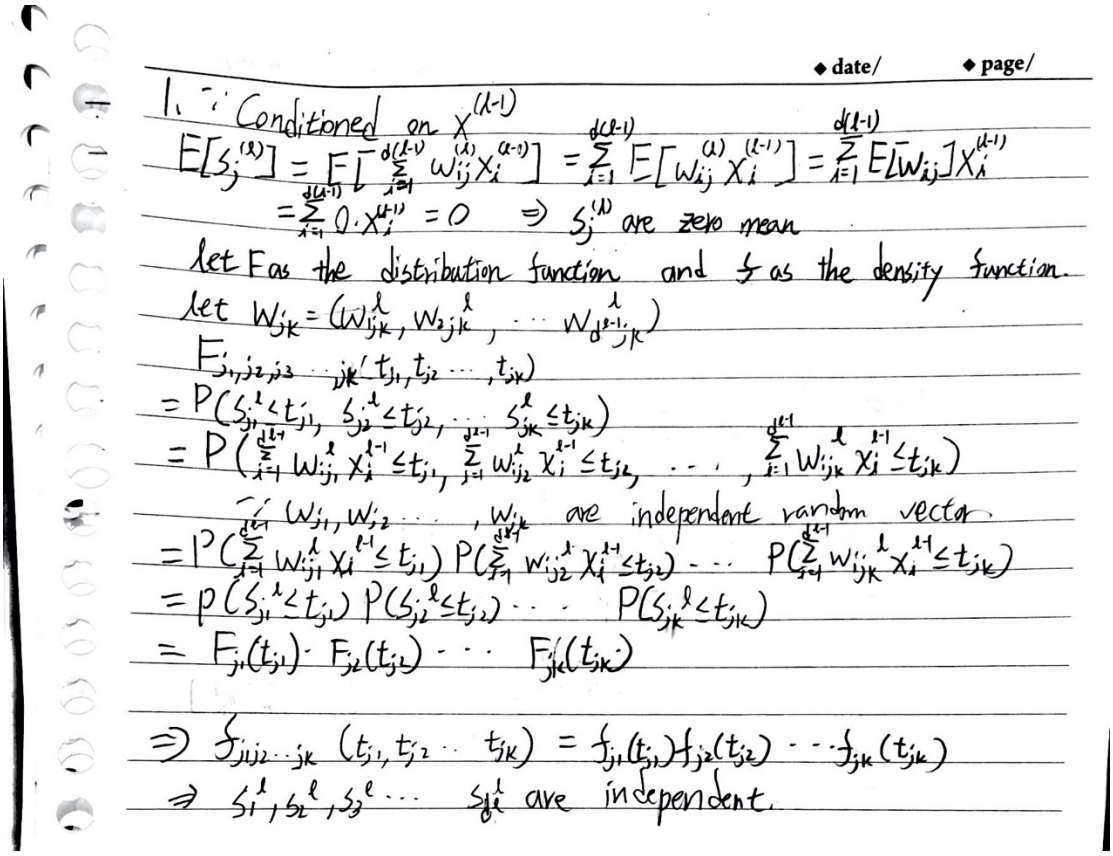
http://www.work.caltech.edu/~htlin/publication/doc/infkernel.pdf?fbclid=IwAR0y3-BpeP3_nwaz9tLFtJZh3-XiGTFv7M9CvkurYlgkj_MQbJx9tWfav54
https://pouannes.github.io/blog/initialization/?fbclid=IwAR1JtHrBBNL68MnBHFQ4its6vNzb1kKoNsCVG8pu9IVdqq-rFHZY7tG8WPU
http://d2l.ai/chapter_multilayer-perceptrons/numerical-stability-and-init.html?fbclid=IwAR11jhydSiNDLjKGn3euWHVlRtku39han8rJd9iimJCUbmVhlBoxZChdLn8



◆ date/ ◆ page/

1. Conditioned on $x^{(l-1)}$

$$E[s_j^{(l)}] = E\left[\sum_{i=1}^{d^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)}\right] = \sum_{i=1}^{d^{(l-1)}} E[w_{ij}^{(l)} x_i^{(l-1)}] = \sum_{i=1}^{d^{(l-1)}} E[w_{ij}] x_i^{(l-1)}$$

$$= \sum_{i=1}^{d^{(l-1)}} 0 \cdot x_i^{(l-1)} = 0 \Rightarrow s_j^{(l)} \text{ are zero mean}$$

let F as the distribution function and f as the density function.

let $W_{jk} = (w_{jk}^{l}, w_{2jk}^{l}, \cdots w_{d^{l-1}jk}^{l})$

$$F_{j_1,j_2,j_3 \cdots j_K}(t_{j_1}, t_{j_2} \cdots, t_{j_K})$$

$$= P(s_{j_1}^{l} < t_{j_1}, s_{j_2}^{l} \le t_{j_2}, \cdots s_{j_K}^{l} \le t_{j_K})$$

$$= P\left(\sum_{i=1}^{d^{l-1}} w_{j_1}^{l} x_i^{l-1} \le t_{j_1}, \sum_{i=1}^{d^{l-1}} w_{j_2}^{l} x_i^{l-1} \le t_{j_2}, \cdots, \sum_{i=1}^{d^{l-1}} w_{j_K}^{l} x_i^{l-1} \le t_{j_K}\right)$$

$\because w_{j_1}, w_{j_2} \cdots, w_{j_K}$ are independent random vector.

$$= P\left(\sum_{i=1}^{d^{l-1}} w_{j_1}^{l} x_i^{l-1} \le t_{j_1}\right) P\left(\sum_{i=1}^{d^{l-1}} w_{j_2}^{l} x_i^{l-1} \le t_{j_2}\right) \cdots P\left(\sum_{i=1}^{d^{l-1}} w_{j_K}^{l} x_i^{l-1} \le t_{j_K}\right)$$

$$= P(s_{j_1}^{l} \le t_{j_1}) P(s_{j_2}^{l} \le t_{j_2}) \cdots P(s_{j_K}^{l} \le t_{j_K})$$

$$= F_{j_1}(t_{j_1}) \cdot F_{j_2}(t_{j_2}) \cdots F_{j_K}(t_{j_K})$$

$$\Rightarrow f_{j_1 j_2 \cdots j_K}(t_{j_1}, t_{j_2} \cdots t_{j_K}) = f_{j_1}(t_{j_1}) f_{j_2}(t_{j_2}) \cdots f_{j_K}(t_{j_K})$$

$\Rightarrow s_1^{l}, s_2^{l}, s_3^{l} \cdots s_{d^l}^{l}$ are independent.

◆ date/ ◆ page/

2. $\mathrm{Var}[x_i^{(l-1)}] = E[(x_i^{l-1})^2] - E[x_i^{l-1}]^2$

$\Rightarrow E[(x_i^{l-1})^2] = \sigma_x^2 + \bar{x}^2$

$\mathrm{Var}[w_{ij}^l] = E[(w_{ij}^l)^2] - E[w_{ij}^l]^2$

$\Rightarrow E[(w_{ij}^l)^2] = \sigma_w^2$

$$\begin{aligned}
\mathrm{Var}[s_j^l] &= E[(s_j^l)^2] - E[s_j^l]^2 \\
&= E\left[\left(\sum_{i=1}^{d^{l-1}} w_{ij}^l x_i^{l-1}\right)^2\right] \\
&= E\left[\sum_{i=1}^{d^{l-1}}\sum_{k=1}^{d^{l-1}} w_{ij}^l w_{kj}^l x_i^{l-1} x_k^{l-1}\right] \\
&= E\left[\sum_{i=1}^{d^{l-1}} (w_{ij}^l)^2 (x_i^{l-1})^2 + \sum_{i\neq k} w_{ij}^l w_{kj}^l x_i^{l-1} x_k^{l-1}\right] \\
&= \sum_{i=1}^{d^{l-1}} E[(w_{ij}^l)^2 (x_i^{l-1})^2] + \sum_{i\neq k} E[w_{ij}^l w_{kj}^l x_i^{l-1} x_k^{l-1}] \\
&\because s_j^l,\ w_{ij}^l,\ x_i^{l-1}\text{'s independence} \\
&= \sum_{i=1}^{d^{l-1}} E[(w_{ij}^l)^2] E[(x_i^{l-1})^2] + \sum_{i\neq k} E[w_{ij}^l] E[w_{kj}^l] E[x_i^{l-1}] E[x_k^{l-1}] \\
&= d^{l-1}\sigma_w^2 (\sigma_x^2 + \bar{x}^2) + \sum_{i\neq k} 0 \cdot 0 \cdot \bar{x} \cdot \bar{x} \\
&= d^{l-1}\sigma_w^2 (\sigma_x^2 + \bar{x}^2)
\end{aligned}$$

3. $\mathrm{ReLU}(x) \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} = \max(x, 0)$

$E[(x_i^{l-1})^2] = E[(\max(s_i^{l-1}, 0)^2]$

let F as the distribution function and f as the density function

$$F_{x_i^{l-1}}(t) = P(\max(s_i^{l-1}, 0) \leq t) = \begin{cases} P(s_i^{l-1} \leq t) = F_{s_i^{l-1}}(t) & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

$$\Rightarrow f_{x_i^{l-1}}(t) = F'_{x_i^{l-1}}(t)\,dt = \begin{cases} F'_{s_i^{l-1}}(t) = f_{s_i^{l-1}}(t) & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

$$\begin{aligned}
\Rightarrow E[(x_i^{l-1})^2] &= \int_{-\infty}^{\infty} t^2 f_{x_i^{l-1}}(t)\,dt = \int_{-\infty}^{0} t^2 f_{x_i^{l-1}}(t)\,dt + \int_{0}^{\infty} t^2 f_{x_i^{l-1}}(t)\,dt \\
&= \int_{0}^{\infty} t^2 f_{x_i^{l-1}}(t)\,dt \quad \because s_i^{l-1} \text{ is symmetric.} \\
&= \frac{1}{2}\int_{-\infty}^{\infty} t^2 f_{s_i^{l-1}}(t)\,dt \\
&= \frac{1}{2} E[(s_i^{l-1})^2] \quad \#
\end{aligned}$$

◆ date/ ◆ page/

4: By problem 2,3, we have:

$$E[(S_j^{l-1})^2] = 2E[(x_i^{l-1})^2] = 2(\sigma_x^2 + \bar{x}^2)$$

$$\Rightarrow Var[S_j^{l-1}] = E[(S_j^{l-1})^2] - E[S_j^{l-1}]^2$$
$$= 2(\sigma_x^2 + \bar{x}^2) - 0^2$$

$$\therefore Var[S_j^l] = d_{l-1}\sigma_w^2(\sigma_x^2 + \bar{x}^2)$$
$$= \frac{d_{l-1}}{2}\sigma_w^2 \cdot 2(\sigma_x^2 + \bar{x}^2)$$
$$= \frac{d_{l-1}}{2}\sigma_w^2 \cdot Var[S_j^{l-1}]$$

5. leaky ReLU(x) $\begin{cases} x & \text{if } x \geq 0 \\ a \cdot x & \text{if } x < 0 \end{cases}$ $\max(x, a \cdot x)$

Let F as the distribution function and f as the density function.

$$F_{x_i^{l-1}}(t) = P(\max(S_i^{l-1}, a \cdot S_i^{l-1}) \leq t) = \begin{cases} P(S_i^{l-1} \leq t) = F_{S_i^{l-1}}(t) & \text{if } t \geq 0 \\ P(S_i^{l-1} \leq \frac{t}{a}) = F_{S_i^{l-1}}(\frac{t}{a}) & \text{if } t < 0 \end{cases}$$

$$\Rightarrow f_{x_i^{l-1}}(t) = F'_{x^{l-1}}(t)dt = \begin{cases} F'_{S_i^{l-1}}(t) = f_{S_i^{l-1}}(t) & \text{if } t \geq 0 \\ F'_{S_i^{l-1}}(\frac{t}{a}) = \frac{1}{a} f_{S_i^{l-1}}(\frac{t}{a}) & \text{if } t < 0 \end{cases}$$

$$\Rightarrow E[(X_i^{l-1})^2] = \int_{-\infty}^{0} t^2 f_{x_i^{l-1}}(t)dt + \int_{0}^{\infty} t^2 f_{x_i^{l-1}}(t)dt$$
$$= \int_{-\infty}^{0} \frac{t^2}{a} f_{S_i^{l-1}}(\frac{t}{a})dt + \int_{0}^{\infty} t^2 f_{S_i^{l-1}}(t)dt$$
$$= a^2 \int_{-\infty}^{0} t^2 f_{S_i^{l-1}}(t)dt + \int_{0}^{\infty} t^2 f_{S_i^{l-1}}(t)dt$$

$\because S_i^{l-1}$ is symmetric

$$= \frac{a^2+1}{2} \int_{-\infty}^{\infty} t^2 f_{S_i^{l-1}}(t)dt$$
$$= \frac{a^2+1}{2} E[(S_i^{l-1})^2]$$

Also by problem 2. $E[(S_i^{l-1})^2] = \frac{2}{a^2+1} E[(X_i^{l-1})^2] = \frac{2}{a^2+1}(\sigma_x^2 + \bar{x}^2)$

$$\Rightarrow Var[S_i^{l-1}] = E[(S_i^{l-1})^2] - E[S_i^{l-1}]^2 = \frac{2}{a^2+1}(\sigma_x^2 + \bar{x}^2)$$

By problem $\Rightarrow (\sigma_x^2 + \bar{x}^2) = \frac{a^2+1}{2} Var[S_i^{l-1}]$

By problem 2 we have. $Var[S_j^l] = d^{l-1}\sigma_w^2(\sigma_x^2 + \bar{x}^2) = \frac{d^{l-1}\sigma_w^2(a^2+1)}{2} Var[S_i^{l-1}]$

$\Rightarrow$ 取 $\sigma_w^2 = \frac{2}{d^{l-1}(a^2+1)}$, and all the reserve show pp~pq,

we have $Var[S_j^l] = Var[S_j^{l-1}]$ #1

6

$V_1 = (1-\beta)\Delta t$

$V_2 = \beta(1-\beta)\Delta_1 + (1-\beta)\Delta_2$

$V_3 = \beta^2(1-\beta)\Delta_1 + \beta(1-\beta)\Delta_2 + (1-\beta)\Delta_3$

guess $V_k = \sum_{i=1}^{k} \beta^{k-i}(1-\beta)\Delta_i$

when $k=1,2,3$, assumption holds

Assume that when $k=n$, assumption also hold.

$V_n = \sum_{i=1}^{n} \beta^{n-i}(1-\beta)\Delta_i$

when $k=n+1$

$V_{n+1} = \beta V_n + (1-\beta)\Delta_{n+1}$

$= \sum_{i=1}^{n} \beta^{n-i+1}(1-\beta)\Delta_i + (1-\beta)\Delta_{n+1}$

$= \sum_{i=1}^{n+1} \beta^{n+1-i}(1-\beta)\Delta_i$, assumption holds!

By mathematical induction,

$\forall T \in \mathbb{N}$, $V_T = \sum_{i=1}^{T} \beta^{T-i}(1-\beta)\Delta_i$

$\Rightarrow \alpha_t = \beta^{T-1}(1-\beta)$ #

7. $\because 0 < \beta < 1$

$\therefore \log_2 \beta < 0$

$\alpha_1 \leq \frac{1}{2}$

$\Rightarrow \beta^{T-1}(1-\beta) \leq \frac{1}{2}$

$\Rightarrow \log_2(\beta^{T-1}(1-\beta)) \leq \log_2 \frac{1}{2}$

$\Rightarrow (T-1)\log_2\beta + \log_2(1-\beta) \leq -1$

$\Rightarrow (T-1) \geq \frac{-1-\log_2(1-\beta)}{\log_2\beta}$

$\Rightarrow T \geq \frac{-1-\log_2(1-\beta)}{\log_2\beta} - 1$

Ans : $T = \left\lceil \frac{-1-\log_2(1-\beta)}{\log_2\beta} - 1 \right\rceil$

◆ date/ ◆ page/

8. $\alpha'_t = \frac{\alpha_t}{\sum_{t=1}^{T} \alpha_t}$

$= \frac{\beta^{T-t}(1-\beta)}{\sum_{t=1}^{T} \beta^{T-t}(1-\beta)}$

$= \frac{\beta^{T-t}(1-\beta)}{(1-\beta)\sum_{t=1}^{T} \beta^{T-t}}$

$= \beta^{T-t}\frac{1-\beta}{1-\beta^T}$ #

9: $\because 0<\beta<1$

$\therefore \log_2 \beta < 0$

$\alpha'_1 \le \frac{1}{2}$

$\Rightarrow \beta^{T-1}\frac{1-\beta}{1-\beta^T} \le \frac{1}{2}$

$\Rightarrow \frac{1-\beta^{-1}}{1-\beta^{-T}} \le \frac{1}{2}$

$\Rightarrow 2(1-\beta^{-1}) \le 1-\beta^{-T}$

$\Rightarrow \beta^{-T} \le 2\beta^{-1}-1$

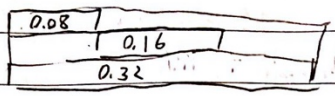
$\Rightarrow T\ln\beta^{-1} \le \ln(2\beta^{-1}-1)$

$\Rightarrow T \ge \frac{\ln(2\beta^{-1}-1)}{\ln\beta^{-1}}$

Ans: $T = \left\lceil \frac{\ln(2\beta^{-1}-1)}{\ln\beta^{-1}} \right\rceil$

10.

◆ date/ ◆ page/

11 We need at least 2 classifier to wrongly classify then G will classify the data wrong.
Therefore, the lower bound would be 0, as the three classifier didn't over lap.
The upper bound would be.


with the most overlap;
As a result, $0 \le E_{out}(G) \le 0.24$

12. Because K is odd, we need at least $\frac{K+1}{2}$ classifier to wrongly classify, the G will wrongly classify.
Let the data size as N;
then we will have $e_n N$ data wrongly classify by g_k.
Thus, we will have $\sum_{n=1}^{K} e_n N$ data wrongly classify by $g_1 g_2 \cdots g_k$
When the datas are correctly classify by G are also correctly classify by $g_1, g_2 \cdots g_k$ and the datas are wrongly classify by G are only wrongly classify by $\frac{K+1}{2}$ of the classifier, $g_1, g_2 \cdots g_k$, at this situation the upper bound of the data wrongly classify would be $\frac{\sum_{n=1}^{k} e_n N}{\frac{K+1}{2}} = \frac{2}{K+1} \sum_{n=1}^{k} e_n N$.

$\Rightarrow$ one of the upper bound of $E_{out}(G)$ is $\frac{\frac{2}{K+1}\sum_{n=1}^{k} e_n N}{N} = \frac{2}{K+1}\sum_{n=1}^{k} e_n$ //

◆ date/ ◆ page/

13. If the results of every sample are independent;
the possibility that at sample $N=pN$ the data hadn't been sample once is $\left(\frac{N-1}{N}\right)^{pN}=\left(1-\frac{1}{N}\right)^{pN}\approx e^{-p}$ is N is large enough
Then, the possibility for the data to get sample at least once is $(1-e^{-p})$ (e.g. This is the approx results by wolfram: $\left(1-\frac{a}{N}\right)^{pN}\approx e^{-ap}$)
Let X_i represents that whether the ith data got sample, 1 means it got sample at least once, 0 means it didn't got sample.
We now consider k random variable X_i, where there are s times successed and t times failed, $s+t=k$. Also let $x_{i1}, x_{i2}\cdots x_{ik}=\{0,1\}$

$\Rightarrow P(X_{i1}=x_{i1})P(X_{i2}=x_{i2})\cdots P(X_{ik}=x_{ik})\approx (1-e^{-p})^s(e^{-p})^t$

$\therefore P(X_{i1}=x_{i1}, X_{i2}=x_{i2}, \cdots X_{ik}=x_{ik})$

$=\left(\frac{N-t}{N}\right)^{pN}-C_1^s\left(\frac{N-(t+1)}{N}\right)^{pN}+C_2^s\left(\frac{N-(t+2)}{N}\right)^{pN}-\cdots$

$+(-1)^s C_s^s\left(\frac{N-(t+s)}{N}\right)^{pN}$

$\approx e^{-tp}-C_1^s e^{-(t+1)p}+C_2^s e^{-(t+2)p}-\cdots+(-1)^s C_s^s e^{-(t+s)p}$

$=e^{-tp}\left(1-C_1^s e^{-p}+C_2^s e^{-2p}-\cdots+(-1)^s C_s^s e^{-sp}\right)$

$=e^{-tp}(1-e^{-p})^s$

$=P(X_{i1}=x_{i1})P(X_{i2}=x_{i2})\cdots P(X_{ik}=x_{ik})$

Which means that they are independent random variable.

$\Rightarrow$ The expected value that datas got sample at least once is $(1-e^{-p})N=N-(N\cdot e^{-p})$ #

◆ date/ ◆ page/

14 Fist consider two special case: all classified as 1 or -1. i.e.
$\theta \le L$, $\theta > R$.
Then lee us consider cases where $L < \theta \le R$.
Notice that these decision stumps are same in a specific range: $k < \theta \le k+1$, $k = L+1, L+2 \cdots, R$
this means on a special dimension, we have
$2 \cdot (R-L)$ different decision stumps (2 is because $\{-1, +1\}$)
For d dimension. we have $2 \cdot (R-L) \cdot d + 2$ different ds.
this is the 2 special case mentioned at the beginning
$\Rightarrow 2 \cdot (5-0) \cdot 4 + 2 = 42$ //

15. First, notice that $K_{ds}(X, X') = |G| - 2 \cdot diff(X, X')$, where $diff(X, X')$ means the number of different decision stumps that $g(X) \neq g(X')$, $g \in G$. Consider a dimension i, then

$$diff(x_i, x_i') \begin{cases} x_i = x_i', 0 \\ x_i \neq x_i', 2 \cdot |x_i - x_i'| \end{cases} \Rightarrow diff(X, X') = \sum_{i=1}^{d} 2 \cdot |x_i - x_i'|$$

$$\Rightarrow K_{ds}(X, X') = 2 \cdot (R-L) \cdot d + 2 - 2 \cdot \sum_{i=1}^{d} 2 \cdot |x_i - x_i'|$$
$$= 2(R-L)d + 2 - 4 \sum_{i=1}^{d} |x_i - x_i'|$$

16. $\int_L^R g_{b,\beta}(x)\, g_{s,\theta}(x')\, d\theta$

$= \int_L^R \text{sign}(x_i-\theta)\,\text{sign}(x_i'-\theta)\, d\theta$

let $\text{sign}(x_i-\theta) = s_i(\theta)$, $\text{sign}(x_i'-\theta) = s_i'(\theta)$

$= \int_L^{\min(x_i,x_i')} s_i(\theta)s_i'(\theta)d\theta + \int_{\min(x_i,x_i')}^{\max(x_i,x_i')} s_i(\theta)s_i'(\theta)d\theta + \int_{\max(x_i,x_i')}^{R} s_i(\theta)s_i'(\theta)d\theta$

when $\theta \in [\min, \max)$ $s_i(\theta)s_i'(\theta) < 0$

else $s_i(\theta)s_i'(\theta) > 0$

$= \int_L^{\min} (+1)d\theta + \int_{\min}^{\max} (-1)d\theta + \int_{\max}^{R} (+1)\, d\theta.$

$= (\min - L) + (\min - \max) + (R - \max)$

$= (R-L) - 2(\max - \min)$

$= (R-L) - 2|x_i - x_i'|$

now, we summation over i

$\sum_{i=1}^{d} ((R-L) - 2|x_i - x_i'|)$

$= d(R-L) - 2\sum_{i=1}^{d} |x_i - x_i'|$

because $s \in \{+1, -1\}$

$\Rightarrow$ we need to 2^d times ?

$2^d d(R-L) - 4\sum_{i=1}^{d} |x_i - x_i'|$ #

◆ date/ ◆ page/

17. Actually, I like all of them. Prof. Lin, uses the simplest way to let us understand the hardest concepts. If I have to choose one, I would have to say the newly added course "Activation in Deep Learning" It gave me a better understand on the other ML courses I taken. Prof Lee's ML course is more on the application aspect, where as Prof Lin is more theoretical. With the combination of the two courses, I think I reached a higher level on ML.

18. It would also be the newly added course of Deep Learning. This is because the videos often have background noises, which made learning even harder. Also, personally, I like to review/preview with courses' powerpoint., I'm not used to hand written notes.