Homework #3

B07902028 資工二 林鶴哲

1. First we show that conditioned on $\mathbf{x}^{(l-1)}$, $s_j^{(l)}$ is zero-mean random variables.

$$\begin{aligned}
E\left[s_j^{(l)}|\mathbf{x}^{(l-1)}\right] &= E\left[\sum_{i=1}^{d^{(l-1)}} w_{ij}x_i^{(l-1)}|\mathbf{x}^{(l-1)}\right] \\
&= \sum_{i=1}^{d^{(l-1)}} E\left[w_{ij}x_i^{(l-1)}|\mathbf{x}^{(l-1)}\right] \\
&= \sum_{i=1}^{d^{(l-1)}} x_i^{(l-1)} E\left[w_{ij}|\mathbf{x}^{(l-1)}\right] && (\text{ Since } E[x_i^{(l-1)}|\mathbf{x}^{(l-1)}] = x_i^{(l-1)}) \\
&= \sum_{i=1}^{d^{(l-1)}} x_i^{(l-1)} E[w_{ij}] && (\text{ Since } w_{ij} \text{ and } \mathbf{x}^{(l-1)} \text{ are independent}) \\
&= 0 && (\text{ Since all } E[w_{ij}^{(l)}] = 0)
\end{aligned}$$

Then, we'd show that conditioned on $\mathbf{x}^{(l-1)}$, all $s_j^{(l)}$ are independent. Let $P(event)$ denote the probability that an event (or certain value) happens.
For simplicity, let $d = d^{(l)}$. Our goal is to show that for any pair, 3-tuple, ..., d-tuple of $\{s_1^{(l)}, \dots s_d^{(l)}\}$, denoted by $(s_1, s_2, ..., s_n)$, we have

$$P(s_1 = a_1, s_2 = a_2, \dots s_n = a_n) = P(s_1 = a_1)P(s_2 = a_2)\dots P(s_n = a_n)$$

for any fixed $(a_1, \dots, a_n) \subset \mathbb{R}^n$, which is exactly the definition of independence.

Let W_j contain those $\mathbf{w}_j^{(l)}$'s (which is a $d^{(l-1)}$-dimensional vector) such that $(\mathbf{x}^{(l-1)})^T\mathbf{w}_j^{(l)} = a_j$. Then we have $P(s_j = a_j) = P(\mathbf{w}_j^{(l)} \in W_j)$. Hence

$$\begin{aligned}
P(s_1 = a_1, s_2 = a_2, \dots s_n = a_n) &= P(\mathbf{w}_1^{(l)} \in W_1, \mathbf{w}_2^{(l)} \in W_2, \dots, \mathbf{w}_n^{(l)} \in W_n) \\
&= P(\mathbf{w}_1^{(l)} \in W_1)P(\mathbf{w}_2^{(l)} \in W_2)\dots P(\mathbf{w}_n^{(l)} \in W_n) \quad (\text{Since } w_{ij}^{(l)} \text{ are independent}) \\
&= P(s_1 = a_1)P(s_2 = a_2)\dots P(s_n = a_n)
\end{aligned}$$

2.

$$\begin{aligned}
\mathrm{Var}(s_j^{(l)}) &= \mathrm{Var}\left(\sum_{i=1}^{d^{(l-1)}} w_{ij}^{(l)}x_i^{(l-1)}\right) \\
&= \sum_{i=1}^{d^{(l-1)}} \mathrm{Var}\left(w_{ij}^{(l)}x_i^{(l-1)}\right) \\
&= \sum_{i=1}^{d^{(l-1)}} \left(E\left[(w_{ij}^{(l)})^2(x_i^{(l-1)})^2\right] - E\left[w_{ij}^{(l)}x_i^{(l-1)}\right]^2\right) \\
&= \sum_{i=1}^{d^{(l-1)}} \left((0^2 + \sigma_w^2)(\overline{x}^2 + \sigma_x^2) - (0 \cdot \overline{x})^2\right) && (\text{Since } w_{ij}^{(l)} \text{ and } x_i^{(l-1)} \text{ are independent}) \\
&= d^{(l-1)}(\sigma_w^2)(\overline{x}^2 + \sigma_x^2)
\end{aligned}$$

3. Since $s_j^{(l-1)}$ is symmetric, $P(s_j^{(l-1)} \geq 0) = P(s_j^{(l-1)} < 0) = \frac{1}{2}$.
That is, $P(\max\left(s_i^{(l-1)}, 0\right) = s_i^{(l-1)}) = P(\max\left(s_i^{(l-1)}, 0\right) = 0) = \frac{1}{2}$. Thus, we have

$$\begin{aligned} E\left[(x_i^{(l-1)})^2\right] &= E\left[\max\left(s_i^{(l-1)}, 0\right)^2\right] \\ &= \frac{1}{2}E\left[(s_i^{(l-1)})^2\right] + \frac{1}{2}E\left[0\right] = \frac{1}{2}E\left[(s_i^{(l-1)})^2\right] \end{aligned}$$

4. By problem 2, write $\mathrm{Var}(s_j^{(l)})$ as

$$\begin{aligned} \mathrm{Var}(s_j^{(l)}) &= d^{(l-1)}\sigma_w^2(\sigma_x^2 + \overline{x}^2) \\ &= d^{(l-1)}\sigma_w^2 E\left[(x_i^{(l-1)})^2\right] \\ &= \frac{d^{(l-1)}}{2}\sigma_w^2 E\left[((s_i^{l-1}))^2\right] && \text{(By what we've proved in problem 3)} \\ &= \frac{d^{(l-1)}}{2}\sigma_w^2\left(E\left[((s_i^{l-1}))^2\right] + \overline{x}^2\right) && \text{(By } \overline{x} = 0\text{)} \\ &= \frac{d^{(l-1)}}{2}\sigma_w^2 \mathrm{Var}(s_i^{(l-1)}) \end{aligned}$$

5. Use the result of problem 2, we re-write the problem as

$$\begin{aligned} \mathrm{Var}(s_j^{(l)}) = d^{(l-1)}{\sigma_w}^2 E\left[(x_i^{(l-1)})^2\right] &\overset{\text{want}}{=} \mathrm{Var}(s_i^{(l-1)}) \\ &= E[(s_i^{(l-1))^2}] - E[s_i^{(l-1)}]^2 = E[(s_i^{(l-1)})^2] \end{aligned} \tag{5.1}$$

Then, compute $E\left[(x_i^{(l-1)})^2\right]$ and $E\left[(s_i^{(l-1)})^2\right]$:

$$\begin{aligned} E\left[(x_i^{(l-1)})^2\right] &= \int_{-\infty}^{\infty} x^2 f(x)dx \\ &= \int_{-\infty}^{0} x^2 f(x)dx + \int_{0}^{\infty} x^2 f(x)dx \\ &= \int_{-\infty}^{0} a^2 s^2 f(s)ds + \int_{0}^{\infty} s^2 f(s)ds \\ &= (a^2+1)\int_{0}^{\infty} s^2 f(s)ds \end{aligned}$$

$$E\left[(s_i^{(l-1)})^2\right] = 2\int_{0}^{\infty} s^2 f(s)ds$$

Then, we can write (5.1) as

$$d^{(l-1)}\sigma_w^2(a^2+1)\cancel{\int_{0}^{\infty} s^2 f(s)ds} = 2\cancel{\int_{0}^{\infty} s^2 f(s)}$$

Thus, if we set each w with zero mean and variance $\frac{2}{d^{(l-1)}(a^2+1)}$(Normal distribution is a distribution that comes in use, following the assumption of the problem.), the effect of leaky ReLU $x_j^{(l)} = \max\left(s_j^{(l)}, a \cdot s_j^{(l)}\right)$ would leads to our desired result, i.e, $\mathrm{Var}(s_j^{(l)}) = \mathrm{Var}(s_i^{(l-1)})$.

6. Observe that

$$\begin{aligned}\mathbf{v}_1 &= (1-\beta)\Delta_1 \\ \mathbf{v}_2 &= \beta\mathbf{v}_1 + (1-\beta)\Delta_2 \\ \mathbf{v}_3 &= \beta\mathbf{v}_2 + (1-\beta)\Delta_3 \\ &= \beta^2(1-\beta)\mathbf{v}_1 + \beta(1-\beta)\mathbf{v}_2 + (1-\beta)\mathbf{v}_3\end{aligned}$$

So, guess $\mathbf{v}_t = \sum_{t=1}^{T} \beta^{T-t}(1-\beta)\Delta_t$. We prove the statement by mathematical induction.
It's trivial that the summation holds for $t = 1$. When $t = k + 1$,

$$\begin{aligned}\mathbf{v}_{k+1} &= \beta\mathbf{v}_k + (1-\beta)\Delta_{k+1} \\ &= \sum_{t=1}^{k} \beta^{k-t+1}(1-\beta)\Delta_t + (1-\beta)\Delta_{k+1} && \text{(By inductive hypothesis)} \\ &= \sum_{t=1}^{k+1} \beta^{k+1-t}(1-\beta)\Delta_t\end{aligned}$$

So, $\alpha_t = \beta^{T-t}(1-\beta)$.

7. Solve $\alpha_1 = \beta^{T-1}(1-\beta) \leq \frac{1}{2}$, we have

$$\begin{aligned}\beta^{T-1} &\leq \frac{1}{2(1-\beta)} \\ \Rightarrow T - 1 &\geq \frac{-\ln(2(1-\beta))}{\ln\beta} \\ \Rightarrow T &\geq \left(1 + \frac{-\ln(2(1-\beta))}{\ln\beta}\right)\end{aligned}$$

Hence, the smallest T is $\max\left(\lceil 1 + \frac{-\ln(2(1-\beta))}{\ln\beta} \rceil, 1\right)$.

8. By direct computation,

$$\alpha'_t = \frac{\alpha_t}{\sum_{t=1}^{T}\alpha_t} = \frac{\alpha_t}{(1-\beta)\sum_{t=0}^{T} -1\beta^t} = \frac{\alpha_t}{(1-\beta)\frac{1-\beta^T}{1-\beta}} = \frac{(1-\beta)\beta^{T-t}}{1-\beta^T} = \frac{\beta^{T-1}}{\sum_{t=0}^{T-1}\beta^t}$$

9. Solve $\frac{(1-\beta)\beta^{T-1}}{1-\beta^T} \leq \frac{1}{2}$, we obtain

$$\begin{aligned}&\beta^{T-1} - \beta^T \leq \frac{1}{2} - \frac{1}{2}\beta^T \\ \Rightarrow\ &\beta^{T-1}(1 - \frac{1}{2}\beta) \leq \frac{1}{2} \\ \Rightarrow\ &\beta^{T-1} \leq \frac{1}{2-\beta} \\ \Rightarrow\ &T \geq 1 - \frac{\ln(2-\beta)}{\ln\beta}\end{aligned}$$

Hence the smallest T is $\max\left(\lceil 1 - \frac{\ln(2-\beta)}{\ln\beta} \rceil, 1\right)$.

10. Expand the target function, we obtain

$$\begin{aligned}
&\min_{\mathbf{w}} E_{\mathbf{p}}(\mathbf{y}^T\mathbf{y} - 2(\mathbf{w}\odot\mathbf{p})^T X^T\mathbf{y} + (\mathbf{w}\odot\mathbf{p})^T X^T X(\mathbf{w}\odot\mathbf{p})) \\
\Rightarrow\quad &\min_{\mathbf{w}} E_{\mathbf{p}}(-2(\mathbf{w}\odot\mathbf{p})^T X^T\mathbf{y} + (\mathbf{w}\odot\mathbf{p})^T X^T X(\mathbf{w}\odot\mathbf{p})) \\
&= \min_{\mathbf{w}} -(\mathbf{w}^T X^T\mathbf{y}) + E_{\mathbf{p}}((\mathbf{w}\odot\mathbf{p})^T X^T X(\mathbf{w}\odot\mathbf{p})) && \text{(Since } E_{\mathbf{p}}(\mathbf{w}\odot\mathbf{p}) = \tfrac{1}{2}\mathbf{w}\text{)} \\
&= \min_{\mathbf{w}} -(\mathbf{w}^T X^T\mathbf{y}) + E_{\mathbf{p}}\left(\sum_{i=1}^{d}\sum_{i=1}^{d} w_i p_i w_j p_j (X^T X)_{ji}\right) \\
&= \min_{\mathbf{w}} -(\mathbf{w}^T X^T\mathbf{y}) + \frac{1}{4}\mathbf{w}^T X^T X\mathbf{w} + \frac{1}{4}\sum_{i=1}^{d}\sum_{k=1}^{N} w_i^2 (X_{ki})^2 \quad (*) \text{ (Since } E_{\mathbf{p}}(p_i p_j) = \tfrac{1}{4} \text{ for } i\neq j,\ \tfrac{1}{2} \text{ for } i = j\text{)}
\end{aligned}$$

We called $(*)$ the target function $T(\mathbf{w})$. Then we compute the partial derivatives of $T(\mathbf{w})$. Since it's optimal solution, the gradient of T should be $\mathbf{0}$.

$$\begin{aligned}
\frac{\partial T}{\partial w_i} &= -(X^T\mathbf{y})_i + \frac{1}{2}(X^T X\mathbf{w})_i + \frac{1}{2}\sum_{k=1}^{N} w_i (X_{ki})^2 = 0 \\
&\Rightarrow\ \frac{1}{2}(X^T X\mathbf{w})_i + \frac{1}{2}\sum_{k=1}^{N} w_i (X_{ki})^2 = (X^T\mathbf{y})_i
\end{aligned}$$

Write the system of equations to $Z\mathbf{w} = X^T\mathbf{y}$, where

$$Z = \frac{1}{2}X^T X + \frac{1}{2}\begin{pmatrix} \sum_{k=1}^{N}(X_{k1})^2 & 0 & \dots & 0 \\ 0 & \sum_{k=1}^{N}(X_{k2})^2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & \sum_{k=1}^{N}(X_{kd})^2 \end{pmatrix}$$

Hence the solution for optimal $\mathbf{w} = Z^\dagger X^T\mathbf{y}$.

11. Let A (respectively, B, C) be the set of data that g_1 (respectively, g_2, g_3) makes an mistakes on. From the problem description, we know that A (respectively, B, C) accounts for 0.08 (respectively, 0.16, 0.32) of the testing data.
If A, B, C are disjoint (which is possible, since $0.08 + 0.16 + 0.32 < 1$), then $E_{out}(G) = 0$ since no data is predicted wrong by at least two g's, which is a minimum.
If A, B are disjoint, $A \subset C$ and $B \subset C$, then $E_{out}(G) = 0.08 + 0.16 = 0.24$, which is the maximum of $E_{out}(G)$. Since if $E_{out}(G) > 0.24$, all the data should be predicted wrong on either g_1 or g_2 (or both), but A and B only takes 0.24 of the whole testing data, which is absurd.
By the above derivation, we have $0 \le E_{out}(G) \le 0.24$.

12. We prove the statement by contradiction.
Let $E = \frac{2}{K+1}\sum_{k=1}^{K} e_k$. Suppose E is not an upper bound of $E_{out}(G)$, i.e, $E_{out}(G)$ can be larger than E, say $E_{out}(G) = E'$ under the assumption of the problem with $E' > E$. In other words, G predicts wrong on E' of testing data.
Since G is a uniform blending classifier, each data that G predicts wrong should also be predicted wrong by at least $\frac{K+1}{2}$ g_k's. So, we have the summation of $E_{out}(g_k)$

$$\sum_{k=1}^{K} E_{out}(g_k) \ge E' \cdot \frac{K+1}{2} > E \cdot \frac{K+1}{2} = \sum_{k=1}^{K} e_k,$$

which is a contradiction.

13. In each time of pN sampling, the probability that a data isn't chosen is $\frac{N-1}{N}$. So the probability that a data isn't chosen through the whole pN sampling is $\left(\frac{N-1}{N}\right)^{pN}$. Thus, the expected number of data which are sampled at least once is $N\left(1-\left(\frac{N-1}{N}\right)^{pN}\right)$. Since N is large, we take limit and let $N \to \infty$ and obtain

$$\lim_{N\to\infty}\left(\frac{N-1}{N}\right)^{pN} = \lim_{N\to\infty}\left(1-\frac{1}{N}\right)^{pN} = e^{-p} \qquad \text{(By the definition of } e\text{)}$$

Hence, when N is large,

$$N\left(1-\left(1-\frac{1}{N}\right)^{pN}\right) \approx N(1-e^{-p}).$$

14. First, we focus on the extreme case: simply choose $\theta = -1$, then $g_{1,i,\theta}(\mathbf{x}) = 1$ and $g_{-1,i,\theta}(\mathbf{x}) = -1$ for all $\mathbf{x}$ and any dimension i. So there are two extreme case decision stumps: all $g(\mathbf{x})$ is 1 (or -1).
Then, for $i \in \{1,2,3,4\}$, since $x_i \in \{0,1,2,3,4,5\}$, let $\theta \in \{0.5, 1.5, 2.5, 3.5, 4.5\}$, then $g_{1,i,\theta}$ and $g_{-1,i,\theta}$ are all different decision stumps for the integer input vectors $\mathcal{X}$.
Thus, when discussing the number of decision stumps (cases lead to all positive of negative y's are excluded), there are total

$$(num\ of\ i) \times (num\ of\ s) \times (num\ of\ \theta) = 4 \times 2 \times 5 = 40$$

different decision stumps.
Hence for the case of the problem, there are total $2 + 40 = 42$ decision stumps.

15. Let $\mathcal{G}^+$ denote the set containing those g_t's such that $g_t(\mathbf{x})g_t(\mathbf{x}') = 1$ and $\mathcal{G}^-$ denote the set containing those g_t's such that $g_t(\mathbf{x})g_t(\mathbf{x}') = -1$.
Our goal is to find

$$\begin{aligned}(\phi_{ds}(\mathbf{x})^T)(\phi_{ds}(\mathbf{x}')) &= \sum_{t=1}^{|\mathcal{G}|} g_t(\mathbf{x})g_t(\mathbf{x}') \\ &= |\mathcal{G}^+| \times 1 + |\mathcal{G}^-| \times (-1) \\ &= |\mathcal{G}| - 2|\mathcal{G}^-|.\end{aligned}$$

By the analysis of the previous problem, we know that $|\mathcal{G}| = 2 + 2d(R-L)$.
Now, we'd like to find $|\mathcal{G}^-|$.
For $s \in \{+1, -1\}$ and $i \in \{1, \ldots, d\}$, WLOG, we assume $x_i \le x_i'$.
When $L \le \theta \le x_i$ or $x_i' < \theta \le R$,

$$g_{s,i,\theta}(\mathbf{x})g_{s,i,\theta}(\mathbf{x}') = \begin{cases}(+1)(+1) & , s = +1 \\ (-1)(-1) & , s = -1\end{cases} = 1$$

when $x_i < \theta \le x_i'$,

$$g_{s,i,\theta}(\mathbf{x})g_{s,i,\theta}(\mathbf{x}') = \begin{cases}(-1)(+1) & , s = +1 \\ (+1)(-1) & , s = -1\end{cases} = -1.$$

Hence, for any i, there are

$$2 \times |x_i - x_i'|$$

decision stumps in $|\mathcal{G}^-|$. So we know that

$$|\mathcal{G}^-| = \sum_{i=1}^{d} 2|x_i - x_i'| = 2\|\mathbf{x} - \mathbf{x}'\|.$$

Hence we obtain

$$(\phi_{ds}(\mathbf{x})^T)(\phi_{ds}(\mathbf{x}')) = |\mathcal{G}| - 2|\mathcal{G}^-| = 2 + 2d(R-L) - 4\|\mathbf{x} - \mathbf{x}'\|$$

16. Let $\mathcal{G}$ denote all different decision stumps for $\mathcal{X}$. Each decision stump in $\mathcal{G}$ can be indexed by a 3-tuple (s, i, θ), where $s \in \{-1, +1\}, i \in \{1, \ldots, d\}$ and $\theta \in [L, R]$. We collect these indexes in an index set $\mathcal{C}$. Since $\mathcal{C}$ is $2 \times d$ line segments in $\mathbb{R}^3$, $\mathcal{C}$ is clearly measurable.
Now, let $\phi_{\mathbf{x}}(s, i, \theta) = g_{(s,i,\theta)}(\mathbf{x})$, we know that $(\phi_{\mathbf{x}})^2 = 1$ for any $\mathbf{x} \in \mathcal{X}$, so is $\phi_{\mathbf{x}}$ is $\mathcal{L}^2$ integrable. Thus, the infinite dimensional inner product can be computed by the inner product function defined on $\mathcal{L}^2$ space, i.e,

$$(\phi_{ds}(\mathbf{x}))^T(\phi_{ds}(\mathbf{x}')) = \int_{\mathcal{C}} \phi_{\mathbf{x}}(s, i, \theta)\phi_{\mathbf{x}'}(s, i, \theta)d(s, i, \theta)$$

Since i and s are only discrete values, the above integral can be written as

$$\begin{aligned}
(\phi_{ds}(\mathbf{x}))^T(\phi_{ds}(\mathbf{x}')) &= \sum_{s\in\{1,-1\}} \sum_{i=1}^{d} \int_{[L,R]} g_{(s,i,\theta)}(\mathbf{x})g_{(s,i,\theta)}(\mathbf{x}')d\theta \\
&= 2\sum_{i=1}^{d} \int_{L}^{R} g_{(s,i,\theta)}(\mathbf{x})g_{(s,i,\theta)}(\mathbf{x}')d\theta \\
&= 2\sum_{i=1}^{d} \int_{L}^{R} \text{sign}(x_i - \theta)\text{sign}(x_i' - \theta)d\theta \\
&= 2\sum_{i=1}^{d} ((R - L) - |x_i - x_i'|) \times 1 + |x_i - x_i'| \times (-1)) \qquad (*) \\
&= 2d(R - L) - 4\|\mathbf{x} - \mathbf{x}'\|,
\end{aligned}$$

where $(*)$ is by the length of θ in $[L, R]$ that leads to $\text{sign}(x_i - \theta)\text{sign}(x_i' - \theta) = -1$ is $|x_i - x_i'|$, and the length of θ in $[L, R]$ that lead to $\text{sign}(x_i - \theta)\text{sign}(x_i' - \theta) = 1$ is $L - R - |x_i - x_i'|$.

17. The lecture I like the most is the kernel SVM (MLT lecture 3). In my path of learning, I always find it fascinating to transform an infinite-dimensional problem to a finite-dimensional one. Kernel method "bags" feature transform and inner product together, which is somehow implicitly but powerful to me. By coincidence, my contribution in final project is also SVM :-) Although it seems that SVM doesn't perform well in the final project, I still like the lecture that introduces the kernel.

18. The lecture I like least is intro to CNN, lectured on 5/1 (maybe the same with some students). That's our very first time to receive most knowledge not by video. As the teacher said in the following class, somehow both the teacher and me can't get the right tempo through the skeleton slides, or it might because it's online teaching. However, since it's a new and important topic to machine learner, given deep learning everywhere, it's a pity that that lecture didn't give me really impressive construction of concepts. But the good thing is that the following lectures are great. :-)