frac{p(x)}{q(x)}f(x)\right] \\ &= \frac{1}{N}\left(\int \frac{p(x)^2 f(x)^2}{q(x)}dx - \left(\int p(x)f(x)dx\right)^2\right).\end{aligned}$$

Since

$$\int \frac{p(x)^2 f(x)^2}{q(x)}dx \int q(x)dx \geq \left(\int p(x)|f(x)|dx\right)^2,$$

and the equality holds if and only if $q(x) \propto p(x)|f(x)|$, we know that the variance is minimized when $q(x) \propto p(x)|f(x)|$.

Problem 4

(1) Suppose that the sampling gives

$$T(s \to s') = c\exp\left(-\frac{(s-s')^2}{\sigma^2}\right),$$

we can immediately find that the Markov Chain satisfies the detailed balance property since

$$\frac{T(s \to s')}{T(s' \to s)} = \frac{\alpha(s \to s')}{\alpha(s' \to s)} = \frac{\min\left(1, \frac{p(s')}{p(s)}\right)}{\min\left(1, \frac{p(s)}{p(s')}\right)} = \frac{p(s')}{p(s)}$$

Moreover, we can check the ergodicity of the Markov Chain by checking that

$$\min_z \min_{\pi(z')>0} \frac{T(z \to z')}{\pi(z')} = \min_{z,z',p(z')\neq 0} \frac{q(z \to z') \min\left(1, \frac{p(z')}{p(z)}\right)}{p(z')}.$$

Now, notice that

$$\min_{z,z',p(z')\neq 0} \frac{q(z \to z')}{p(z')} = \min_{z,z',p(z')\neq 0} \frac{c \exp\left(-\frac{(z-z')^2}{2\sigma^2}\right)}{p(z')} > 0,$$

$$\min_{z,z',p(z')\neq 0} \frac{q(z \to z')\frac{p(z')}{p(z)}}{p(z')} = \min_{z,z',p(z')\neq 0} \frac{q(z' \to z)}{p(z)} > 0,$$

we can know that

$$\min_{z,z',p(z')\neq 0} \frac{q(z \to z') \min\left(1, \frac{p(z')}{p(z)}\right)}{p(z')} > 0.$$

Thus, it is a valid Markov chain.

(2) Since the way of updating is

$$q(s_i \to s_i') = p(s_i'|s_{j\neq i}), \alpha(s_i \to s_i') = \min\left(1, \frac{p(s_i')q(s_i' \to s_i)}{p(s_i)q(s_i \to s_i')}\right),$$

and the other parts of the algorithms are the same, we know that Gibbs sampling is a case of Metropolis-Hasting sampling.

Now, we only have to calculate the acceptance rate. In fact, we have

$$\frac{p(s_i')q(s_i' \to s_i)}{p(s_i)q(s_i \to s_i')} = \frac{p(s_i')}{p(s_i)} \frac{p(s_i|s_{j\neq i}')}{p(s_i'|s_{j\neq i})} = \frac{p(s_i')}{p(s_i)} \frac{p(s_i)}{p(s_i')} = 1.$$

Thus, we know that Gibbs sampling is a case of Metropolis-Hasting sampling, and the acceptance rate is always 1.

(3) We consider the transition from v to u. In the sampling step i, the per-step transition is actually

$$(u_1, u_2, \cdots, u_{i-1}, v_i, v_{i+1}, \cdots, v_n) \to (u_1, u_2, \cdots, u_{i-1}, u_i, v_{i+1}, \cdots, v_n),$$

which has the probability

$$p_{v\to u,i} = p(u_i|u_1, \cdots, u_{i-1}, v_{i+1}, \cdots, v_n)$$

by definition. We can then find

$$q(v \to u) = \prod_{i=1}^{n} p_{v\to u,i} = \prod_{i=1}^{n} \frac{p(u_1, \cdots, u_{i-1}, u_i, v_{i+1}, \cdots, v_n)}{\sum_x p(u_1, \cdots, u_{i-1}, x, v_{i+1}, \cdots, v_n)}.$$

Now, we show that the stationary distribution is indeed π. Firstly, we show a more general requirement as a counterpart of the detailed balance property in this specific scenario:

$$\begin{aligned} & \pi^{(t+1)}(s_1, s_2, \cdots, s_n) = \pi^{(t)}(s_1, s_2, \cdots, s_n) \\ \Longleftarrow \quad & \pi^{(t)}(s_1, s_2, \cdots, s_n) = p(s_1, s_2, \cdots, s_n), \end{aligned}$$

where π denotes the distribution of the coordinates for the sample s. (In other words, this means that the distribution p is stationary under the sampling process.) In fact, given that $\pi^{(t)}(s_1, s_2, \cdots, s_n) = p(s_1, s_2, \cdots, s_n)$, we can have

$$\begin{aligned} & \pi^{(t+1)}(s_1, s_2, \cdots, s_n) \\ = & \sum_u q(u \to s)\pi^{(t)}(u_1, u_2, \cdots, u_n) \\ = & \sum_u \prod_{i=1}^{n} \frac{p(s_1, \cdots, s_{i-1}, s_i, u_{i+1}, \cdots, u_n)}{\sum_x p(s_1, \cdots, s_{i-1}, x, u_{i+1}, \cdots, u_n)} \cdot p(u_1, u_2, \cdots, u_n) \\ = & \sum_{u_1, \cdots, u_n} \frac{p(s_1, u_2, \cdots, u_n)}{\sum_x p(x, u_2, \cdots, u_n)} \frac{p(s_1, s_2, u_3, \cdots, u_n)}{\sum_x p(s_1, x, u_3, \cdots, u_n)} \cdots \frac{p(s_1, \cdots, s_n)}{\sum_x p(s_1, \cdots, s_{n-1}, x)} \cdot p(u_1, u_2, \cdots, u_n) \\ = & \sum_{u_1, \cdots, u_n} \frac{p(u_1, u_2, \cdots, u_n)}{\sum_x p(x, u_2, \cdots, u_n)} \frac{p(s_1, u_2, \cdots, u_n)}{\sum_x p(s_1, x, u_3, \cdots, u_n)} \cdots \frac{p(s_1, \cdots, s_{n-1}, u_n)}{\sum_x p(s_1, \cdots, s_{n-1}, x)} \cdot p(s_1, \cdots, s_n) \\ = & \, p(s_1, \cdots, s_n) \sum_{u_n} \frac{p(s_1, \cdots, s_{n-1}, u_n)}{\sum_x p(s_1, \cdots, s_{n-1}, x)} \cdots \sum_{u_2} \frac{p(s_1, u_2, \cdots, u_n)}{\sum_x p(s_1, x, u_3, \cdots, u_n)} \sum_{u_1} \frac{p(u_1, u_2, \cdots, u_n)}{\sum_x p(x, u_2, \cdots, u_n)} \\ = & \, p(s_1, \cdots, s_n), \end{aligned}$$

where the last equation is due to summation in the sequence $u_1, u_2, \cdots, u_n$. So the property we claimed is indeed true.

We can then regard this property as "like" the detailed balance property in the Metropolis-Hastings algorithm. On the other hand, the problem already provides that the Markov chain can access all states under the sampling. Thus, we can conclude that p is the unique stationary distribution for cyclic Gibbs sampling, which is the same as the random-order sampling. (In fact, one general theorem actually states that if a Markov chain with finite state space is irreducible (hence it is recurrent), then the stationary

distribution is unique.)

Problem 5

(1)

(a) False. Similarly, as we have done on the last homework, we choose $\mathbb{P}(B|A)$ to be nonzero only when $B = A$, and $\mathbb{P}(C|B)$ to be nonzero only when $C = B$. Thus, for $\mathbb{P}(A, C)$ to be nonzero, we must have $A = C$, so A and C must not be independent.

(b) True. Given B,

$$\mathbb{P}(A, C|B) = \frac{\mathbb{P}(A, B, C)}{\mathbb{P}(B)} = \frac{1}{\mathbb{P}(B)}\mathbb{P}(A)\mathbb{P}(B|A)\mathbb{P}(C|B) = \mathbb{P}(A|B)\mathbb{P}(C|B),$$

where $\mathbb{P}(A|B), \mathbb{P}(C|B)$ are functions only depending on A, C given B.

(c) False. We have

$$\mathbb{P}(A, C|D) = \frac{\mathbb{P}(A, C, D)}{\mathbb{P}(D)}, \mathbb{P}(A|D) = \frac{\mathbb{P}(A, D)}{\mathbb{P}(D)}, \mathbb{P}(C|D) = \frac{\mathbb{P}(C, D)}{\mathbb{P}(D)}.$$

Thus, if we also choose the conditional probabilities as in (a), we can find that $\mathbb{P}(A, C|D)$ is nonzero only when $A = C$, so A and C are not independent given D.

(d) False. We have

$$\mathbb{P}(A, C|B, D) = \frac{1}{\mathbb{P}(B, D)}\mathbb{P}(B|A)\mathbb{P}(C|B)\mathbb{P}(D|C, A).$$

If we choose the functions such that $\mathbb{P}(B|A), \mathbb{P}(C|B)$ both only depend on B, and $\mathbb{P}(D|C, A)$ only depends on whether $A = C$, we can then know that $\mathbb{P}(A, C|B, D)$ only depend on whether $A = C$, which means that A and C are not independent given B, D.

(e) False. We can choose $\mathbb{P}(C|B)$ is only nonzero when $B = C$, $\mathbb{P}(D|C, A) = \mathbb{P}(D|C)$ is independent of A and is nonzero only when $D = C$. Thus, we know that $\mathbb{P}(B, D)$ is nonzero only when $D = B$. Thus, B and D are not independent.

(f) False. We may choose that both $\mathbb{P}(B|A)$ and $\mathbb{P}(D|A, C)$ don't depend on A, then the problem reduces to (e).

(g) False. We pick $\mathbb{P}(C|B) = \mathbb{P}(C)$ to be independent to B, and $\mathbb{P}(D|A,C) \overset{5}{=} \mathbb{P}(D|A)$ to be independent to C and only nonzero when $D = A$. Also, we let $\mathbb{P}(B|A)$ be nonzero only when $A = B$. Then we can know that

$$\mathbb{P}(B, D|C) = \frac{1}{\mathbb{P}(C)}\mathbb{P}(B, C, D) = \frac{1}{\mathbb{P}(C)}\mathbb{P}(B|A)\mathbb{P}(D|A)$$

is only nonzero when $B = A = D$. Thus, B, D are not independent given C.

(h) True. We have

$$\mathbb{P}(B, D|A, C) = \frac{1}{\mathbb{P}(A, C)}\mathbb{P}(A)\mathbb{P}(B|A)\mathbb{P}(D|A, C) = \mathbb{P}(B|A, C)\mathbb{P}(D|A, C),$$

where $\mathbb{P}(B|A, C), \mathbb{P}(D|A, C)$ are functions only depending on B, D given A, C.

(2) The likelihood is

$$\begin{aligned}
\mathbb{P}(B, C, D|A) &= \mathbb{P}(B|A)\mathbb{P}(C|B)\mathbb{P}(D|A, C) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(B-A)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(C-B)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(D-(C+A))^2}{2}\right) \\
&= \frac{1}{(2\pi)^{\frac{3}{2}}} \exp\left(-\frac{2A^2 + 2B^2 + 2C^2 + D^2}{2} + (D+B)(C+A) - AC\right).
\end{aligned}$$

For the posterior, we have

$$\begin{aligned}
\mathbb{P}(A|B, C, D) &= \frac{\mathbb{P}(A, B, C, D)}{\mathbb{P}(B, C, D)} \\
&= \frac{1}{\mathbb{P}(B, C, D)} \frac{1}{4\pi^2} \exp\left(-\frac{3A^2 + 2B^2 + 2C^2 + D^2}{2} + (D+B)(C+A) - AC\right),
\end{aligned}$$

and we can calculate

$$\begin{aligned}
\mathbb{P}(B, C, D) &= \int_A \mathbb{P}(B, C, D|A)\mathbb{P}(A)dA \\
&= \int_A \frac{1}{4\pi^2} \exp\left(-\frac{3A^2 + 2B^2 + 2C^2 + D^2}{2} + (D+B)(C+A) - AC\right) dA \\
&= \frac{1}{\sqrt{3}(2\pi)^{\frac{3}{2}}} \exp\left(-\frac{5}{6}(B^2 + C^2) - \frac{1}{3}D^2 + \frac{1}{3}BD + \frac{2}{3}C(B+D)\right).
\end{aligned}$$

Thus, we have

$$\mathbb{P}(A|B, C, D) = \sqrt{\frac{3}{2\pi}} \exp\left(-\frac{3}{2}\left(A - \frac{D+B-C}{3}\right)^2\right).$$

Problem 6

(1) Let the kernel be $(2k+1) \times (2k+1)$. We consider the dependencies of pixel values by inversing the calculation process and consider each round of computation separately.

In the first round, the pixel at (x_0, y_0) can be only influenced by the pixel lowest as $(x_0 + k, y_0 - 1)$. Thus, $(x_0 + 1, y_0), (x_0 + 2, y_0), \cdots, (x_0 + k, y_0)$ can not influence (x_0, y_0) in the first round. For the second round, each pixel listed above extends to a new region, so we can then replace x_0, y_0 by $x_0 + k, y_0 - 1$ in the previous argument and find that the pixel at $(x_0 + k + 1, y_0 - 1), \cdots, (x_0 + 2k, y_0 - 1)$ can't influence (x_0, y_0). Notice that $(x_0 + k + 1, y_0), \cdots, (x_0 + 2k, y_0)$ also can't influence (x_0, y_0).

We may repeat this process for times and conclude that for $y_0 - l$, the pixels $(x_0 + kl, y_0 - l), \cdots, (x_0 + (l+1)k, y_0 - l)$ can't influence (x_0, y_0). Thus, the "turning points" between the pixels that can influence (x_0, y_0) and the pixels that can't influence (x_0, y_0) is a folded line $(x_0+1, y_0) \to (x_0+k+1, y_0) \to (x_0+k+1, y_0-1) \to (x_0+2k+1, y_0-1) \to \cdots$. (More precisely, this line is the left-top-most boundary of the pixels that can't influence (x_0, y_0).) We then get a sawtooth-shaped receptive field.

(2) We can mimic the method of Gated PixelCNN, which uses both a vertical stack and a horizontal stack to calculate the generating pixels and avoid blind spots.

For each layer computation (except the first layer for which the central pixel is masked), the first step is to calculate the vertical stack. On that stack, we define a kernel of size $(2k+1) \times (k+1)$ such that the vertical stack value $v_l(x, y)$ (after l iterations) of pixel (x, y) depends on $v_{l-1}([x-k : x+k+1], [y-k : y+1])$ (Here the bracket '[]' is close on the left but open on the right). The next step is letting the final value $f_l(x, y)$ of pixel (x, y) depend horizontally on $f_{l-1}(x, y), \cdots, f_{l-1}(x-k, y)$ and the vertical stack value $v_l(x, y-1)$. In summary, our per-layer computation process is:

$$v_l(x, y) = \sum_{i=-k}^{k} \sum_{j=0}^{k} v_{l-1}(x+i, y-j) w_{ij} \qquad (w_{00} = 0);$$

$$f_l(x, y) = \sum_{i=0}^{k} f_{l-1}(x-i, y) w'_{ij} + v_l(x, y-1).$$

We now demonstrate that this computation process will not lead to blind spots while maintaining the autoregressive property. In fact, we can find that $v_l(x, y)$ depends on all $v_0(x_1, y_1)$ where $y_1 \leq y$. Thus, $f_l(x, y)$ depends on all values at (x_1, y_1) where $y_1 < y$. Moreover, we can notice that $f_l(x, y)$ depends on $f_{l-1}(x-1, y), \cdots, f_{l-1}(x-k, y)$ and hence the values at $(0, y), \cdots, (x-1, y)$. This makes sure that $f(x, y)$ depends on all the values of previous pixels. Finally, we can see that it maintains the autoregressive property since there is no way for $f_l(x, y)$ to depend on the values at pixel (x_1, y_1) where $x_1 > x, y_1 = y$ or $y_1 > y$.