5. Approximation: 2-Layer ReLU net

Recap

- Last class, we showed that 3-layer ReLU net are universal approximators
  - Covered *d*-dimensional inputs
  - Constructive proof
    - i.e., explicit construction given
  - L1 norm
  - Lipschitz function

Today

- We prove that 2-layer sigmoid / ReLU networks are universal approximators
  - Non-constructive proof
    - i.e., no explicit construction will be given
  - Uniform norm (!)
  - Continuous function (!)

Regret

- **Question.** In the last class, why did we need three layers?

Regret

- **Question.** In the last class, why did we need three layers?

  - Each d-dimensional pulse $\mathbf{1}\{\mathbf{x} \in R_i\}$ was a 3-layer net
    - Each R_i is a hypercube

$$\mathbf{1}[\mathbf{x} \in R_i] = \prod_{i=1}^{d} \mathbf{1}[\mathbf{x}_j \in [a_i, b_i]]$$

    - **1st hidden layer.** Construct a 1D pulse
    - **2nd hidden layer.** Conduct a "multiplication" of 1D pulses
      - Add 1D pulses
      - Subtract $d - 1$
      - ReLU out negative parts

Regret

- **Question.** In the last class, why did we need three layers?
  - Each d-dimensional pulse $\mathbf{1}\{\mathbf{x} \in R_i\}$ was a 3-layer net
    - Each R_i is a hypercube

$$\mathbf{1}[\mathbf{x} \in R_i] = \prod_{i=1}^{d} \mathbf{1}[\mathbf{x}_j \in [a_i, b_i]]$$

    - **1st hidden layer.** Construct a 1D pulse
    - **2nd hidden layer.** Conduct a "multiplication" of 1D pulses
      - Add 1D pulses
      - Subtract $d - 1$
      - ReLU out negative parts

Can we remove this?

Regret

- **Question.** In the last class, why did we need three layers?
  - Each d-dimensional pulse $\mathbf{1}\{\mathbf{x} \in R_i\}$ was a 3-layer net
    - Each R_i is a hypercube

$$\mathbf{1}[\mathbf{x} \in R_i] = \prod_{i=1}^{d} \mathbf{1}[\mathbf{x}_j \in [a_i, b_i]]$$

    - **1st hidden layer.** Construct a 1D pulse
    - **2nd hidden layer.** Conduct a "multiplication" of 1D pulses
      - Add 1D pulses
      - Subtract $d - 1$
      - ReLU out negative parts

Can we remove this? Yes, if two-layer nets are "closed under multiplication"

Formalisms

- To formalize this, consider a set of all two-layer networks
  - The set of depth-2, width-m nets will be defined as:

$$\mathscr{F}_{\sigma,d,m} = \left\{ \mathbf{x} \mapsto \sum_{i=1}^{m} a_i \sigma(\mathbf{x}^\top \mathbf{w}_i + b_i) \;\middle|\; \mathbf{w}_i \in \mathbb{R}^d, b_i \in \mathbb{R}, a_i \in \mathbb{R}, i \in [m] \right\}$$

  - The set of depth-2 nets will be defined as:

$$\mathscr{F}_{\sigma,d} = \bigcup_{m \in \mathbb{N}} \mathscr{F}_{\sigma,d,m}$$

- Let's study the properties of these hypothesis spaces

Hypothesis space as an algebra

Claim 1. The set $\mathscr{F}_{\sigma,d,m}$ is closed under scalar multiplication, i.e.,

$$f \in \mathscr{F}_{\sigma,d,m}, c \in \mathbb{R}_{\neq 0} \quad \longrightarrow \quad (c \cdot f) \in \mathscr{F}_{\sigma,d,m}$$

Hypothesis space as an algebra

Claim 2. The set $\mathscr{F}_{\sigma,d}$ is closed under addition, i.e.,

$$f, g \in \mathscr{F}_{\sigma,d} \quad \longrightarrow (f+g) \in \mathscr{F}_{\sigma,d}$$

Hypothesis space as an algebra

Claim 3 (Cosine). The set $\mathscr{F}_{\cos,d}$ is closed under multiplication, i.e.,

$$f, g \in \mathscr{F}_{\cos,d} \quad \longrightarrow (f \cdot g) \in \mathscr{F}_{\cos,d}$$

Stone-Weierstrass

- Now, we describe our main technical tool
  - Will not prove it, sadly

Theorem 2.2. (Stone-Weierstrass).

Let a set of functions $\mathscr{F}$ be given as follows.

- Each $f \in \mathscr{F}$ is continuous
- For any x, there exists $f \in \mathscr{F}$ such that $f(x) \neq 0$
- For any $x \neq x'$, there exists $f \in \mathscr{F}$ such that $f(x) \neq f(x')$
- $\mathscr{F}$ is closed under multiplications and vector space operations (i.e., algebra)

Then, $\mathscr{F}$ is a universal approximator:

For every continuous $f : \mathbb{R}^d \to \mathbb{R}$ and $\varepsilon > 0$, there exists $f \in \mathscr{F}$ with

$$\sup_{x \in [0,1]^d} |f(x) - g(x)| \leq \varepsilon$$

Exercises

Exercise 1. (Cosines).

Show that $\mathscr{F}_{\cos,d}$ satisfies the Stone-Weierstrass conditions, and thus universal approximators

- Each $f \in \mathscr{F}$ is continuous
- For any x, there exists $f \in \mathscr{F}$ such that $f(x) \neq 0$
- For any $x \neq x'$, there exists $f \in \mathscr{F}$ such that $f(x) \neq f(x')$
- $\mathscr{F}$ is closed under multiplications and vector space operations

Exercises

Exercise 2. (Exponentials).

Show that $\mathscr{F}_{\exp,d}$ satisfies the Stone-Weierstrass conditions, and thus universal approximators

- Each $f \in \mathscr{F}$ is continuous
- For any x, there exists $f \in \mathscr{F}$ such that $f(x) \neq 0$
- For any $x \neq x'$, there exists $f \in \mathscr{F}$ such that $f(x) \neq f(x')$
- $\mathscr{F}$ is closed under multiplications and vector space operations

Sigmoidal functions

- Now, we state our main result today

Theorem 2.3. (Hornik, Stinchecombe, and White, 1989)

Suppose that $\sigma : \mathbb{R} \to \mathbb{R}$ is sigmoidal, i.e.,

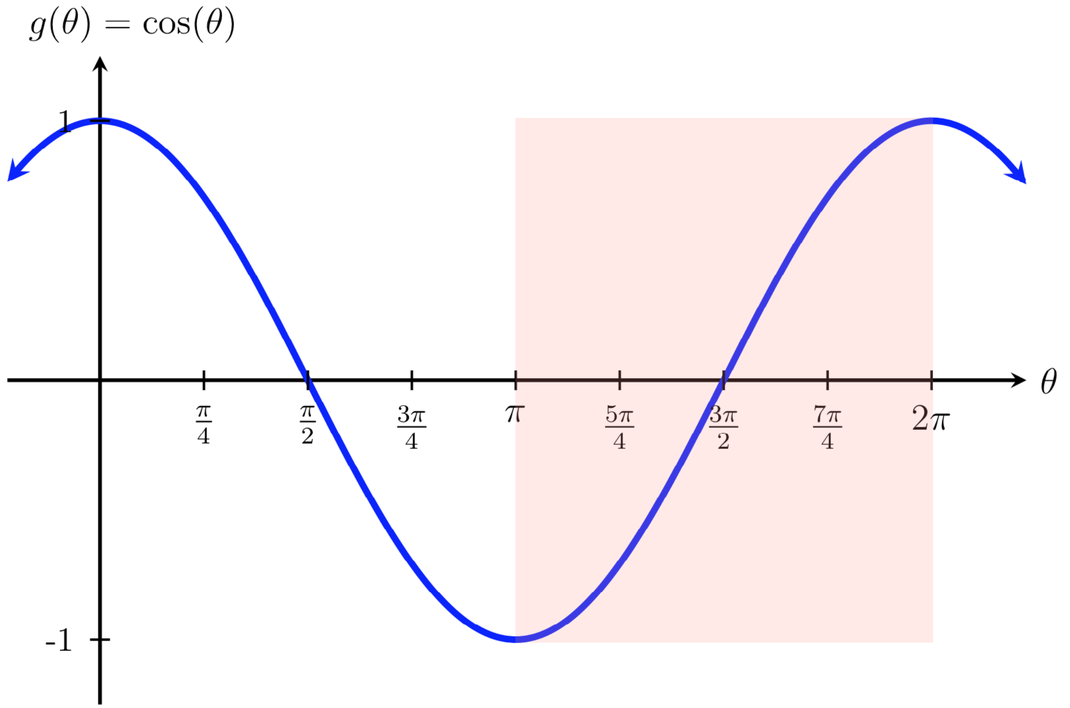
- Continuous
- Nondecreasing
- $\lim_{x \to -\infty} \sigma(x) = 0$ and $\lim_{x \to +\infty} \sigma(x) = 1$

Then, $\mathscr{F}_{\sigma,d}$ is universal.

- Unfortunately, validating SW conditions for general sigmoid is harder than it looks...

Sigmoidal functions

- Instead of a direct proof, we'll go through cosines
  - **Step 1.** Sigmoids can approximate "cosine sigmoids"
  - **Step 2.** Cosine sigmoids can represent cosines
  - **Step 3.** Cosines are universal approximators



Define "Cosine sigmoids" as:

$$\sigma_c(x) = \begin{cases} 0 & \cdots \quad x \leq 0 \\ 1 & \cdots \quad x \geq \pi \\ \dfrac{\cos(x+\pi)+1}{2} & \cdots \quad x \in (0,\pi) \end{cases}$$

Technical Lemma

Lemma A. (Any sigmoid can approximate another sigmoid)

For any sigmoids σ, σ' and $\varepsilon > 0$, there exists $f \in \mathcal{F}_{\sigma,1}$ such that

$$\sup_{x \in \mathbb{R}} |\sigma'(x) - f(x)| < \varepsilon$$

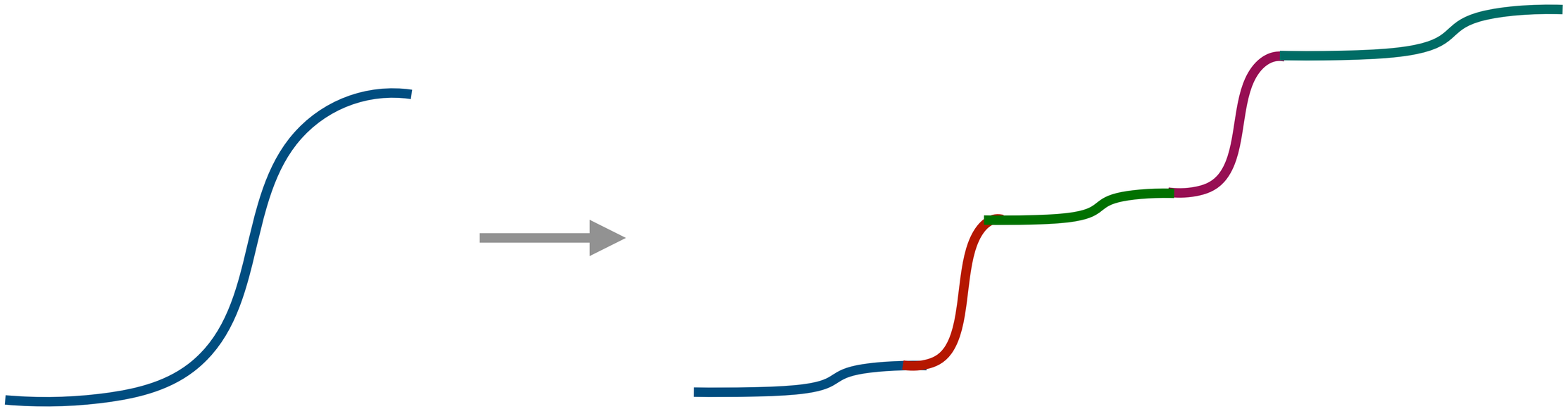
Technical Lemma

Lemma A. (Any sigmoid can approximate another sigmoid)

For any sigmoids σ, σ' and $\varepsilon > 0$, there exists $f \in \mathscr{F}_{\sigma,1}$ such that

$$\sup_{x \in \mathbb{R}} |\sigma'(x) - f(x)| < \varepsilon$$

- **Idea.** Copy-and-paste your sigmoid

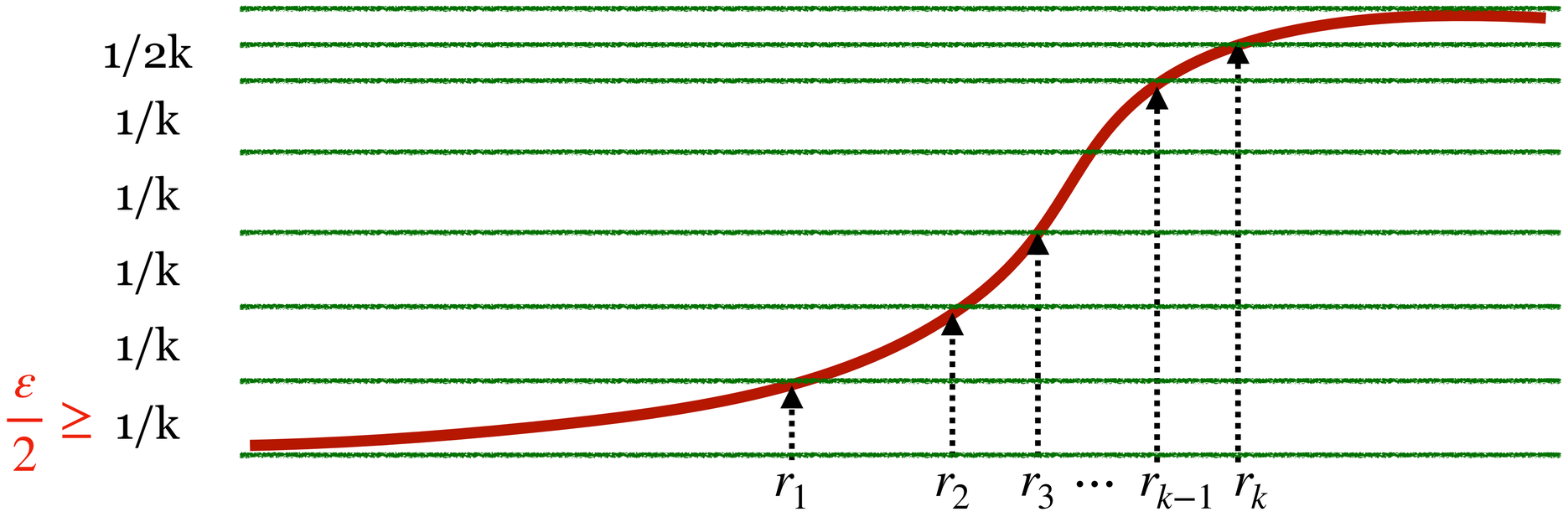


Proof

- **Step 1.** Divide the curve to be fit.
  - Choose some $k > 2/\varepsilon$ (number of pieces)
  - For $j \in \{1, \ldots, k-1\}$, select

$$r_j = \sup \left\{ x \;\middle|\; \sigma'(x) = \frac{j}{k} \right\}$$

(we'll let $r_k = \sup\{x \mid \sigma'(x) = 1 - 1/2k\}$)

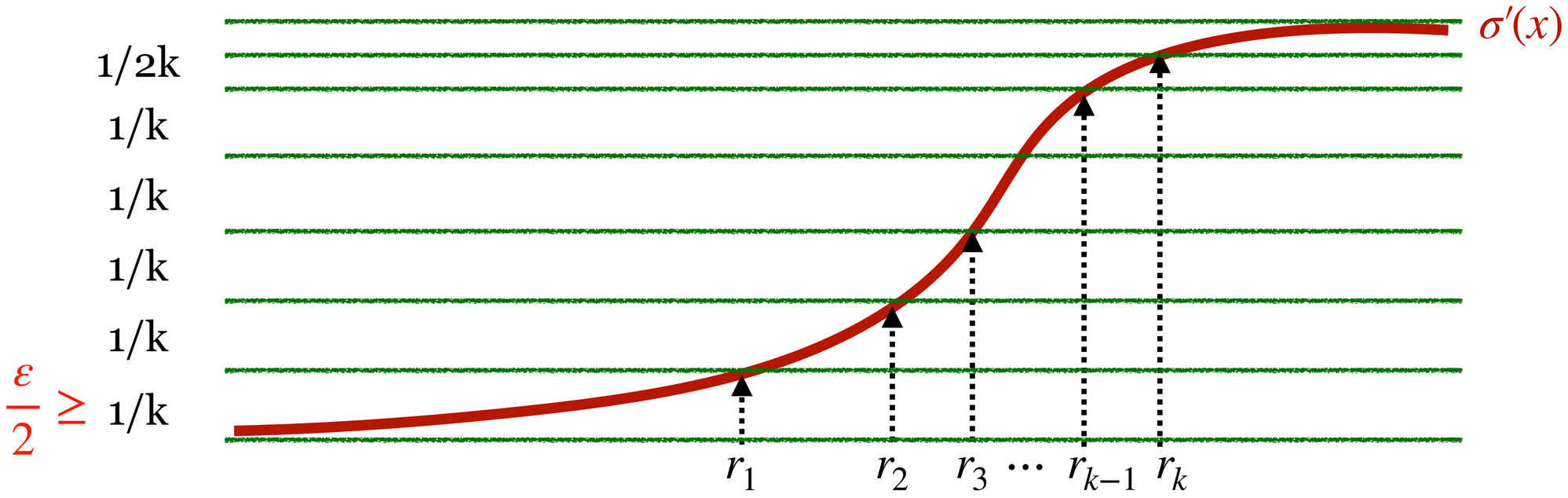


Proof

- **Step 1.** Divide the approximand $\sigma'(x)$
  - Choose some $k > 2/\varepsilon$ (number of pieces)
  - For $j \in \{1, \ldots, k-1\}$, select

$$r_j = \sup \left\{ x \;\middle|\; \sigma'(x) = \frac{j}{k} \right\}$$

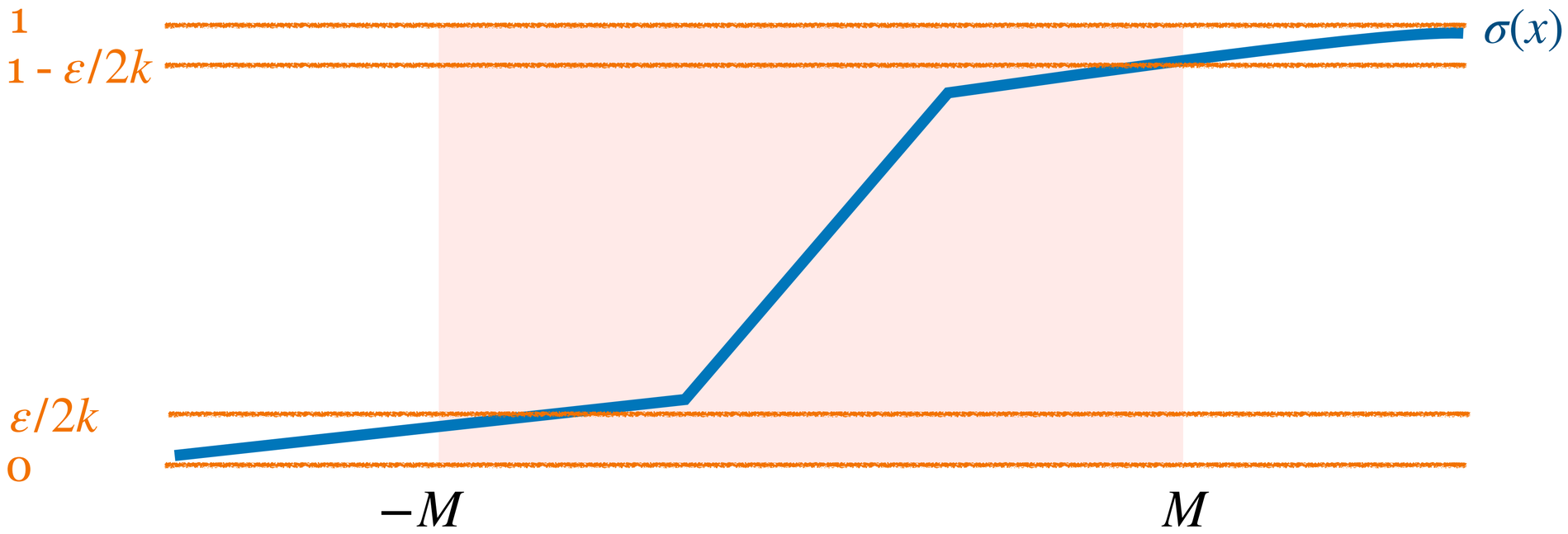
(we'll let $r_k = \sup\{x \mid \sigma'(x) = 1 - 1/2k\}$)



Proof

- **Step 2.** Choose the "effective region" of the approximator $\sigma(x)$
  - Choose M such that

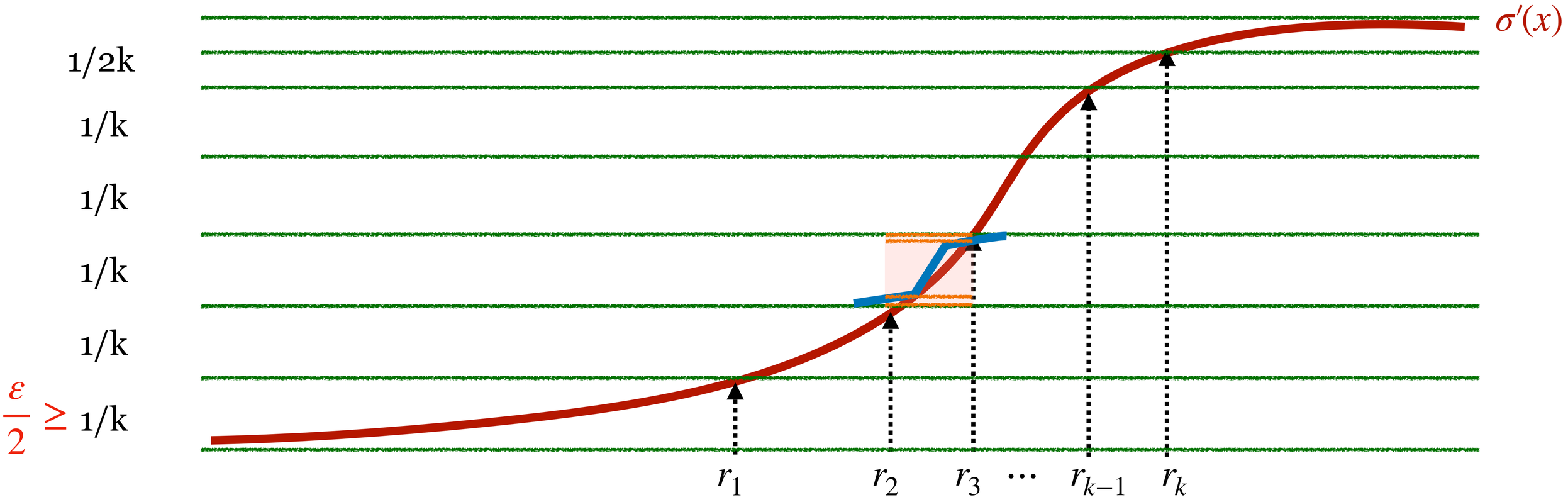
$$\sigma(-M) < \frac{\varepsilon}{2k}, \qquad \sigma(M) < 1 - \frac{\varepsilon}{2k}$$



Proof

- **Step 3.** Fit piecewise functions
  - Use one sigmoid for each of $\sigma'((-\infty, r_1]), \sigma'((r_1, r_2]), \ldots$
    - Blue can deviate from red by

$$1/k + k \times (\text{tail components}) \quad \leq \quad \varepsilon/2 + k \times (\varepsilon/2k) \quad = \quad \varepsilon$$



Technical Lemma

Lemma B. (Any sigmoid can approximate cosine)

For any sigmoids σ, $\varepsilon > 0$, $M > 0$, there exists $f \in \mathscr{F}_{\sigma,1}$ such that

$$\sup_{x \in [-M,M]} |f(x) - \cos(x)| < \varepsilon$$

- **Idea.**
  - Use Lemma A to approximate the cosine sigmoid
  - Overlap cosine sigmoids to get cosine

The Case of ReLU

- Handling ReLU is easy
  - Two ReLUs can generate “hard sigmoid”
  - Hard sigmoids can approximate cosine sigmoid

Next up

- Infinite-width limits of neural networks