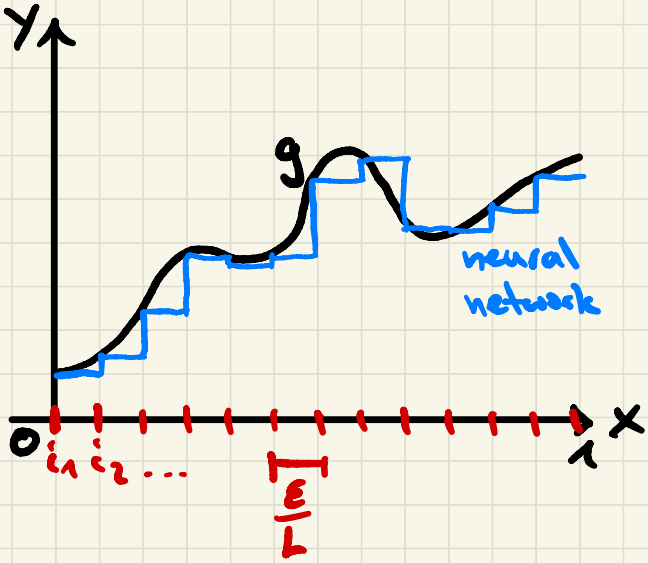
Proof for L-Lipschitz functions

Assume $g: [0,1] \to \mathbb{R}$ is L-Lipschitz. Then we can construct an appropriate 2-layer neural network using two steps:



1. Partition the x-axis into intervals of size $\frac{\varepsilon}{L}$.
2. For each interval k, construct an ANN f_k such that

$$f_k(x) = \begin{cases} g(i_k) & \text{if } x \in [i_k, i_{k+1}] \\ 0 & \text{else} \end{cases}$$

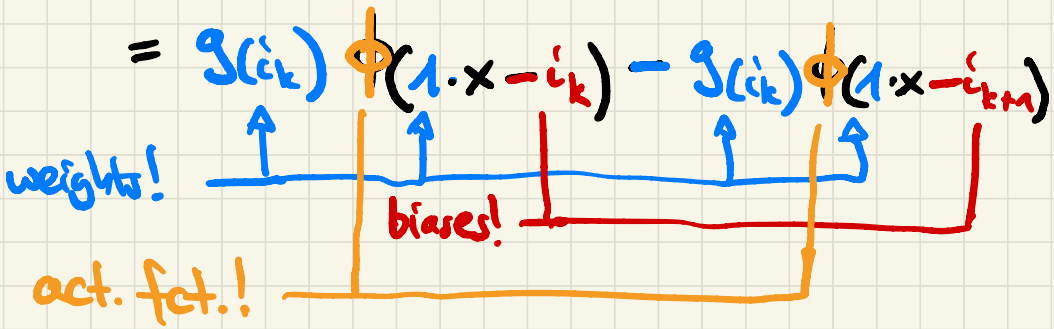
1. is a trivial step. For 2., we choose the following activation function:

$$\phi(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{else} \end{cases}$$

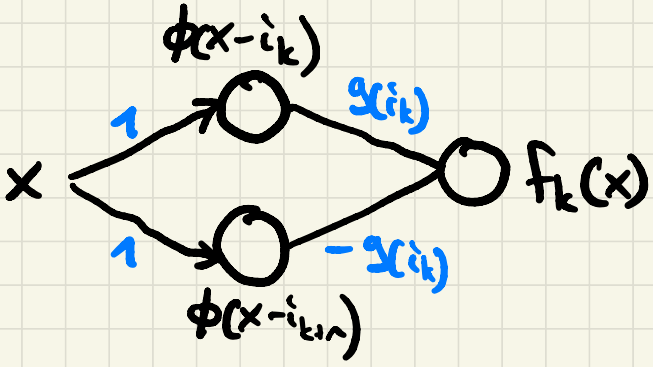
Then the following construction works:

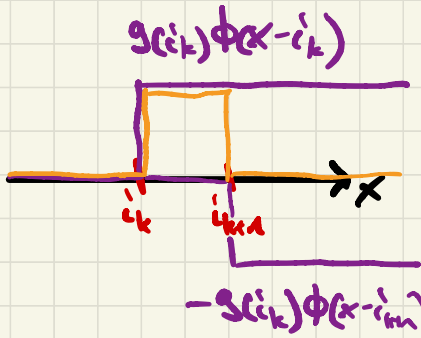
$$f_k(x) = g(i_k) \cdot [\phi(x - i_k) - \phi(x - i_{k+1})]$$

$$= g(i_k)\,\phi(1 \cdot x - i_k) - g(i_k)\,\phi(1 \cdot x - i_{k+1})$$

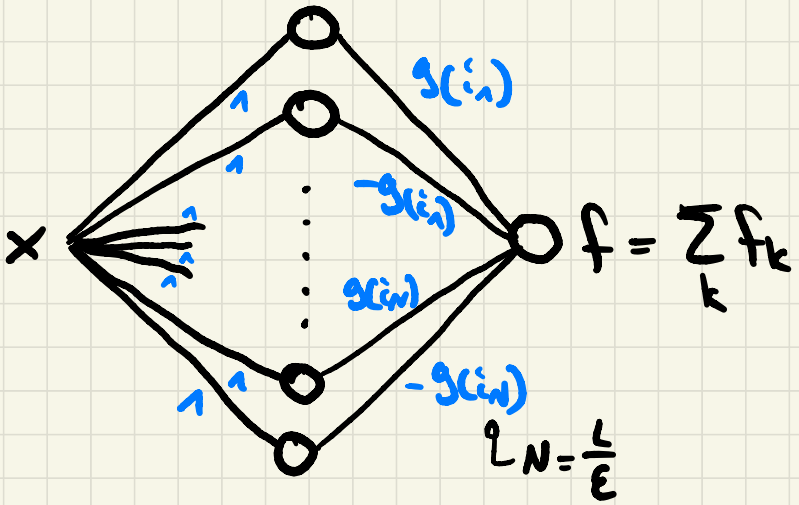


This is a 2-layer neural network!





Our final function is $f(x) = \sum_k f_k(x)$, which is simply a wider 2-layer neural network.



With this construction, we can complete the proof:

$$|f(x) - g(x)|_\infty = \left| \sum_j f_j(x) - g(x) \right|_\infty$$

pick interval with highest deviation →

$$= \max_{i_k} |g(i_k) - g(x)|_\infty$$

g is Lipschitz →

$$\leq L \cdot |i_k - x|_\infty$$

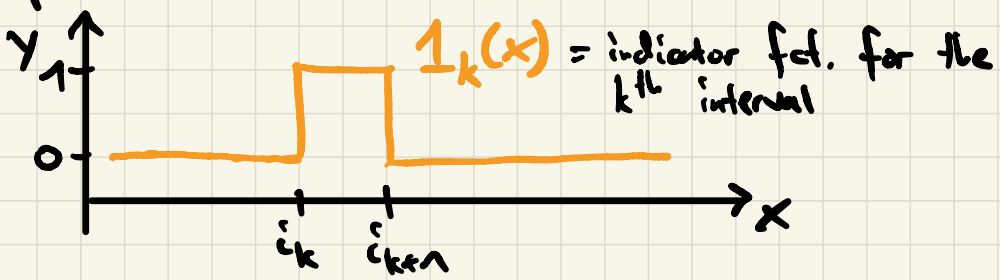
$$\leq L \cdot \frac{\varepsilon}{L}$$

$$= \varepsilon$$

□

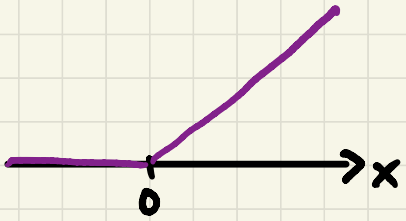
Case of ReLU act. fcts.:

Our proof relied on the fact that we can construct indicator functions for every interval



We achieved this using threshold functions as act. fcts. But: in modern ANNs, these are rarely used. Far more famous are ReLUs:

$$\text{ReLU}(x) = \max(0, x)$$

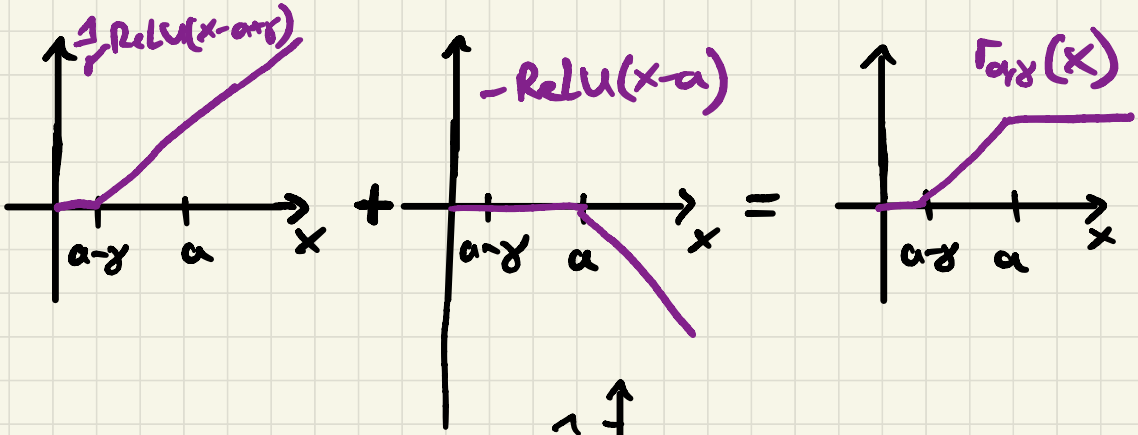


To show that the theorem also holds (up to constants) for ReLUs, we show that 1_k can be constructed using ReLUs!

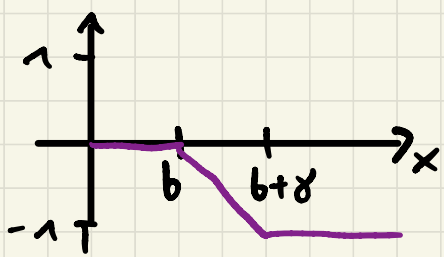
Indicator fct.: Assume $a \in \mathbb{R}$, $\gamma > 0$. Then

$$\Gamma_{a,\gamma}(x) = \frac{1}{\gamma}\left[\mathrm{ReLU}(x-a+\gamma) - \mathrm{ReLU}(x-a)\right]$$

Looks like:

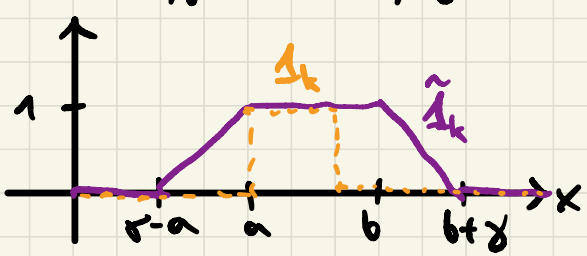


Also note that $\Gamma_{b,-\gamma}$ is:



Thus, we can approximate 1_k, with $a = i_k$, $b = i_{k+1}$,

using $\lambda_k(x) = \Gamma_{a,\gamma}(x) + \Gamma_{b,-\gamma}(x)$

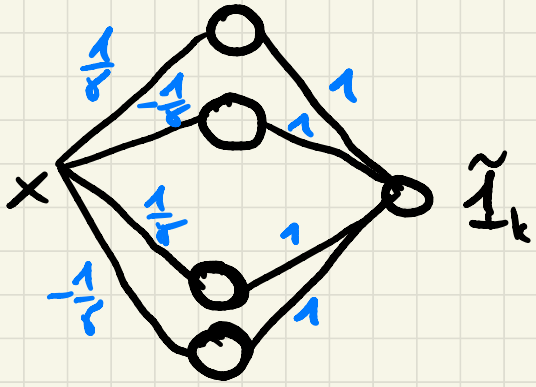


That's a bump!

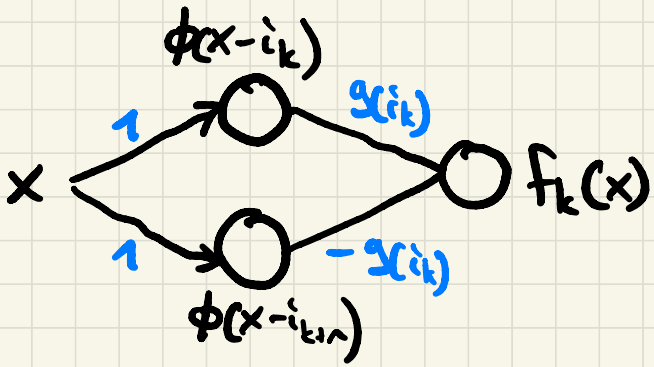
For smaller values of γ, we get closer to the actual indicator fct.! In fact, we can find a γ such that

$$\| 1_k(x) - \tilde{1}_k(x) \|_\infty < \varepsilon$$

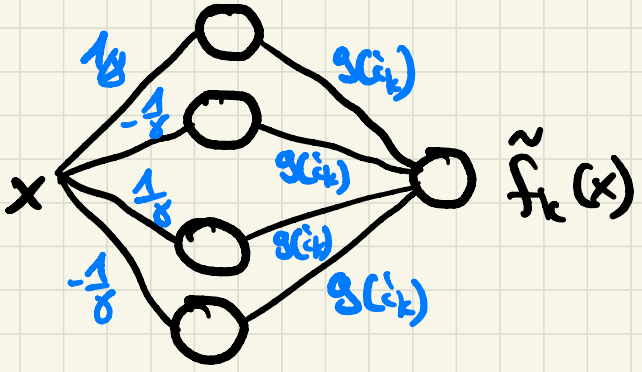
Summary: we can approximate the indicator fct. to given error ε using a 2-layer ReLU neural network!



With threshold act. fct., our block for each interval looked like:



With ReLU, it now looks like:



With $\tilde{f}(x) = \sum_k \tilde{f}_k$, we get:

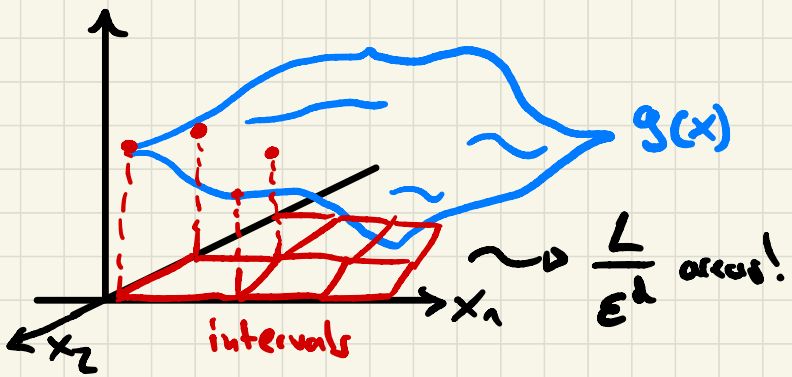
$$|\tilde{f} - g|_\infty \leq \underbrace{|\tilde{f} - f|_\infty}_{\text{set } \delta \text{ s.t. } < \varepsilon} + \underbrace{|f - g|_\infty}_{< \varepsilon} < 2\varepsilon \qquad \square$$

"approx. indicators using ReLUs" — "approx. g using indicator fcts."

Multivariate functions

This is pretty much the same idea as in 1D: we construct d-dimensional indicator regions!

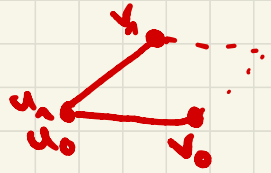
Example in 2D:



So how can we construct a fct. that is 1 on only one tile and 0 otherwise?

Assume our tile has corners $\{(u_j, v_j) \mid j \in [1, d]\}$

We first construct indicators for every dimension using threshold act. fcts. ϕ:



$$h_j(x) = \phi(x - u_j) - \phi(x - v_j)$$

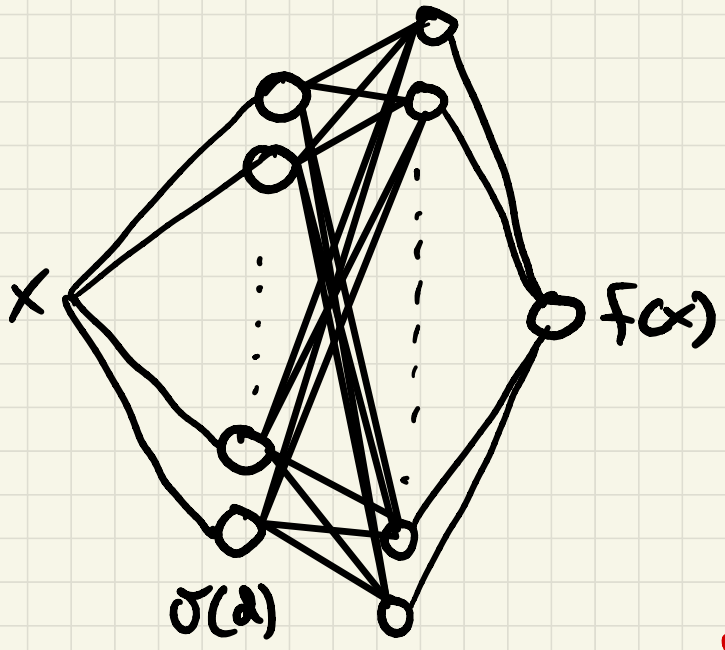
We then check if all interval conditions are satisfied as follows:

$$h(x) = \phi\Big(\underbrace{\sum_{i=1}^{d} h_i(x_i) - (d-1)}_{\text{only } =1 \text{ if } \underline{\text{all}}}\Big)$$

$h_i(x_i) = 1,$

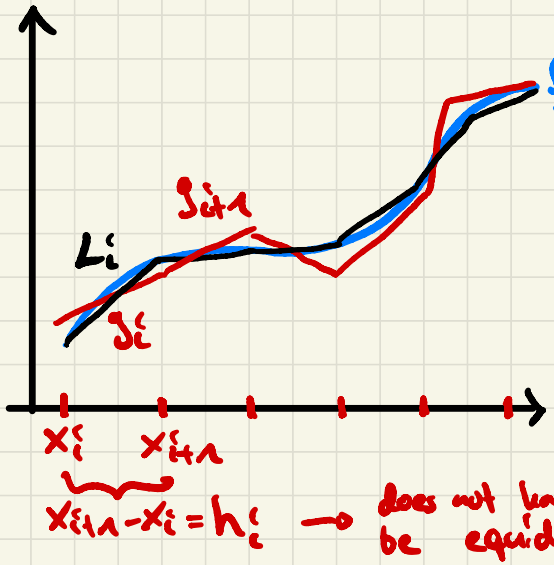
0 otherwise.

Our final network then has 3 layers:



Proof: $\frac{1}{p^2}$ - bound

Let's start with a picture.



f: the function we want to fit

g_i: the pieces of our network

x_i: segmentation of input

L_i: linear interpolation of f using x_i

→ does not have to be equidistant for all i

We assume that the x_i were chosen such that

(A) $\|f(x) - g_i(x)\| \leq \varepsilon$ for $x \in [x_i, x_{i+1}]$ $\forall i$.

The proof consists of two parts:

1) Show that ε is related to the interpol. error of L: $\|e_i\| = \|f(x) - L_i(x)\| \leq 2\varepsilon$

2) Show that the $\max\|e_i\| = \frac{h^2}{8}\|f''(m_i)\| + \sigma(h_i^2)$ with $m_i = \frac{x_i + x_{i+1}}{2}$.

From 1) and 2), we get our result:

$$\|e_i\| \leq 2\varepsilon \quad \text{for all } x \in [x_i, x_{i+n}]$$

this includes $\Leftrightarrow$ the max. error

$$\max \|e_i\| = \frac{h_i^2}{8} \|f''(m_i)\| \leq 2\varepsilon$$

in the limit $h_i \to 0 \triangleq \varepsilon \to 0$!

take $\sqrt{\ }$ $\Leftrightarrow$

$$\frac{h_i}{4} \sqrt{\|f''(m_i)\|} \leq \sqrt{\varepsilon}$$

sum over $\Leftarrow$ all segments

$$\sum_i \frac{h_i}{4} \sqrt{\|f''(m_i)\|} \leq p\sqrt{\varepsilon}$$

$\sum_i$ we sum over p pieces

in the limit, this becomes $\frac{1}{4} \int_{x_i}^{x_{i+n}} \sqrt{f''(h)}\, dh = C$

$$\rightarrow \varepsilon \geq \frac{C}{p^2}$$

Now lets first prove 1):

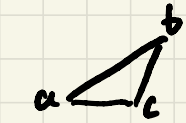
Our interpolation is given by:

$$L_i(x) = f(x_i) + \frac{f(x_{i+n}) - f(x_i)}{x_{i+n} - x_i}(x - x_i)$$

check: $L_i(x_i) = f(x_i)$, $L_i(x_{i+n}) = f(x_{i+n})$ ✓

Then: $\|e_i\| = \|f(x) - L_i(x)\|$ for $x \in [x_i, x_{i+n}]$

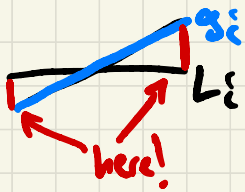
$\overset{\text{add zero}}{=} \|f(x) - g_i(x) + g_i(x) - L_i(x)\|$



$|a-b| \leq |a-c| + |a-b|$

$\overset{\Delta\text{-ineq.}}{\leq} \underbrace{|f - g_i|}_{\text{using (A), this is } \leq \varepsilon} + \underbrace{|g_i - L_i|}_{\text{these are two lines on finite intervals!}}$

$\Rightarrow$ They are furthest apart at the ends



$\Rightarrow |g_i - L_i| \leq \max\{|g_i(x_i) - \overbrace{L_i(x_i)}^{=f(x_i)}|, |g_i(x_{i+n}) - \underbrace{L_i(x_{i+n})}_{=f(x_{i+n})}|\}$

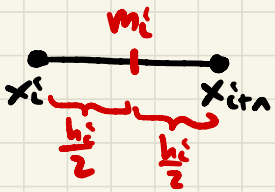
$\leq |g_i(x_i) - f(x_i)|$

$\overset{\text{using (A)}}{\leq} \varepsilon$

$\Rightarrow \|e_i\| \leq 2\varepsilon$ $\blacksquare$

Now the second part. For this we Taylor expand **all** terms in $L_i(x)$ as well as $f(x)$ around $m_i = \frac{x_i + x_{i+1}}{2}$.

Note the following: $x_i = m_i - \frac{h_i}{2}$ $\quad x_{i+1} = m_i + \frac{h_i}{2}$



and $\quad x = m_i + \alpha(x)\frac{h_i}{2}$ for $x \in [x_i, x_{i+1}]$ and $\alpha \in [-1, 1]$

In particular, we get $x - x_i = (\alpha + 1)\frac{h_i}{2}$

Lets look at $L_i(x) = f(x_i) + \frac{f(x_{i+1}) - f(x_i)}{\underbrace{x_{i+1} - x_i}_{= h_i}} \underbrace{(x - x_i)}_{= (\alpha+1)\frac{h_i}{2}}$

Lets Taylor:

$$f(x_i) = f(m_i - \tfrac{h_i}{2}) = f(m_i) - f'(m_i)\frac{h_i}{2} + \frac{f''(m_i)}{2}\frac{h_i^2}{4} + \mathcal{O}(h_i^3)$$

$$f(x_{i+1}) = f(m_i) + f'(m_i)\frac{h_i}{2} + \frac{f''(m_i)}{2}\frac{h_i^2}{4} + \mathcal{O}(h_i^3)$$

$$\Rightarrow L(x_i) = f(m_i) + \alpha\frac{h_i}{2}f'(m_i) + f''(m_i)\frac{h_i^2}{8} + \mathcal{O}(h_i^3)$$

Lets also Taylor expand $f(x)$:

$$f(x) = f(m_i + \alpha \frac{h_i}{2})$$

$$= f(m_i) + f'(m_i)\,\alpha \frac{h_i}{2} + f''(m_i)\,\frac{\alpha^2 h_i^2}{8} + \sigma(h_i^3)$$

Combining these, we get:

$$\|f(x) - L_i(x)\| = \left| \frac{f''(m_i) h_i^2}{8} \right| \cdot |\alpha^2 - 1|$$

Which has its maximum at $\alpha = 0$ (where $|\alpha^2 - 1|$ has its max, since $\alpha \in [-1, 1]$ and $|\alpha^2 - 1| \in [0, 1]$)

□