8. Approximation: Benefits of depth

Recap

- We have shown several universal approximation results
  - Three-layer: O
  - Two-layer: O
  - One-layer: X
    - Thus, two layer is the minimum depth

This lecture

- **Question.** Why are deeper nets often better than shallower ones?

This lecture

- **Question.** Why are deeper nets often better than shallower ones?

- **Answer.** In terms of the approximation, deeper nets are more parameter-efficient

    - In particular, certain depth separation holds:
        - Deep nets can express some function with N neurons
        - Shallow nets cannot, with N neurons
            - Key question. What function is difficult to be learned by shallow nets?

    - We count #neurons here, but anything can be used for separation
        - e.g., norm

Case 1: Wedges

Wedge

- We are interested in the **wedge function** ✏️

$$\Delta(x) = 2 \cdot \sigma\left(x\right) - 4 \cdot \sigma\left(x - \frac{1}{2}\right) + 2 \cdot \sigma\left(x - 1\right)$$
$$= \begin{cases} 2x & \cdots \quad x \in [0,1/2], \\ 2 - 2x & \cdots \quad x \in [1/2,1] \\ 0 & \cdots \quad \text{otherwise} \end{cases}$$

- Expressible with a two-layer ReLU net with 3 neurons

Wedges and Wedges

- Think about the composition

$$\Delta^2(x) = \Delta \circ \Delta(x)$$

- **Question.** What would this function look like? ✏️

Wedges and Wedges and Wedges

- Now, consider the L-time composition

$$\Delta^L(x)$$

- **Question.** What would this look like? ✏️

Depths vs. Width

- For this Δ, we already have some ideas
    - **Deep.** For k wedges, we can express using $O(\log k)$ layers with constant width
    - **Shallow.** For k wedges, you need $O(k)$ neurons

    - Can we formally show that this is "necessary"?

Depths vs. Width

- **Difficulty.** Giving a lower bound for shallow nets
  - Upper bound (Achievability)

$$\min_{s \in S} \ell(s) \leq t$$

    - Easy, find a good s
  - Lower bound (Impossibility)

$$\min_{s \in S} \ell(s) \geq t$$

    - Difficult; check all s?

Main claim

- Here is what we'll prove today

Theorem 5.1.

Let $L \geq 2$. Let $f = \Delta^{L^2+2}$ be a ReLU net with $3L^2 + 6$ nodes and $2L^2 + 4$ layers.

Then, any ReLU net g with $\leq 2^L$ nodes and $\leq L$ layers cannot approximate f, i.e.,

$$\int_{[0,1]} |f(x) - g(x)| \, \mathrm{d}x \geq \frac{1}{32}$$

- What tools can we use?

Tool: Affine Pieces

Tool: Counting Affine Pieces

- **Idea.** We show that shallow nets have small number of affine pieces

Definition (Affine Pieces).

For any univariate function $f : \mathbb{R} \to \mathbb{R}$, let $N_A(f)$ denote the number of affine pieces of f: the minimum cardinality of a partition of $\mathbb{R}$, so that f is affine when restricted to each piece.

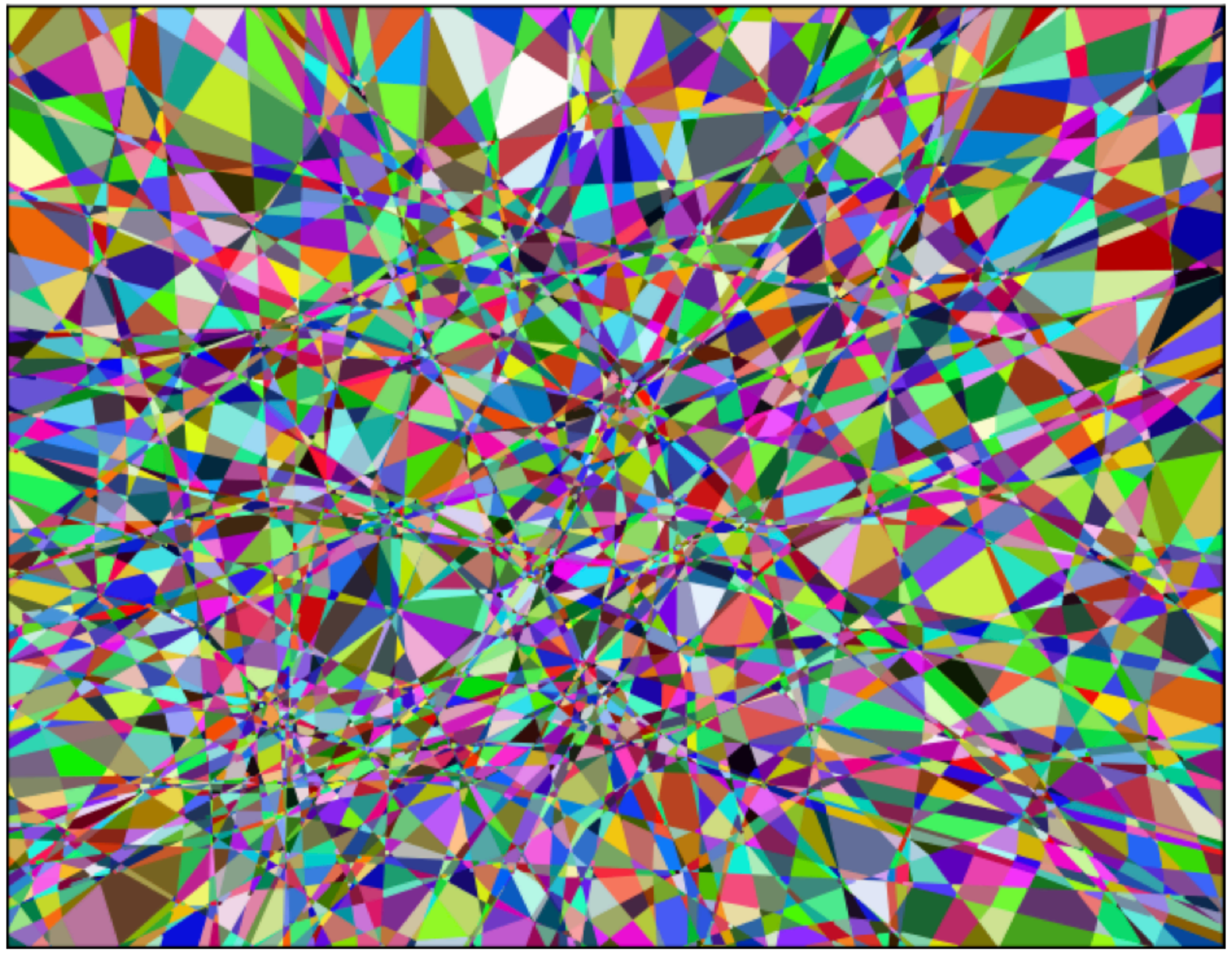


Figure 1. How many linear regions? This figure shows a two-dimensional slice through the 784-dimensional input space of vectorized MNIST, as represented by a fully-connected ReLU network with three hidden layers of width 64 each. Colors denote different linear regions of the piecewise linear network.

Hanin & Rolnick, "Complexity of Linear Regions in Deep Networks," ICML 2019

Basic properties

- We have the following lemma

Lemma 5.2.

Let functions f, g and a scalar c be given. Then, we have:

- $N_A(0) = 1$
- $N_A(c \cdot f) = N_A(f)$ when $c \neq 0$
- $N_A(f + c) = N_A(f)$
- $N_A(f + g) \leq N_A(f) + N_A(g)$
- $N_A(f \circ g) \leq N_A(f) \cdot N_A(g)$

- **Proof idea.** Utilize the partitions, and the definition of linearity

Bounding the number of affine regions

- Using the properties, we can show the following lemma.

Lemma 5.1.

Let $f : \mathbb{R} \to \mathbb{R}$ be a ReLU network with L layers, of widths $m_1, \ldots, m_L$.

- Let $g : \mathbb{R} \to \mathbb{R}$ denote the output of some node at layer i. Then, we have

$$N_A(g) \leq 2^i \cdot \left(\prod_{j<i} m_j \right)$$

- Let $\bar{m} = (m_1 + m_2 + \cdots + m_L)/L$. Then, we have

$$N_A(f) \leq 2^L \cdot \bar{m}^L$$

- Idea?

Bounding the number of affine regions

Let $f : \mathbb{R} \to \mathbb{R}$ be a ReLU network with L layers, of widths $m_1, \ldots, m_L$.

- Let $g : \mathbb{R} \to \mathbb{R}$ denote the output of some node at layer i. Then, we have

$$N_A(g) \leq 2^i \cdot \left(\prod_{j<i} m_j \right)$$

- **Proof idea.** Prove by induction

Bounding the number of affine regions

Let $f : \mathbb{R} \to \mathbb{R}$ be a ReLU network with L layers, of widths $m_1, \ldots, m_L$.

- Let $g : \mathbb{R} \to \mathbb{R}$ denote the output of some node at layer i. Then, we have

$$N_A(g) \leq 2^i \cdot \left(\prod_{j<i} m_j\right)$$

- Let $\bar{m} = (m_1 + m_2 + \cdots + m_L)/L$. Then, we have

$$N_A(f) \leq 2^L \cdot \bar{m}^L$$

- **Proof idea.** With the first claim, suffices to show that

$$\prod m_j \leq \bar{m}^L$$

  - Taking log, suffices to show that

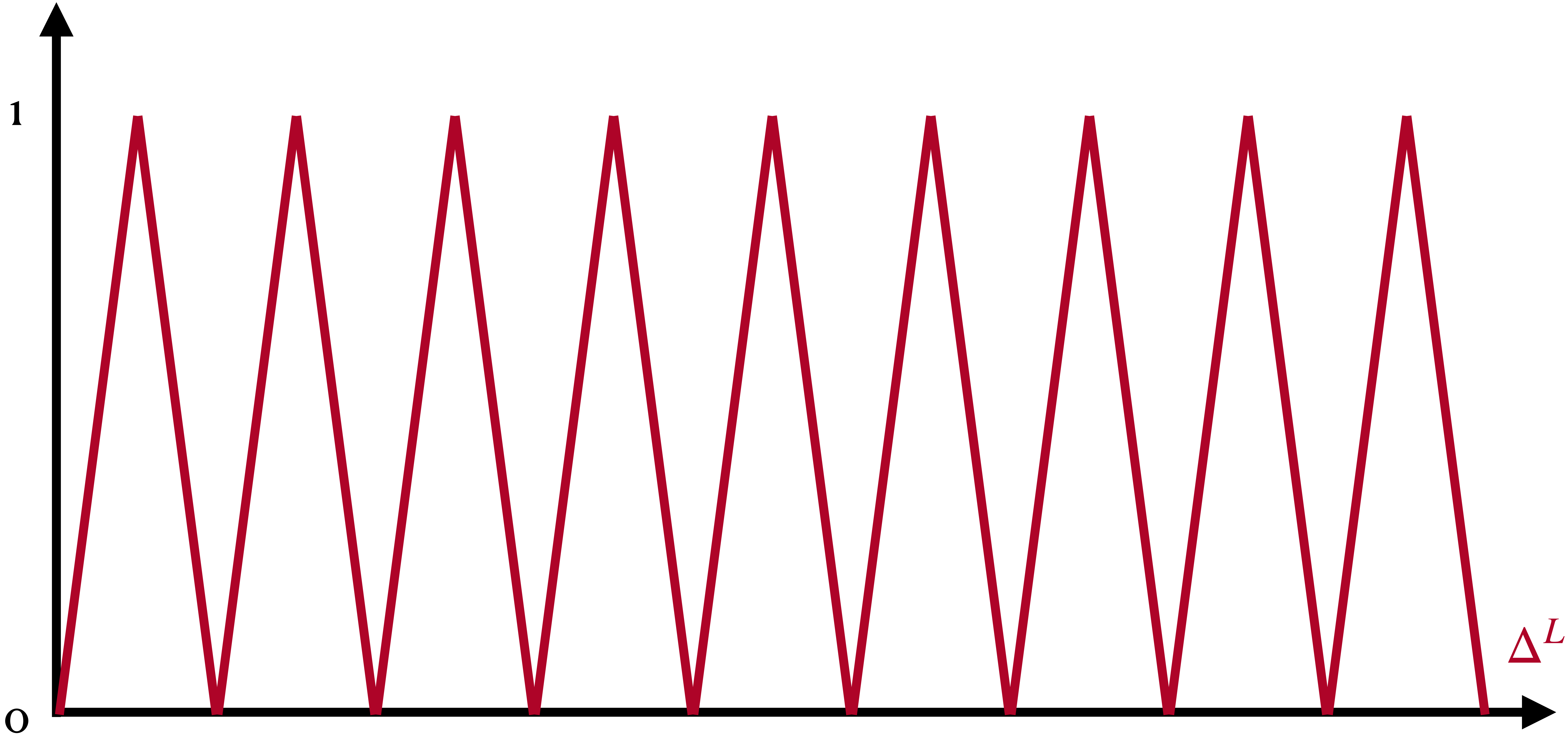
$$\frac{1}{L} \sum \log(m_j) \leq \log(\bar{m})$$

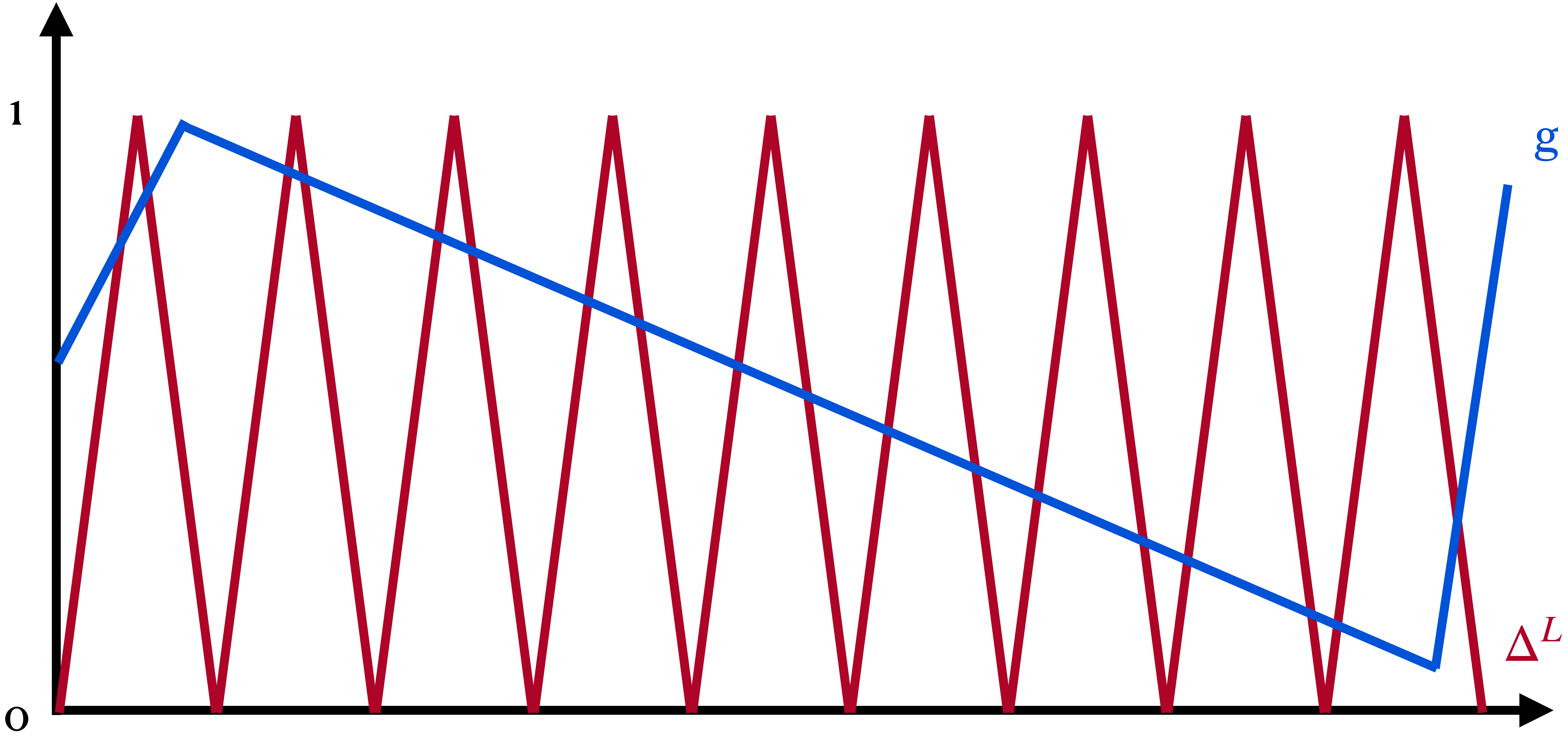
  - Ring a bell?

Bounding the number of affine regions

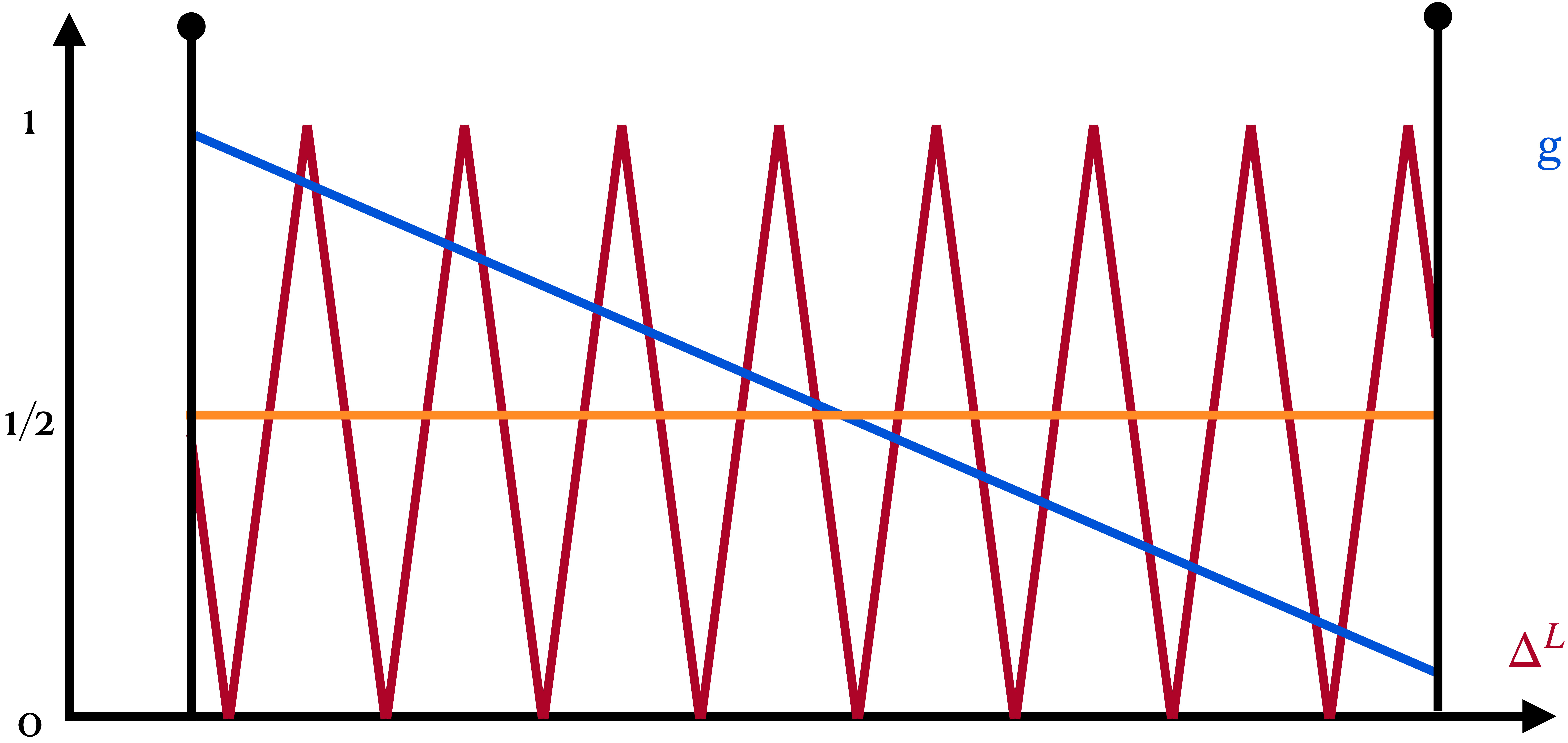
- Now we know that shallow nets have an UB on #affine regions

- **Question.** How do we translate it to a LB on L_1 approximate error for Δ^L?
  - Very neat graphical argument


1
0
Δ^L

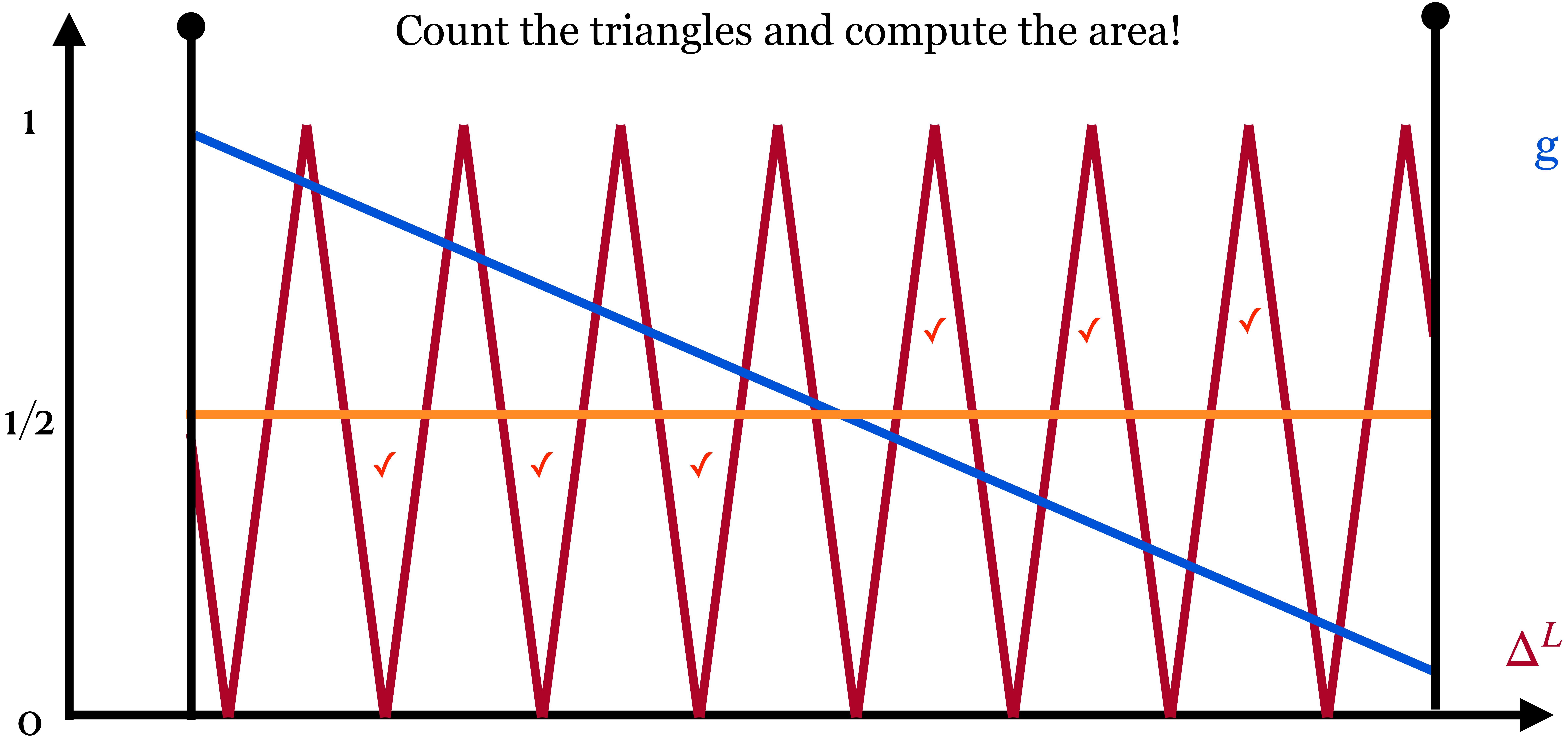


1
0
g
Δ^L



1
1/2
0
g
Δ^L

Count the triangles and compute the area!



Case 2: x^2

x^2

- Wedges were rather special functions
- **Question.** Do depth separation hold for simple functions as well?
- **Answer.** Yes—we'll look at the case of x^2
  - Note. A technique for expressing x^2 can be used to express xy

$$xy = \frac{1}{2}\left((x+y)^2 - x^2 - y^2\right)$$

x^2-approximating ReLU net

- We start from the fact that

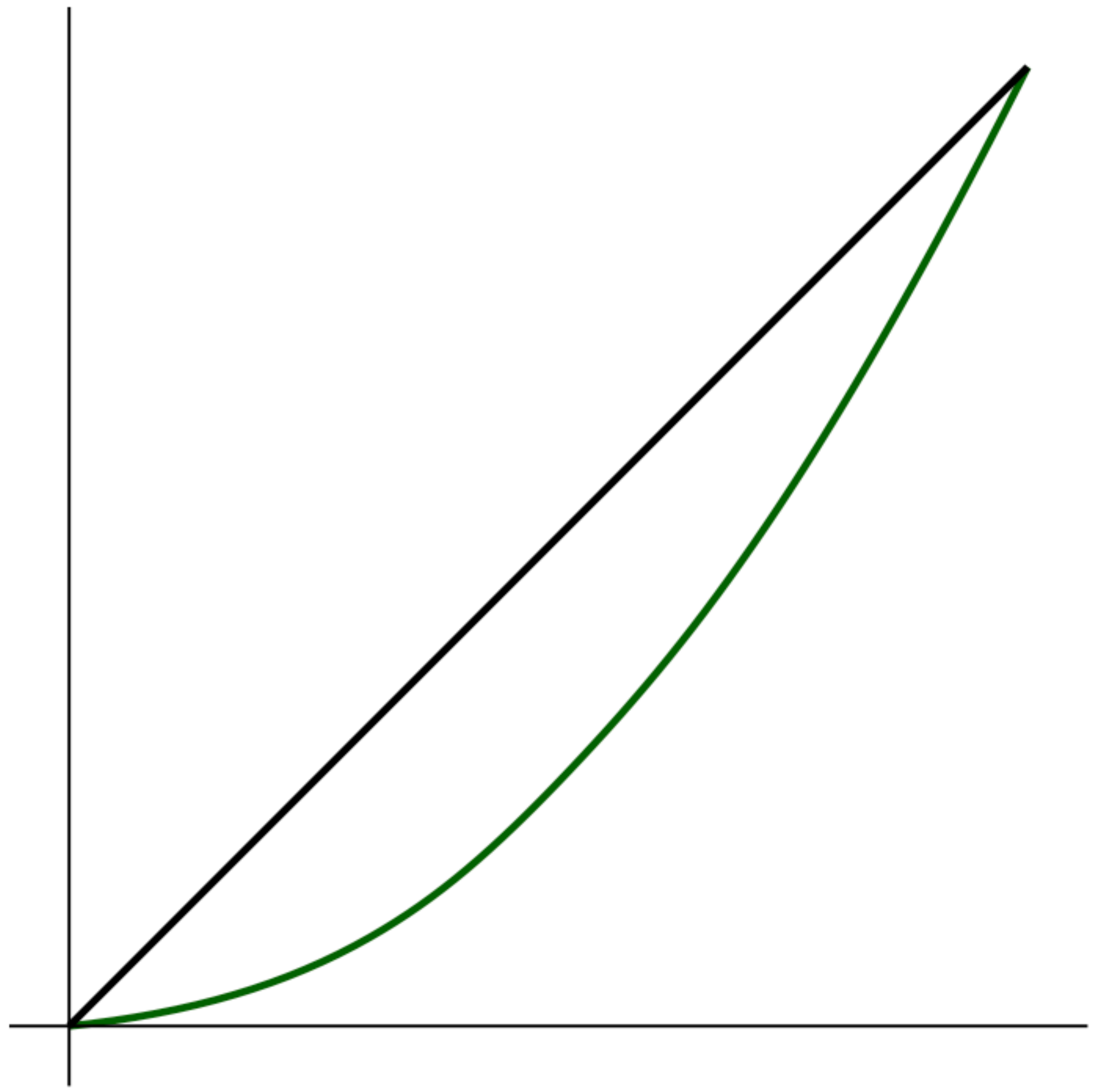
$$x^2 = \int_0^\infty 2 \cdot \sigma(x - b)\, \mathrm{d}b$$

- Recalling the sampling bounds, natural to assume a uniform distribution of neurons
    - Approximate this by piecewise linear approximation, with uniform interval
    - Let $\{h_i\}_{i=1}^n$ be a sequence of piecewise linear approximations on uniform intervals
        - Here, $h_i(x)$ interpolates the partition into 2^i intervals:

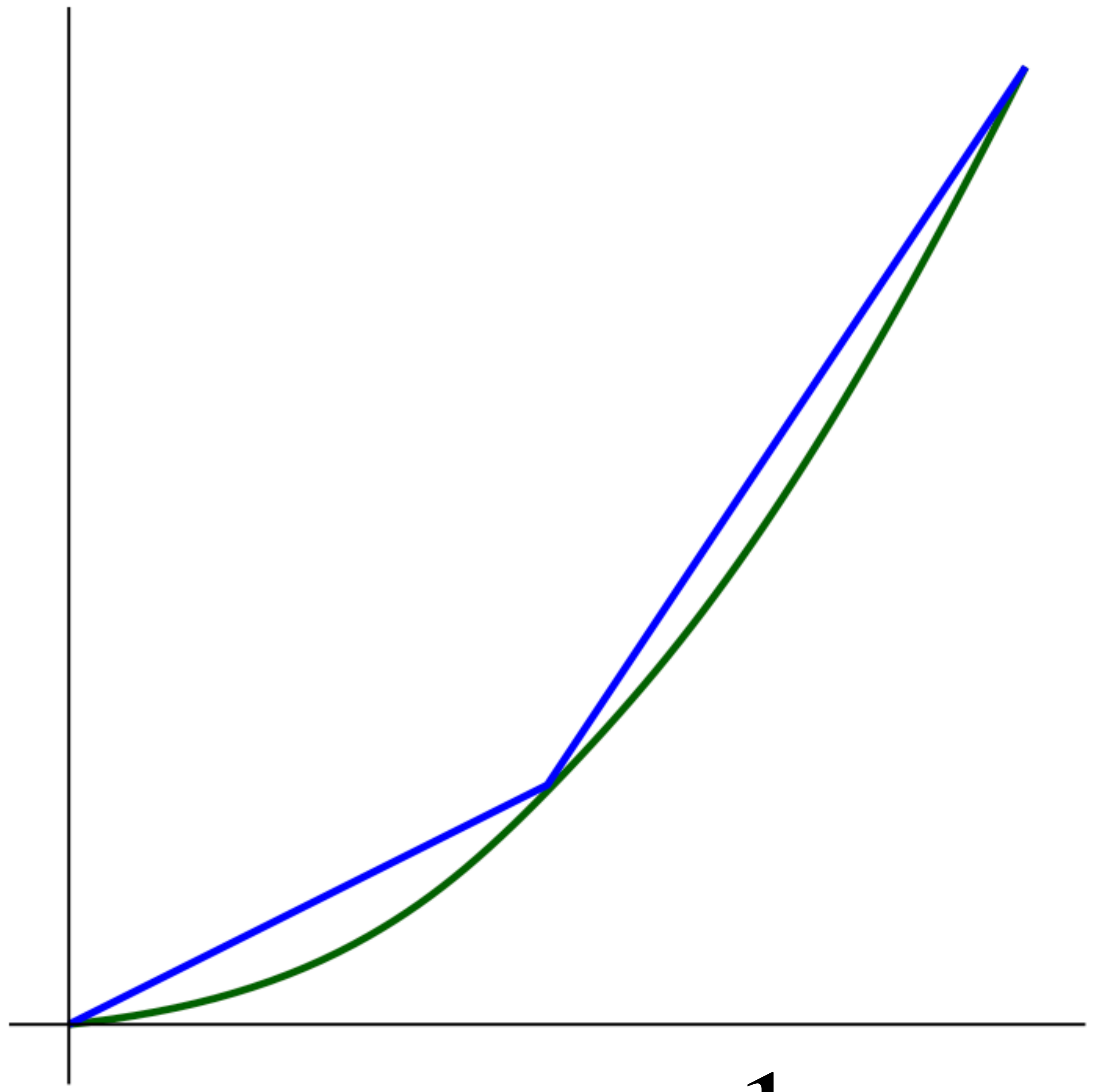
$$S_i := \left(0, \frac{1}{2^i}, \frac{2}{2^i}, \ldots, \frac{2^i}{2^i}\right)$$

Δ-approximators

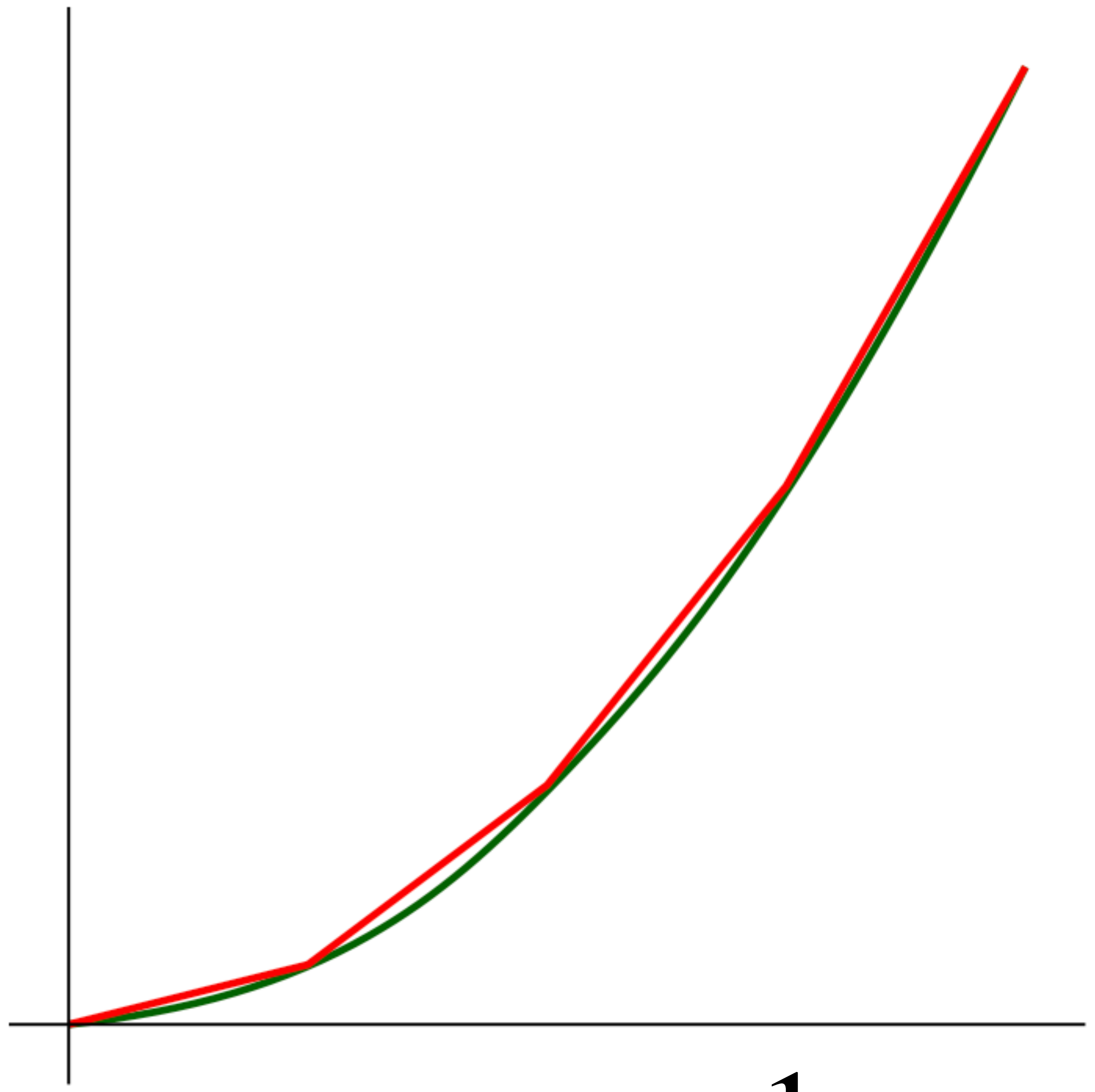
- **Observation.** The approximators h_i can be expressed in terms of wedges



$$h_0 = x$$



$$h_1 = h_0 - \frac{1}{4}\Delta$$



$$h_2 = h_1 - \frac{1}{4^2}\Delta^2$$

Δ-approximators

- **Observation.** The approximators h_i can be expressed in terms of wedges

    - More formally, for any $x \in S_i \setminus S_{i-1}$, we have, for $\epsilon = 1/2^i$:

$$h_i(x) - h_{i-1}(x) = x^2 - \frac{(x-\epsilon)^2 + (x+\epsilon)^2}{2}$$

$$= x^2 - \left(x^2 - \epsilon x + \epsilon x + \epsilon^2\right)$$

$$= -\frac{1}{4^i}$$

- This does not depend on "which x" (i.e., the same height)
    - Thus, making it Δ-like

$$h_i(x) = x - \sum_{j=1}^{i} \frac{\Delta^i}{4^i}$$

What we want

- We want to show three claims

  - **Claim 1.** Deep nets can construct $h_i(x)$ efficiently

  - **Claim 2.** $h_i(x)$ approximates x^2 well

  - **Claim 3.** Shallow nets cannot approximate x^2 well

What we want

- **Claim 1.** Deep nets can construct $h_i(x)$ efficiently

$$h_i(x) = x - \frac{1}{4}\Delta^1 - \frac{1}{4^2}\Delta^2 - \cdots - \frac{\Delta^i}{4^i}$$

  - Can be constructed as many parallel nets
    - Roughly $2i$ layers
    - Roughly $4i$ neurons

What we want

- **Claim 2.** $h_i(x)$ approximates x^2 well
    - More concretely, we claim that

$$\sup_{x\in[0,1]} |h_i(x) - x^2| \leq 4^{-i-1}$$

What we want

- **Claim 2.** $h_i(x)$ approximates x^2 well
    - More concretely, we claim that

$$\sup_{x \in [0,1]} |h_i(x) - x^2| \leq 4^{-i-1}$$

- **Proof idea.**
    - Fix some $x \in [0,1]$
    - We know that $x \in [j\tau, (j+1)\tau]$ for some j, where $\tau = 1/2^i$
    - Then, we can write:

$$h_i(x) = (j\tau)^2 + \frac{((j+1)\tau)^2 - (j\tau)^2}{\tau} \cdot (x - j\tau)$$

What we want

- **Claim 3.** Shallow nets cannot approximate x^2 well
  - More concretely, we claim that:
    Any ReLU net with $\leq L$ layers and $\leq N$ nodes satisfy

$$\int_0^1 (g(x) - x^2)^2 \, \mathrm{d}x \geq \frac{1}{5760 \cdot (2N/L)^{4L}}$$

What we want

- **Claim 3.** Shallow nets cannot approximate x^2 well
  - More concretely, we claim that:
    Any ReLU net with $\leq L$ layers and $\leq N$ nodes satisfy

$$\int_0^1 (g(x) - x^2)^2 \, \mathrm{d}x \geq \frac{1}{5760 \cdot (2N/L)^{4L}}$$

- **Proof idea.** Use the fact that

$$\min_{(c,d)} \int_a^b (x^2 - (cx + d))^2 \, \mathrm{d}x = \frac{(b-a)^5}{180}$$

Next up

- Near-initialization approximation and kernel regime