1 Universal Approximation theorem

$$f_m(x;\theta)=\sum_{k=1}^{m} a_k\,\sigma(b_k x + c_k), \quad \sigma(x)=\max(x,0)$$

$$P_u f=\sum_{k=0}^{u} f(x_k)\,t\left(\frac{x-x_k}{h}\right), \quad t(x)=\max(1-|x|,0)$$

(a) $P_u f$ can be represented as two-layer ReLU network

Proof.

$$P_u f(x)=\sum_{k=0}^{u} f(x_k)\max\left(1-\frac{|x-x_k|}{h},0\right)$$

$$=\sum_{k=0}^{u} f(x_k)\left(\max\left(\frac{x-(x_k-h)}{h},0\right)+\max\left(\frac{x-(x_k+h)}{h},0\right)-2\max\left(\frac{x-x_k}{h},0\right)\right)$$

$$=\sum_{k=0}^{u} f(x_k)\,\sigma\left(\frac{x-(x_k-h)}{h}\right)+\sum_{k=1}^{u} f(x_k)\,\sigma\left(\frac{x-(x_k+h)}{h}\right)+\sum_{k=1}^{u}(-2f(x_k))\,\sigma\left(\frac{x-x_k}{h}\right)$$

#

(b) $f\in C[0,1]$. $\forall\varepsilon>0$, $\exists f_m$, $\sup_{[0,1]}|f(x)-f_m(x;\theta)|\le\varepsilon$

Proof. Using (a), it sufficies to show $\forall\varepsilon>0$, $\exists u$, $\sup_{[0,1]}|f(x)-P_u f(x)|<\varepsilon$

f is uniform continuous on $[0,1]$. Let $\delta>0$ be $|x_1-x_2|\le\delta \Rightarrow |f(x_1)-f(x_2)|\le\varepsilon$

Take $u>\frac{1}{\delta}$. $|f(1)-P_u f(1)|=0$.

Now suppose $x\in[x_k,x_{k+1})$, $k=0,\cdots u-1$

$$|f(x)-P_u f(x)|=\left|f(x)-f(x_k)\,t\left(\frac{x-x_k}{h}\right)-f(x_{k+1})\,t\left(\frac{x-x_{k+1}}{h}\right)\right|$$

$$=\left|f(x)-f(x_k)\left(1-\frac{x-x_k}{h}\right)-f(x_{k+1})\left(1-\frac{x_{k+1}-x}{h}\right)\right|$$

$$=\left|\left(1-\frac{x-x_k}{h}\right)(f(x)-f(x_k))+\left(1-\frac{x_{k+1}-x}{h}\right)(f(x)-f(x_{k+1}))\right|$$

$$\le\left(1-\frac{x-x_k}{h}\right)|f(x)-f(x_k)|+\left(1-\frac{x_{k+1}-x}{h}\right)|f(x)-f(x_{k+1})|$$

$$\le\varepsilon\left(2-\frac{x-x_k+x_{k+1}-x}{h}\right)=\varepsilon \quad \text{We are done. \#}$$

(c) $f(x)=x^2$ $\varepsilon\in(0,1)$ $\exists\ m\le C\varepsilon^{-\frac{1}{2}}$ s.t. $\sup_{[0,1]}|f(x)-f_m(x;\theta)|\le\varepsilon$

Proof. Let $M=\lceil\frac{1}{2\sqrt{\varepsilon}}\rceil$ Consider $f_m(x;\theta)=P_M f$, here $m=3(M+1)\le 6+\frac{3}{2\sqrt{\varepsilon}}$

$\le\frac{10}{\sqrt{\varepsilon}}$

Suppose $x\in[x_k,x_{k+1})$ Let $x=x_k+\Delta$. $\Delta<h$

$|f(x)-f_m(x;\theta)|=|f(x)-P_M f(x)|$

$=\left|(1-\frac{x-x_k}{h})(x^2-x_k^2)-(1-\frac{x_{k+1}-x}{h})(x_{k+1}^2-x^2)\right|$

$=\left|(1-\frac{\Delta}{h})(\Delta^2+2\Delta x_k)-\frac{\Delta}{h}(h-\Delta)(2x_k+h+\Delta)\right|$

$=\left|(1-\frac{\Delta}{h})(\Delta^2+2\Delta x_k-2\Delta x_k-\Delta h-\Delta^2)\right|$

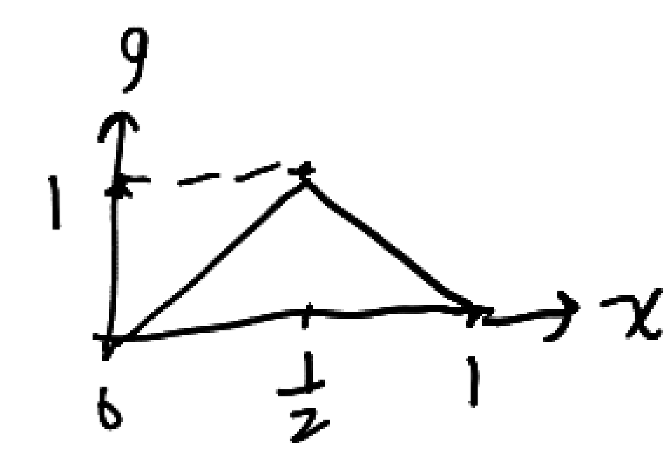
$=(h-\Delta)\Delta\le\frac{h^2}{4}\le\varepsilon$

$\sup_{[0,1]}|f(x)-f_m(x;\theta)|\le\varepsilon$. We are done. #

2 Approximate x^2 with DNN

$f^*(x)=x^2$ in $[0,1]$ $g(x)=t(2x-1)$ which maps $[0,1]$ to $[0,1]$

$g_s(x)=\underbrace{g\circ\cdots\circ g}_{s}(x)$



(a) Show that $\sup_{[0,1]}|P_{2^l}f^*(x)-f^*(x)|\le\frac{C}{2^{2l}}$

Proof. Similar to (c) of last homework,

$|f(x)-P_M f(x)|=\left|(1-\frac{\Delta}{h})(\Delta^2+2\Delta x_k)-\frac{\Delta}{h}(h-\Delta)(2x_k+h+\Delta)\right|$

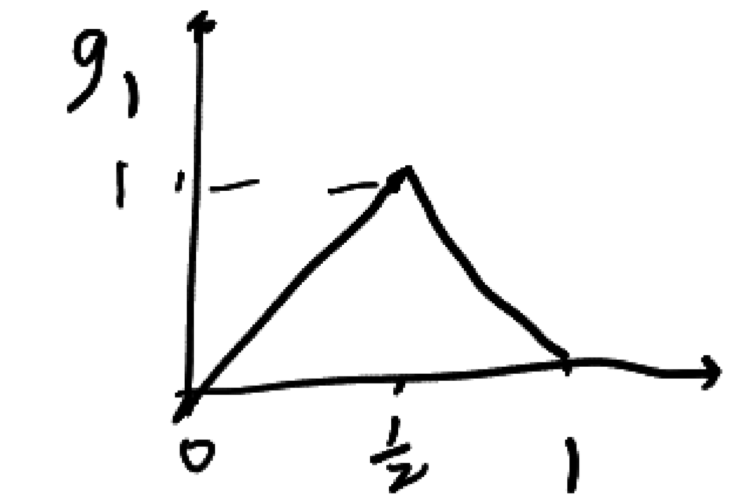
(where $x=x_k+\Delta$, $0\le\Delta<h$, $k=0,1,\cdots M-1$)

$=\left|(1-\frac{\Delta}{h})(\Delta^2+2\Delta x_k-2\Delta x_k-\Delta h-\Delta^2)\right|$

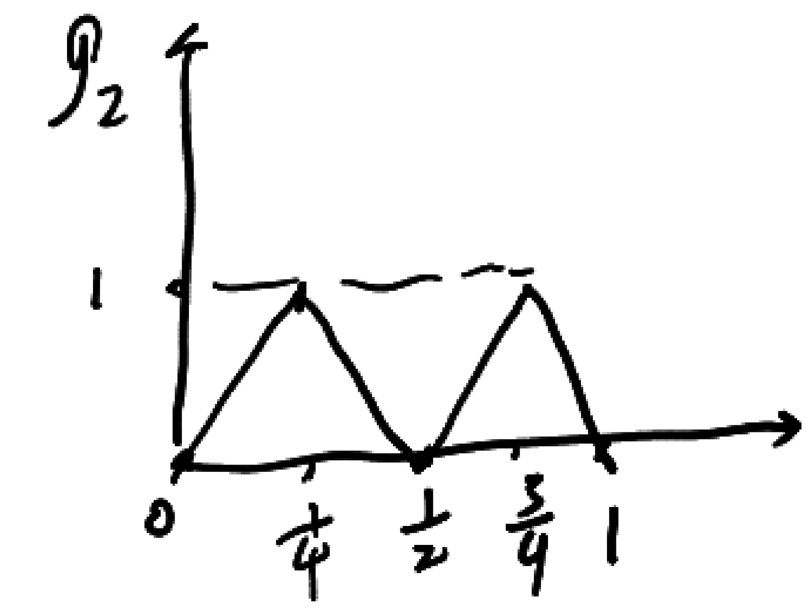
$=(h-\Delta)\Delta\le\frac{h^2}{4}=\frac{1}{4M^2}=\frac{1}{4}\cdot\frac{1}{2^{2l}}$ #

(b) $\ell = 2, 3, \dots$

$$P_{2^{\ell-1}} f^*(x) - P_{2^\ell} f^*(x) = \frac{g_\ell(x)}{2^{2\ell}}, \quad \forall x \in [0,1].$$

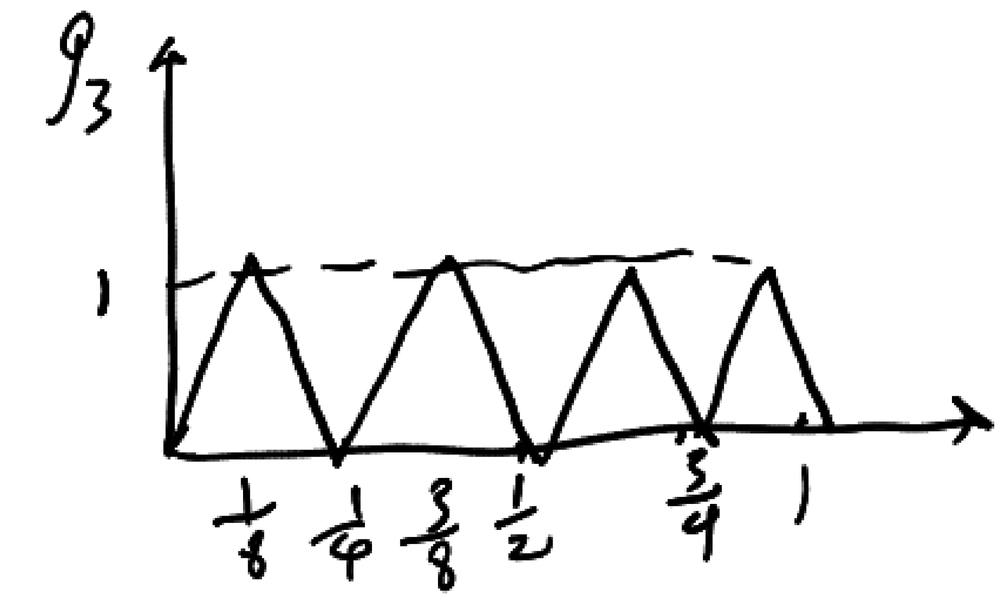


<u>Proof</u>. By induction, it is easy to verify $g_\ell(x)$ is a piecewise-linear function of step $\frac{1}{2^\ell}$ with $g_\ell\left(\frac{2k+1}{2^\ell}\right) = 1$, $k = 0, \dots 2^{\ell-1}-1$ and $g_\ell\left(\frac{k}{2^{\ell-1}}\right) = 0$, $k = 0, \dots 2^{\ell-1}$.



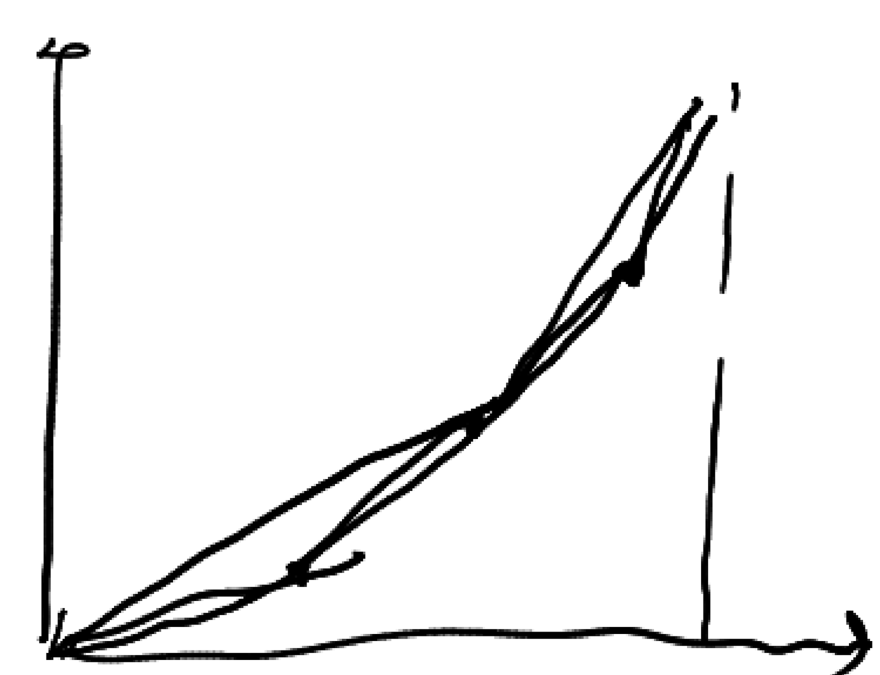
From definition, we also have

$P_{2^{\ell-1}} f^*(x) - P_{2^\ell} f^*(x)$ is piecewise-linear function of step $\frac{1}{2^\ell}$



with $P_{2^{\ell-1}} f^*\left(\frac{k}{2^{\ell-1}}\right) - P_{2^\ell} f^*\left(\frac{k}{2^{\ell-1}}\right) = \left(\frac{k}{2^{\ell-1}}\right)^2 - \left(\frac{k}{2^{\ell-1}}\right)^2 = 0$

$k = 0, \dots 2^{\ell-1}$



$$P_{2^{\ell-1}} f^*\left(\frac{2k+1}{2^\ell}\right) - P_{2^\ell} f^*\left(\frac{2k+1}{2^\ell}\right)$$

$$= \frac{1}{2}\left(\frac{k}{2^{\ell-1}}\right)^2 + \frac{1}{2}\left(\frac{k+1}{2^{\ell-1}}\right)^2 - \left(\frac{2k+1}{2^\ell}\right)^2$$

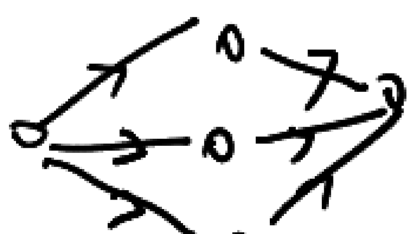
$$= \frac{k^2}{2^{2\ell-1}} + \frac{k^2+2k+1}{2^{2\ell-1}} - \frac{2k^2+2k+\frac{1}{2}}{2^{2\ell-1}} = \frac{1}{2^{2\ell}}, \quad k = 0, \dots 2^{\ell-1}-1$$

We have demonstrated $P_{2^{\ell-1}} f^*(x) - P_{2^\ell} f^*(x) = \frac{g_\ell(x)}{2^{2\ell}}, \ \forall x \in [0,1]$. #

(c) $P_{2^l} f^*$ can be represented as $O(l)$-layer NN with $O(1)$ width

Proof. $t(2x-1) = \max(2x-2, 0) + \max(2x, 0) - 2\max(2x-1, 0)$

$= \sigma(2x-2) + \sigma(2x) - 2\sigma(2x-1)$

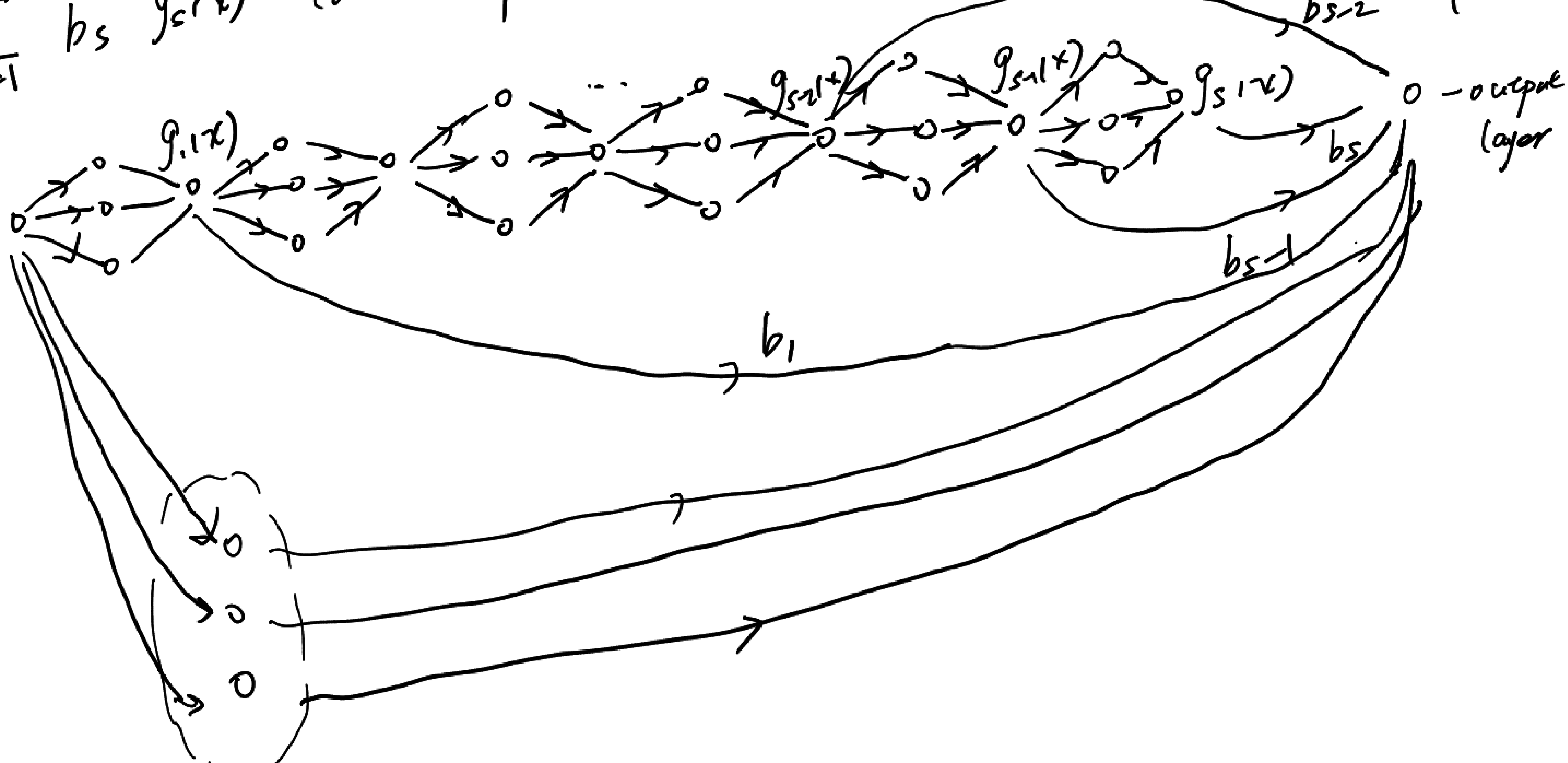


So $g_l(x) = \underbrace{g \circ \cdots \circ g}_{l}(x)$ is $O(l)$ layer-NN with width 3.

From (b), $P_2 f^*(x) - P_{2^l} f^*(x) = \sum_{s=1}^{l} \frac{g_s(x)}{2^{2s}}$

$P_{2^l} f^*(x) = P_2 f^*(x) - \sum_{s=1}^{l} \frac{g_s(x)}{2^{2s}}$ Let $b_s = -\frac{1}{2^{2s}}$

$\sum_{s=1}^{l} b_s g_s(x)$ can be represented as $O(l)$ depth, $O(1)$ width NN with skip connection



As in figure above, $P_{2^l} f^*(x) = P_2 f^*(x) + \sum_{s=1}^{l} b_s g_s(x)$

is a NN with width 6 and depth $2l+1$ with skip connection. #

3 Dropout for linear regression

(a) $f(x;\beta)=\beta^T x \quad \beta, x \in \mathbb{R}^d \quad \hat{R}(\beta)=\frac{1}{n}\sum_{i=1}^{n}(f(x_i;\beta)-y_i)^2$

$\tilde{\beta}=p\beta \quad w_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}^2$.

Show that $\hat{R}_{drop}(\beta)=\hat{R}(\tilde{\beta})+\frac{1-p}{p}\sum_{j=1}^{d}w_j\tilde{\beta}_j^2$

Proof.

$$RHS=\frac{1}{n}\sum_{i=1}^{n}(p\beta^T x_i-y_i)^2+\frac{p(1-p)}{n}\sum_{j=1}^{d}\sum_{i=1}^{n}x_{ij}^2\beta_j^2$$

$$LHS=\mathbb{E}_{\xi\sim\pi}[\hat{R}(\beta\odot\xi)]$$

$$=\mathbb{E}_{\xi\sim\pi}\frac{1}{n}\sum_{i=1}^{n}(f(x_i;\beta\odot\xi)-y_i)^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\xi\sim\pi}(f(x_i;\beta\odot\xi)-y_i)^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}y_i^2-\frac{2}{n}\sum_{i=1}^{n}y_i\mathbb{E}_{\xi\sim\pi}f(x_i;\beta\odot\xi)+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\xi\sim\pi}f^2(x_i;\beta\odot\xi)$$

$$=\frac{1}{n}\sum_{i=1}^{n}y_i^2-\frac{2}{n}\sum_{i=1}^{n}y_i p\beta^T x_i+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\xi\sim\pi}\left(\sum_{j=1}^{d}\xi_j x_{ij}\beta_j\right)^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}y_i^2-\frac{2}{n}\sum_{i=1}^{n}y_i p\beta^T x_i+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\xi\sim\pi}\left(\sum_{j=1}^{d}\xi_j x_{ij}^2\beta_j^2+2\sum_{j<k}\xi_j\xi_k x_{ij}x_{ik}\beta_j\beta_k\right)$$

$$=\frac{1}{n}\sum_{i=1}^{n}y_i^2-\frac{2}{n}\sum_{i=1}^{n}y_i p\beta^T x_i+\frac{1}{n}\sum_{i=1}^{n}p\sum_{j=1}^{d}x_{ij}^2\beta_j^2$$

$$+\frac{2}{n}\sum_{i=1}^{n}\sum_{j<k}p^2 x_{ij}x_{ik}\beta_j\beta_k$$

$$RHS=LHS \iff \frac{1}{n}\sum_{i=1}^{n}p^2\left(\sum_{j=1}^{n}\beta_j x_{ij}\right)^2+\frac{p-p^2}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}x_{ij}^2\beta_j^2=$$

$$\frac{1}{n}\sum_{i=1}^{n}p\sum_{j=1}^{d}x_{ij}^2\beta_j^2+\frac{2}{n}\sum_{i=1}^{n}\sum_{j<k}p^2x_{ij}x_{ik}\beta_j\beta_k$$

It sufficies to prove $\forall\, 1\le i\le n$,

$$p\left(\sum_{j=1}^{d}\beta_j x_{ij}\right)^2+(1-p)\sum_{j=1}^{d}x_{ij}^2\beta_j^2=\sum_{j=1}^{d}x_{ij}^2\beta_j^2+2\sum_{j<k}p\,x_{ij}x_{ik}\beta_j\beta_k$$

This holds obviously. #

(b) This shows dropout traing of $f(x;\beta)$

is regularied training of $f(x;p\beta)$ (ridge regression)

with controlling parameter $\lambda = \frac{1-p}{p}$

i.e. $\hat{R}_{dropout}(\beta) = \hat{R}(p\beta) + \lambda \cdot T(\beta)$

which explains why it improves generalization.

Parameter p controls regularization parameter λ

If $p=1$, no dropout

The smaller p is, the stronger regularization is. #