3. Approximation: Primer & Toy Case

Recap

- Recall the excess risk decomposition

$$R(\hat{f}) - R(f_{\mathrm{GT}})$$
$$\leq \left[R(\hat{f}) - R_n(\hat{f})\right] + \left[R_n(\hat{f}) - R_n(f_{\mathrm{ERM}})\right] + \left[R_n(f^*) - R(f^*)\right] + \boxed{\left[R(f^*) - R(f_{\mathrm{GT}})\right]}$$

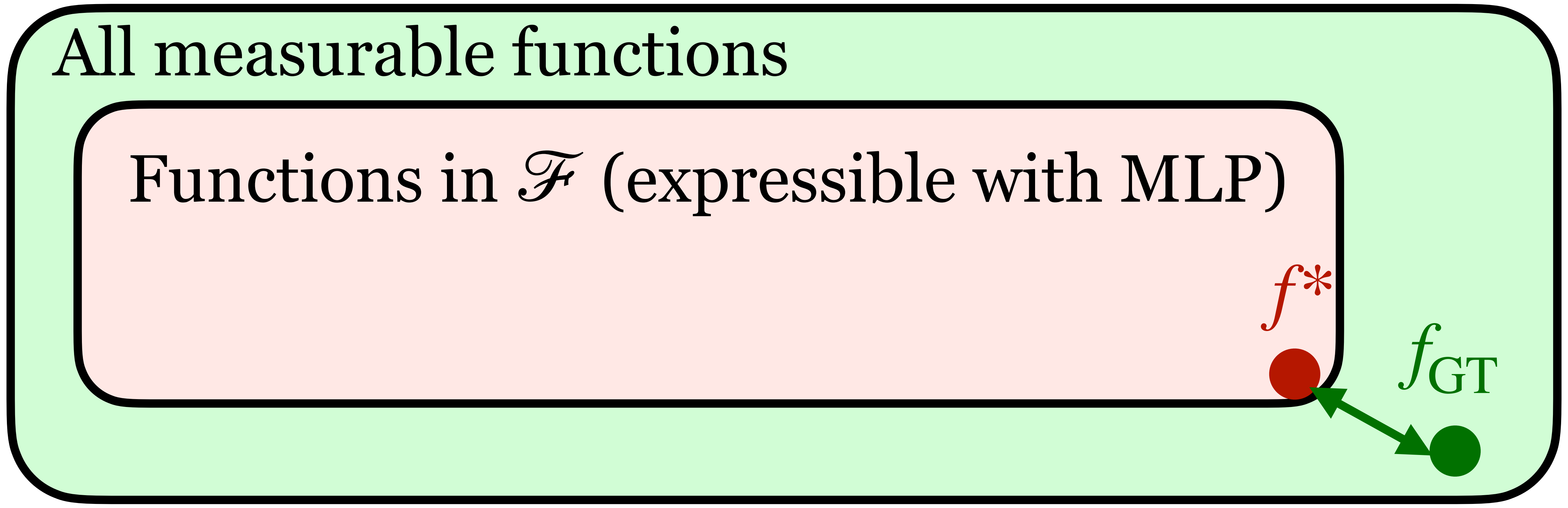
- **Approximation theory** is concerned with controlling the 4th term

$$R(f^*) - R(f_{\mathrm{GT}}) \quad = \quad \inf_{f \in \mathcal{F}} R(f) - \inf_{f \text{ meas.}} R(f)$$

Recap

$$\inf_{f \in \mathcal{F}} R(f) - \inf_{f \text{ meas.}} R(f)$$

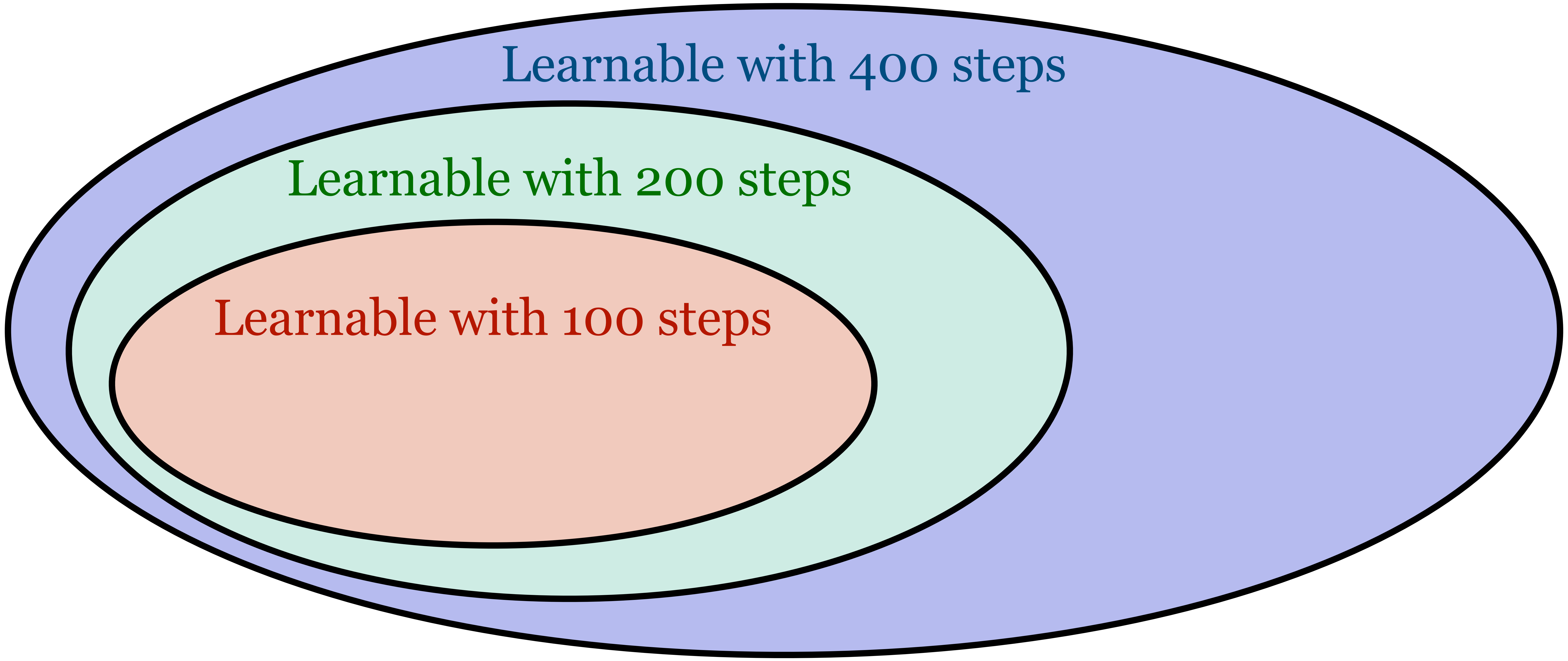
- This quantity measures the richness of the hypothesis space $\mathcal{F}$
  - **If $\mathcal{F}$ is rich.** The gap should be small
  - **If $\mathcal{F}$ is small.** The gap should be large



Recap

$$\inf_{f \in \mathcal{F}} R(f) - \inf_{f \text{ meas.}} R(f)$$

- Fortunately…
  - no {randomness, data} involved
  - less gradient descent involved
    - “less,” because running GD longer means larger $\mathcal{F}$



Quantity of interest

- Still, this quantity *per se* is difficult to analyze

$$\inf_{f \in \mathcal{F}} R(f) - \inf_{f \text{ meas.}} R(f)$$

  - So let us simplify further

- **Issue 1.** Terms are still about P, which we never know

$$\inf_{f \in \mathcal{F}} \mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)] - \inf_{f \text{ meas.}} \mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)]$$

Quantity of interest

- **Simplification 1.** Express it as the "distance of hypotheses"

$$\inf_{f\in\mathcal{F}} R(f) - \inf_{f\text{ meas.}} R(f) = \sup_{g\text{ meas.}} \inf_{f\in\mathcal{F}} \left(\mathbb{E}[\ell(f(X),Y) - \ell(g(X),Y)]\right)$$

$$= \sup_{g\text{ meas.}} \inf_{f\in\mathcal{F}} \left(\int_{\mathcal{X}\times\mathcal{Y}} P_{XY}(x,y)\cdot\left(\ell(f(X),Y) - \ell(g(X),Y)\right) \mathrm{d}x\mathrm{d}y\right)$$

Definition of Lipschitz constant

$$\leq \mathrm{Lip}_{(1)}(\ell)\cdot \sup_{g\text{ meas.}} \inf_{f\in\mathcal{F}} \left(\int_{\mathcal{X}\times\mathcal{Y}} P_{XY}(x,y)\cdot |f(X) - g(X)|\, \mathrm{d}x\mathrm{d}y\right)$$

Hölder's inequality

$$\leq \mathrm{Lip}_{(1)}(\ell)\cdot \|P_{XY}(x,y)\|_q \cdot \sup_{g\text{ meas.}} \inf_{f\in\mathcal{F}} \left(\|f(x) - g(x)\|_p\right)$$

where p and q are Hölder conjugates (i.e., $1/p + 1/q = 1$)

Quantity of interest

- **Simplification 1.** Express it as the "distance of hypotheses"

$$\mathrm{Lip}(\ell) \cdot \|P_{XY}(x, y)\|_q \cdot \sup_{g \text{ meas.}} \inf_{f \in \mathcal{F}} \left(\|f(x) - g(x)\|_p \right)$$

  - A popular choice is to let $p = \infty$
    - Then, $q = 1$ and we get the supremum norm bound:

$$\mathrm{Lip}(\ell) \cdot \sup_{g \text{ meas.}} \inf_{f \in \mathcal{F}} \|f(x) - g(x)\|_\infty$$

  - Otherwise, we can use general p

Quantity of interest

- **Simplification 1.** Express it as the "distance of hypotheses"

$$\mathrm{Lip}(\ell) \cdot \sup_{g \text{ meas.}} \inf_{f \in \mathcal{F}} \|f(x) - g(x)\|_\infty$$

  - Also, in general, we'll ignore the Lipschitz constant
    - because this is something that we cannot control

  - If you are irritated by the fact that $\mathrm{Lip}(\ell)$ does not exist even for ℓ^2 loss
    - simply assume the bounded support $\mathcal{X}, \mathcal{Y}$

Assumptions make you happy ;)

Quantity of interest

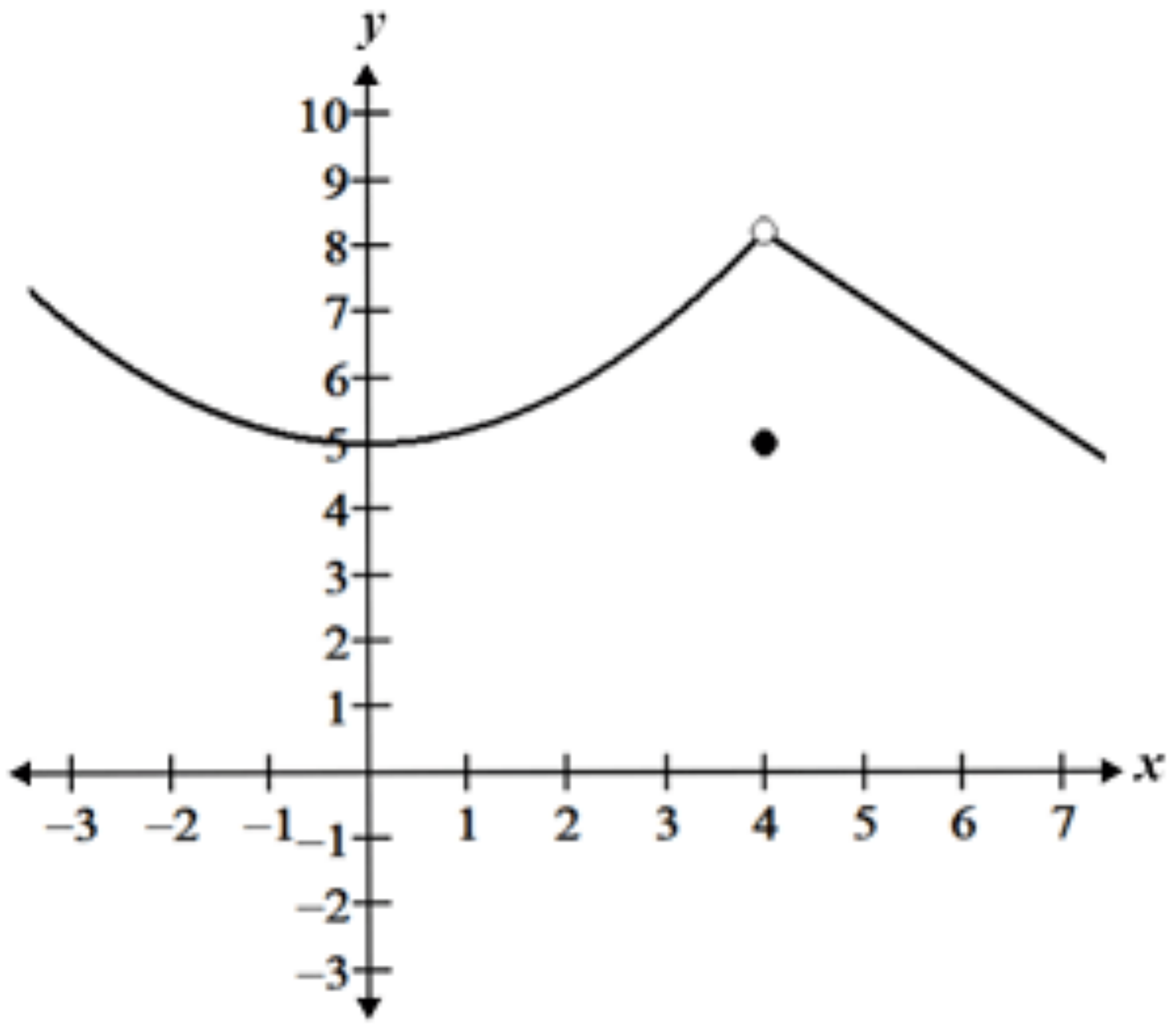
- This is what we have now:

$$\sup_{g \text{ meas.}} \inf_{f \in \mathcal{F}} \|f(x) - g(x)\|_\infty$$

- **Issue 2.** Taking care of "worst measurable g" is too pessimistic

$$\sup_{g \text{ meas.}} \inf_{f \in \mathcal{F}} \|f(x) - g(x)\|_\infty$$

  - Discontinuities can make your function arbitrarily wrong



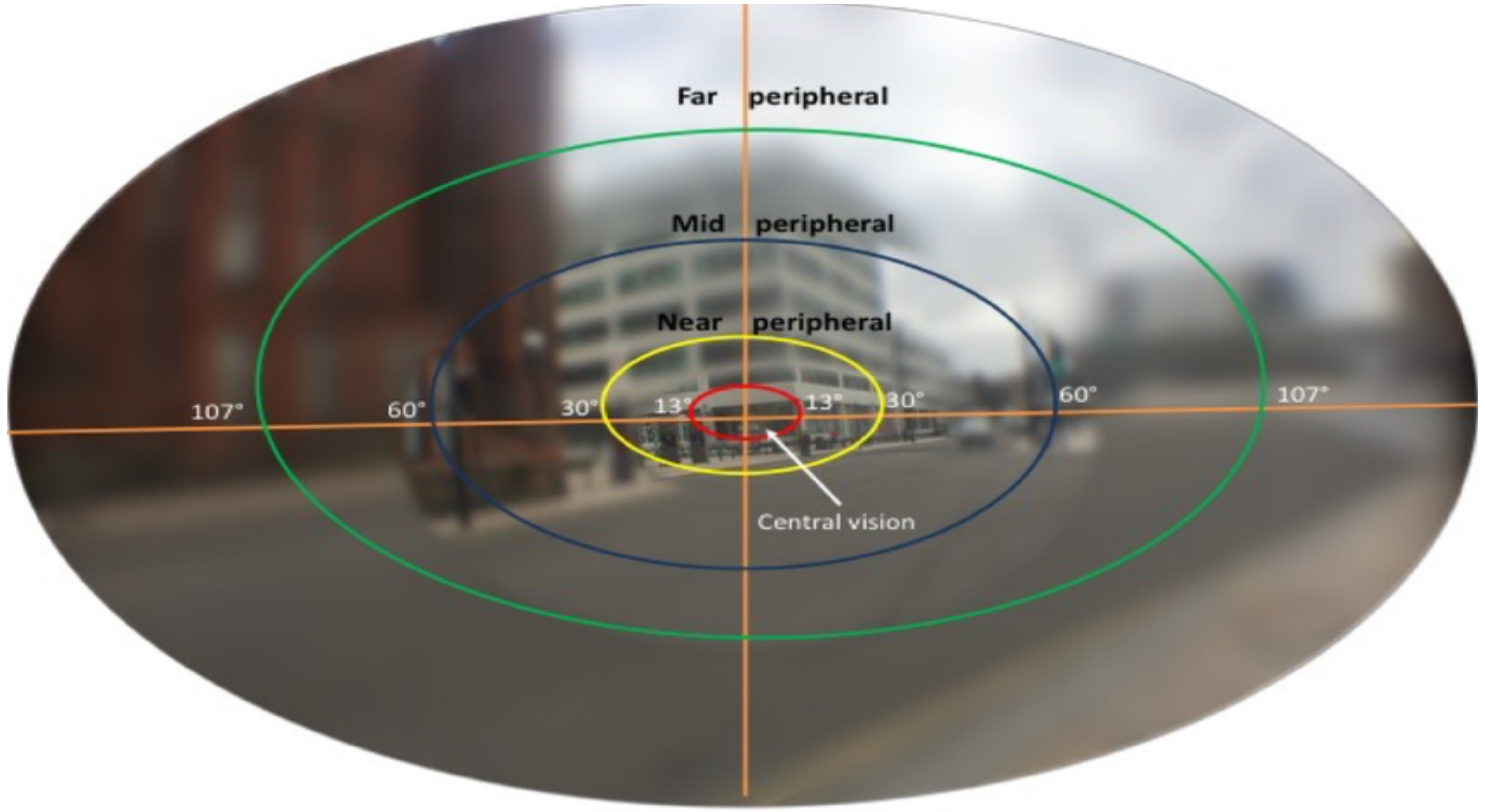
Quantity of interest

- **Simplification 2.** Again, we'll narrow down the to continuous target functions

$$\sup_{g\text{ cont.}} \inf_{f \in \mathcal{F}} \|f(x) - g(x)\|_\infty$$

  - Justification. Ground truth is rarely discontinuous
    - e.g., is human prediction altered by infinitesimal perturbation on input?

$$f(x) \to f(x + \varepsilon)$$



Quantity of interest

- Summing up, this is the quantity that we want to upper/lower-bound for the next few weeks

$$\sup_{g \text{ cont.}} \inf_{f \in \mathcal{F}} \|f(x) - g(x)\|_\infty$$

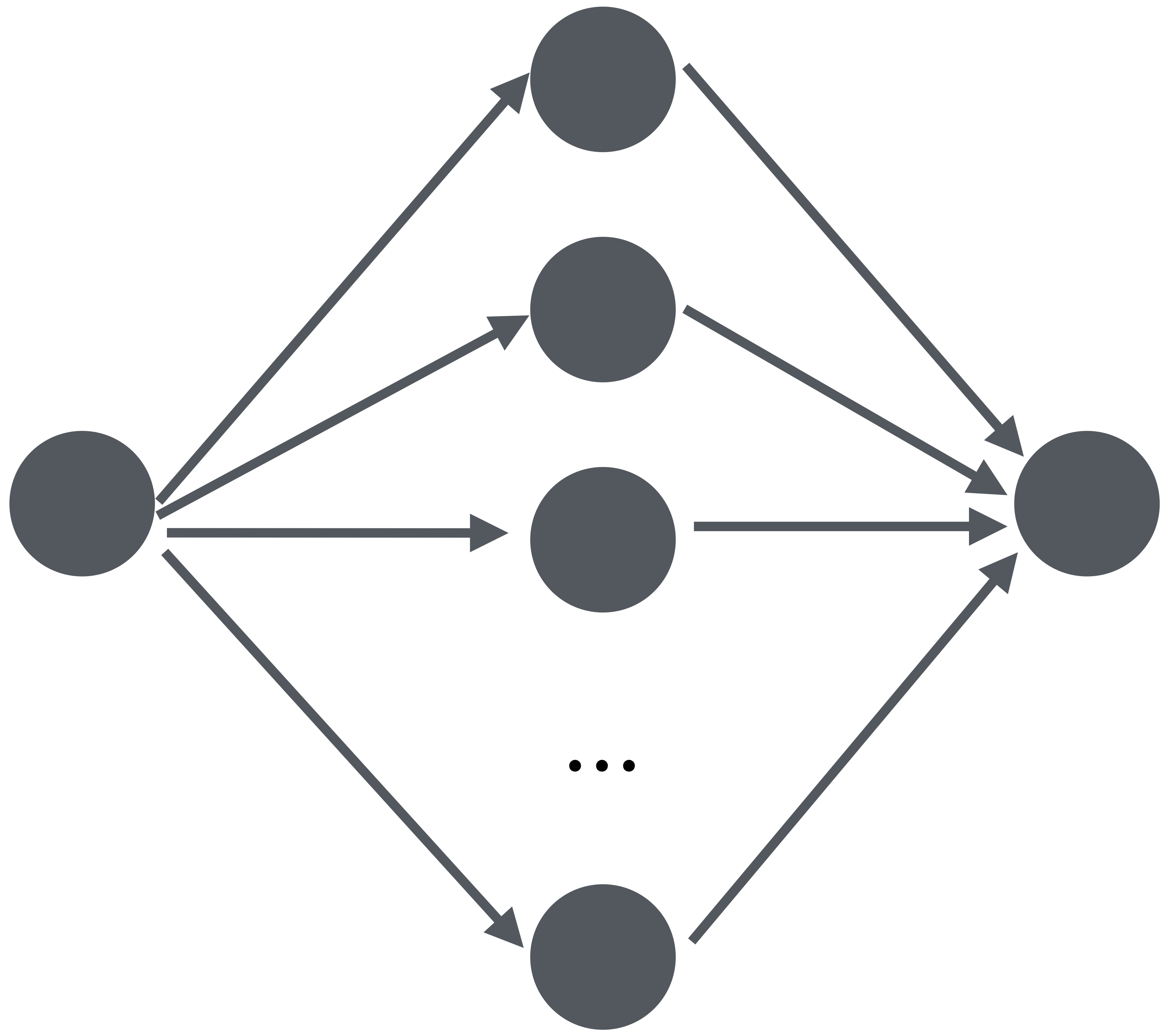
  - called universal approximation results
  - very actively studied in 1980s and 1990s

  - Modern variants include:
    - Are GNNs universal approximators?
    - Are sparse-attention transformers universal approximators?
    - Are mamba-like models universal approximators?
    - Are equivariant networks universal approximators?

The simplest universal approximation theorem

Setup

- To give you an idea, we first study a very simple case
  - 1D inputs
    - $x \in \mathbb{R}, y \in \mathbb{R}$
  - Bounded input domain
    - $x \in [0,1]$
  - Two-layer networks
    - Threshold activation $\sigma(x) = \mathbf{1}\{x \geq 0\}$



- The hypothesis space can be written as:

$$\mathcal{F} = \left\{ \sum_{i=1}^{m} a_i \mathbf{1}\{w_i x + b_i\} \;\middle|\; a_i \in \mathbb{R}, w_i \in \mathbb{R}, b_i \in \mathbb{R} \right\}$$

Result

Proposition 2.1.

Suppose that $g : \mathbb{R} \to \mathbb{R}$ is ρ-Lipschitz. Then, for any $\varepsilon > 0$, there exists a 2-layer network with $\lceil \rho/\varepsilon \rceil$ threshold nodes, so that

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

- Universal approximation is possible, if:
    - certain width and depth conditions are satisfied
    - certain smoothness assumption holds on GT

Proof

Proof.

- Idea: Think about what each neuron represents in threshold neural net ✏️

$$\mathcal{F} = \left\{ \sum_{i=1}^{m} a_i \mathbf{1}\{w_i x + b_i\} \;\middle|\; a_i \in \mathbb{R}, w_i \in \mathbb{R}, b_i \in \mathbb{R} \right\}$$

Proof

Proof.

- Idea: Construct a “histogram”-like approximation of the original function ✏️

Discussion

- While the result is very simple, it contains all the core ideas
  - We broke down GT into basis + small error
  - We used each neuron to approximate the basis
    - Thankfully, this step was exact

- Notice that we have used “Lipschitz assumption” on the GT — a worst-case bound on smoothness
  - **Brainteaser.** If we have a more refined bound, such as total variation, then can we prove a better bound?

Next up

- In the coming lectures, we extend this idea to more complicated cases
  - Two-layer —> Deeper models
  - Threshold —> ReLU and Sigmoid
  - Uniform norm —> L_p norm