Introduction to Machine Learning

13. Learning Theory

Geoff Gordon and Alex Smola
Carnegie Mellon University

http://alex.smola.org/teaching/cmu2013-10-701x
10-701

Carnegie Mellon University

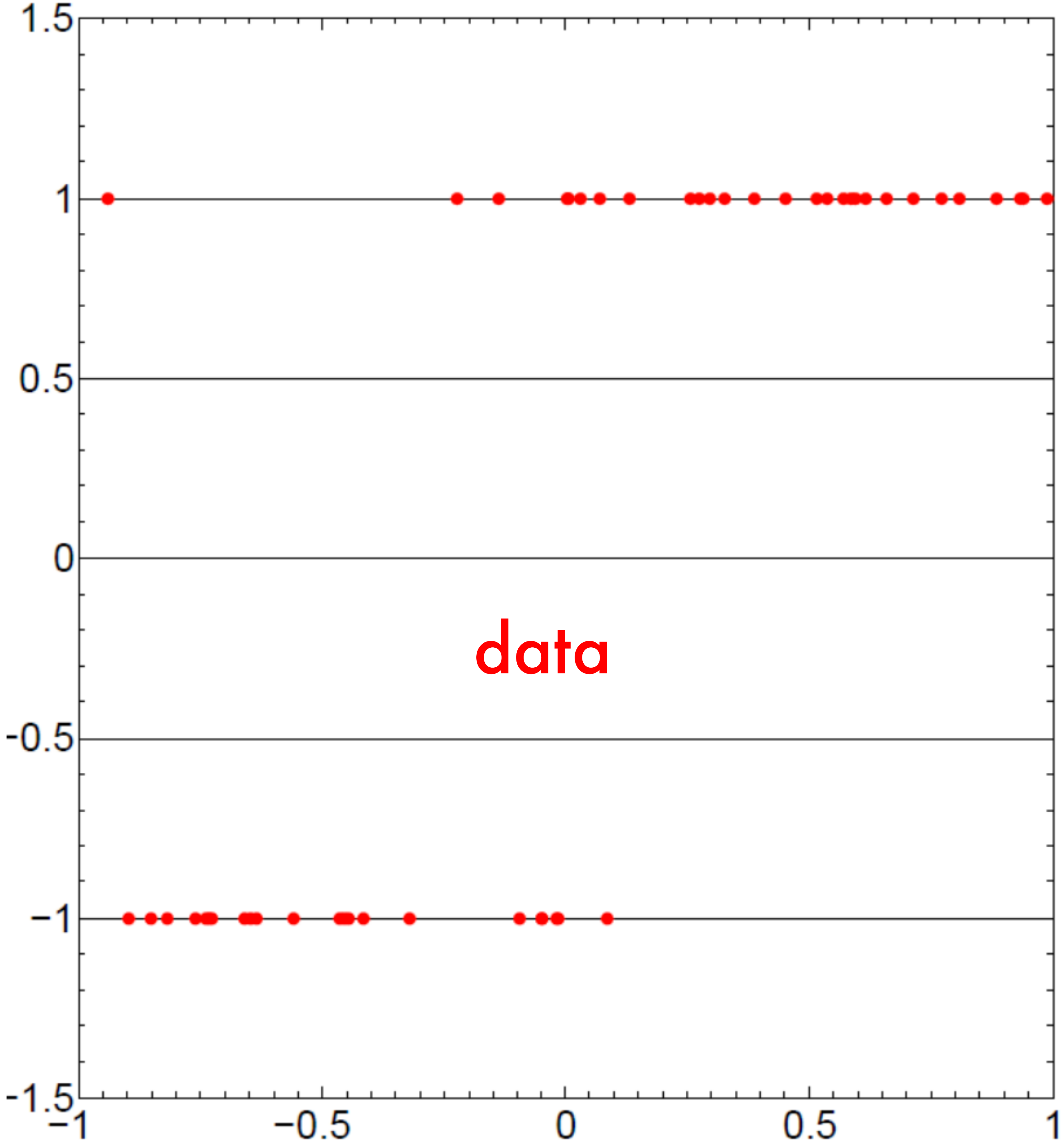
The Problem

- Training
  - Data $\{(x_1, y_1), \ldots (x_m, y_m)\}$ drawn iid from $p(x, y)$
  - Loss function $l(x, y, f(x))$
  - Function class $\mathcal{F} = \{f : \Omega[f] \leq c\}$
  - Empirical risk minimization problem

$$\underset{f \in \mathcal{F}}{\text{minimize}} \frac{1}{m} \sum_{i=1}^{m} l(x_i, y_i, f(x_i))$$

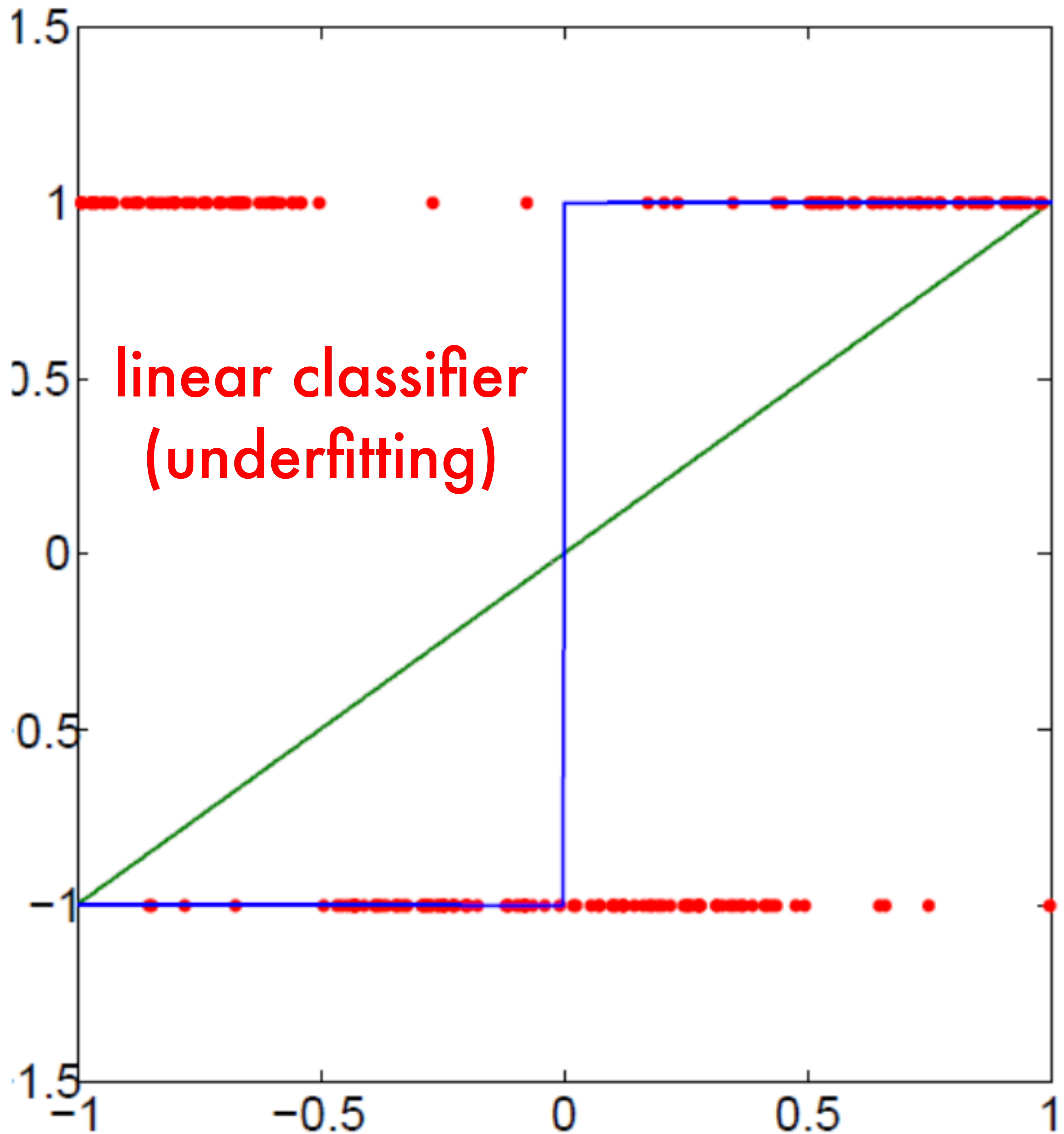
- Testing

$$\mathop{\mathbf{E}}_{(x,y) \sim p(x,y)} [l(x, y, f(x))]$$

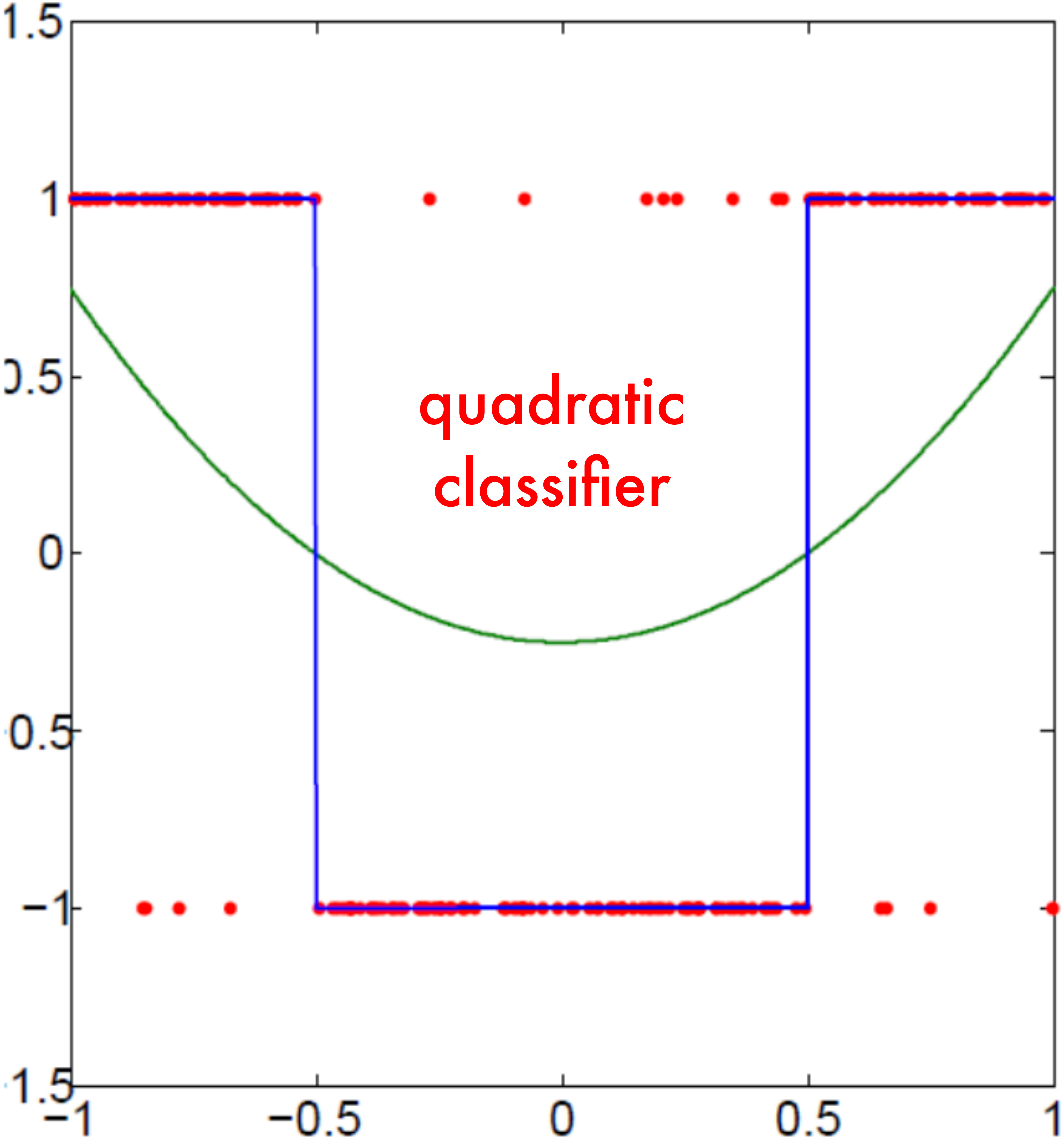


Picture from David Pal

Picture from David Pal

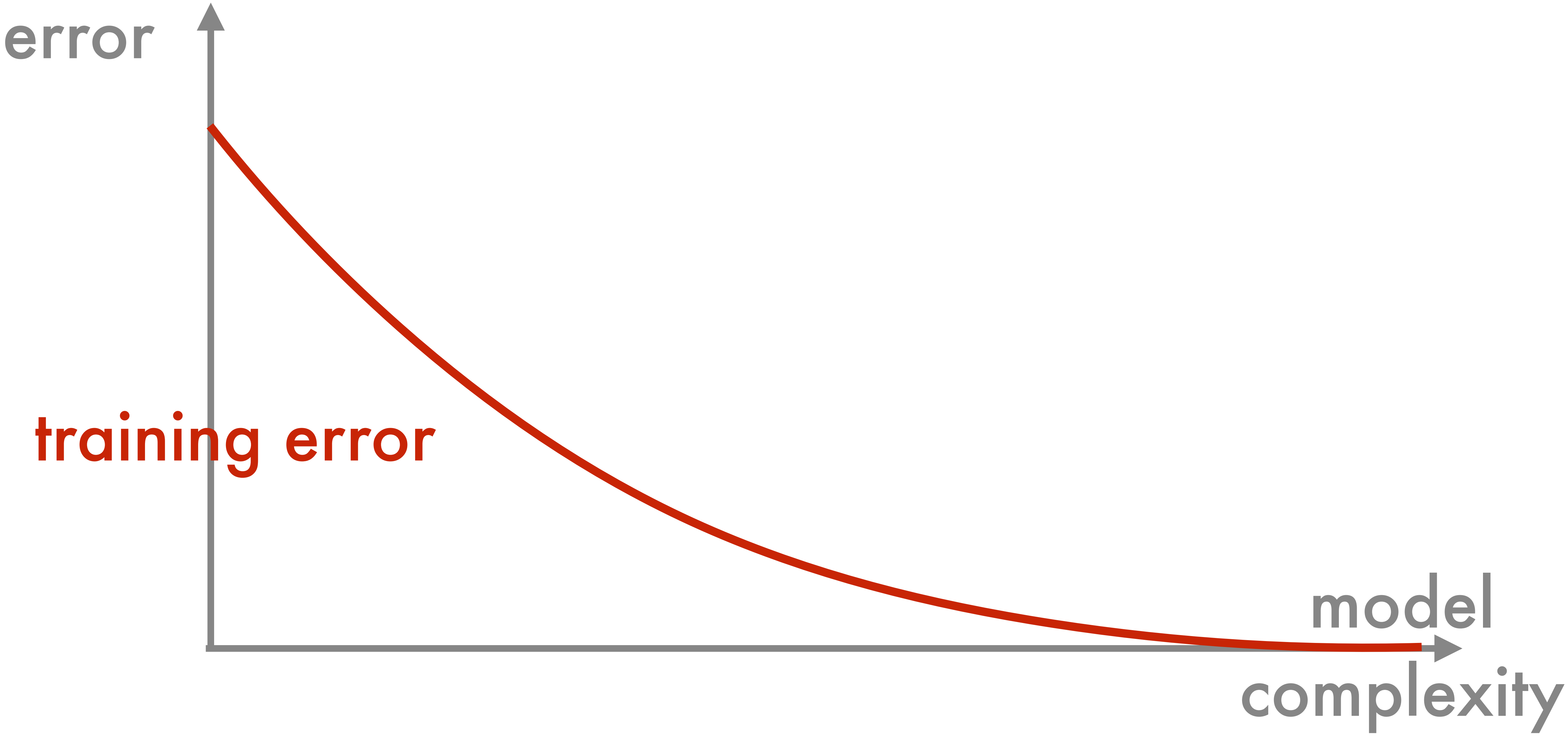


Mellon University


quadratic
classifier
1.5
1
0.5
0
-0.5
-1
-1.5
-1
-0.5
0
0.5
1

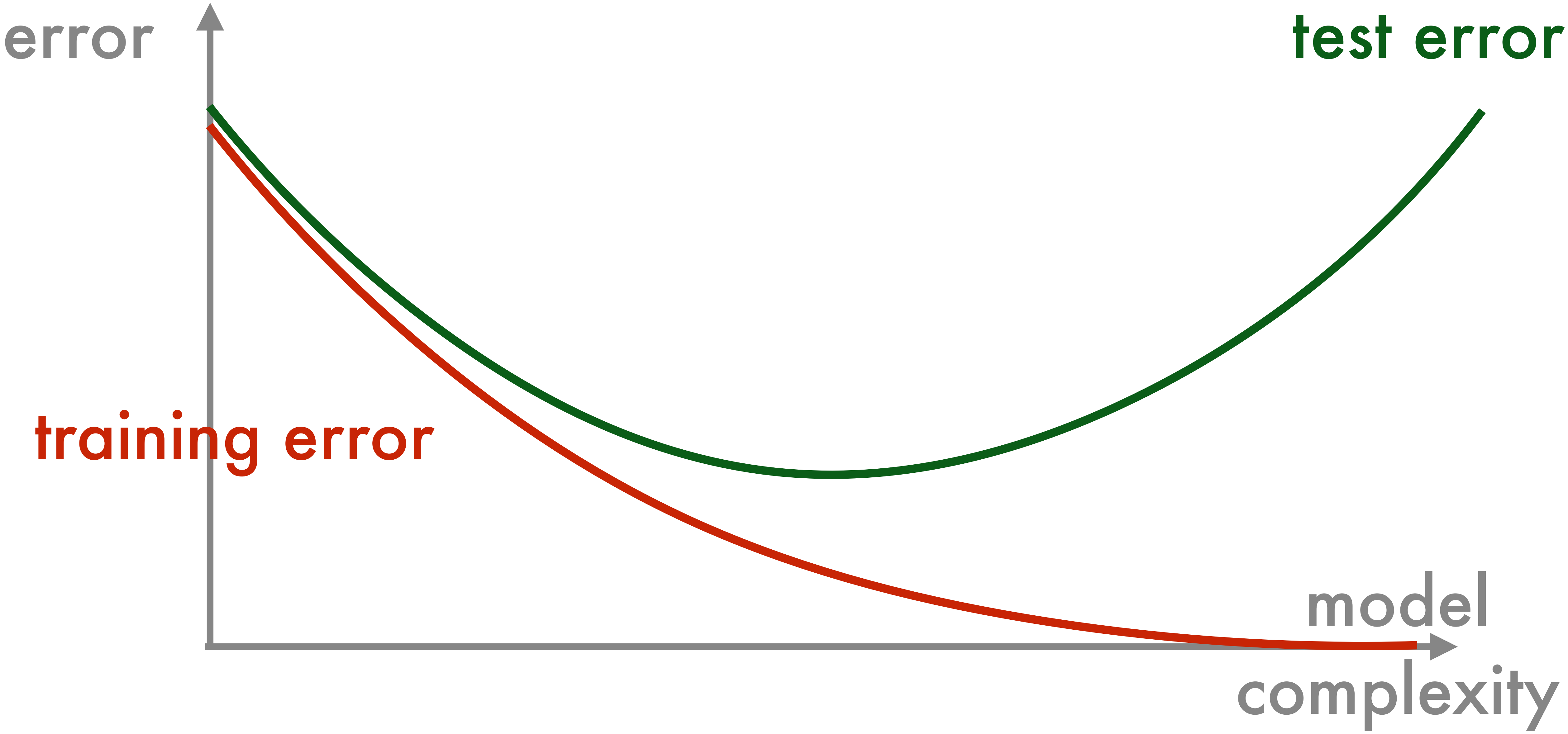
Typical behavior

Typical behavior

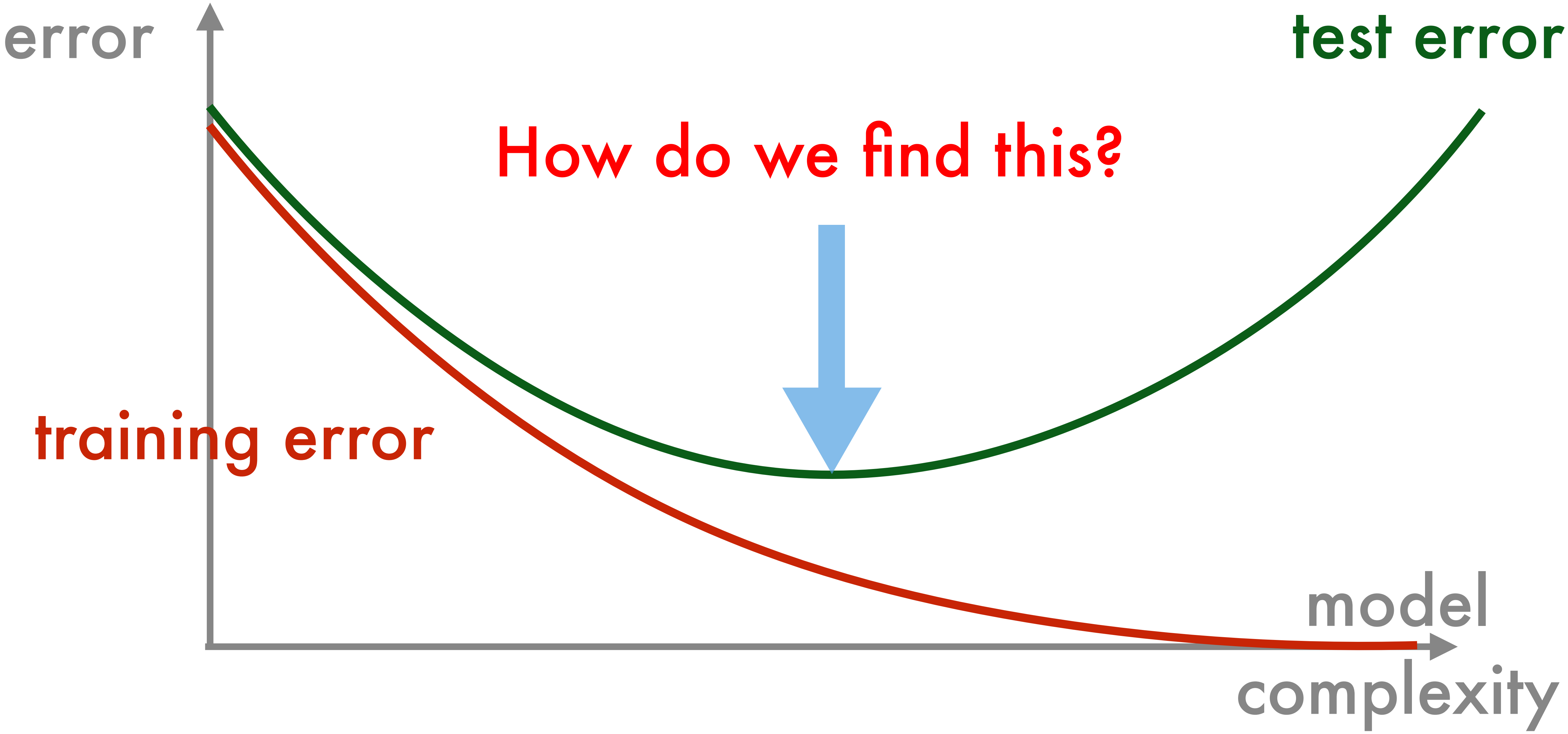


Carnegie Mellon University

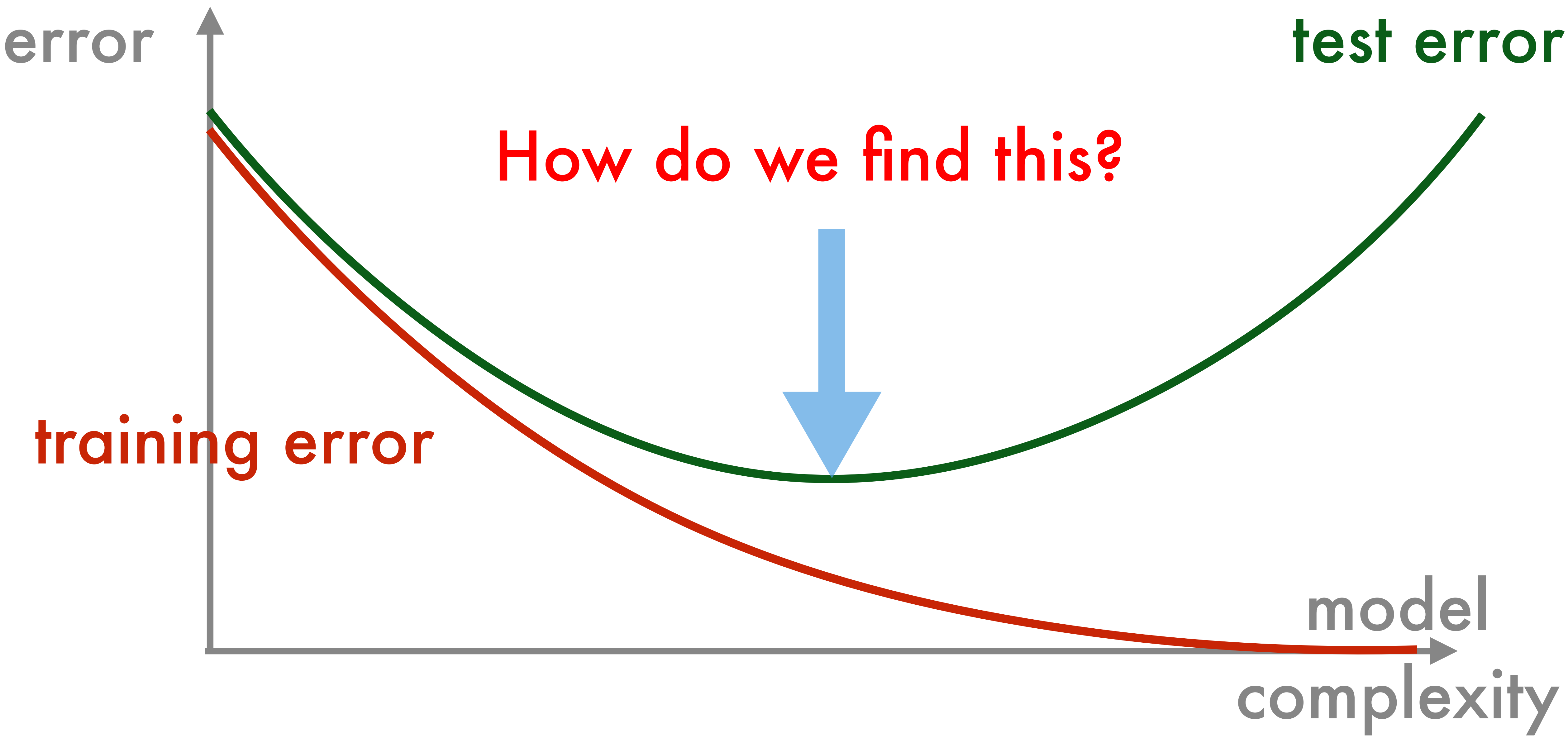
Typical behavior



Typical behavior



Typical behavior



A broken reasoning

- Hoeffding bound for bounded random variable

$$\Pr\left(|\hat{\mu}_m - \mu| > \epsilon\right) \leq 2\exp\left(-\frac{2m\epsilon^2}{c^2}\right).$$

- Function f^* that minimizes empirical risk
- Bounded risk by L
- Apply bound to get with high probability

$$\epsilon \leq L\sqrt{(\log 2/\delta)/2m}$$

A broken reasoning

- Hoeffding bound for bounded random variable

$$\Pr\left(|\hat{\mu}_m - \mu| > \epsilon\right) \leq 2\exp\left(-\frac{2m\epsilon^2}{c^2}\right).$$

- Function f^* that minimizes empirical risk
- Bounded risk by L
- Apply bound to get with high probability

$$\epsilon \leq L\sqrt{(\log 2/\delta)/2m}$$

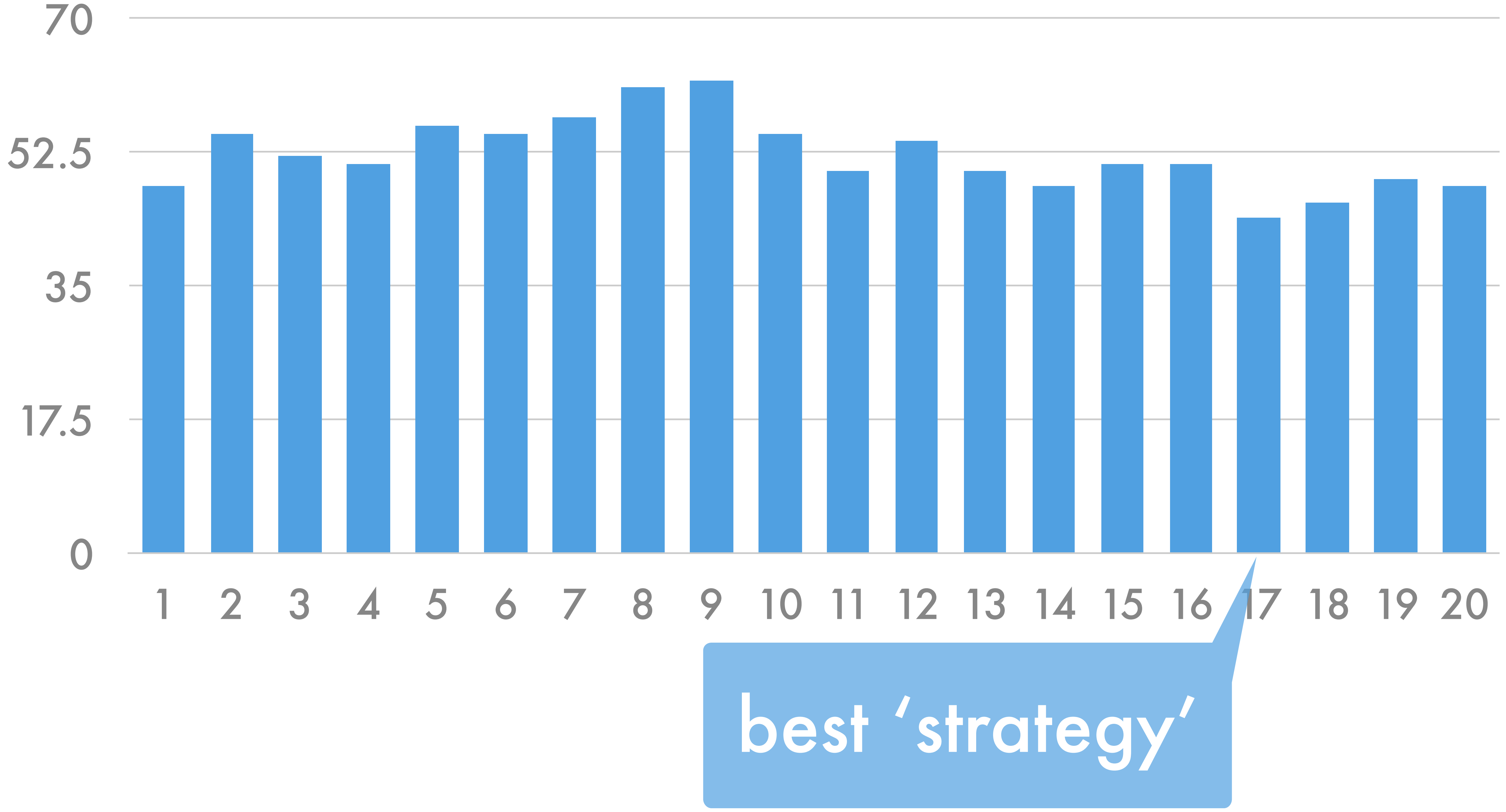
- Why does our bound diverge in reality?

Multiple testing

- Tossing an unbiased coin 10 times

Multiple testing

- Tossing an unbiased coin 100 times



Multiple testing

- We invoke the bound each time we test
- Picking the best out of N gives us N opportunities to get it wrong!
- Union bound

$$\Pr\left\{|R_{\mathrm{emp}}[f] - R[f]| > \epsilon\right\} \leq \sum_{f' \in \mathcal{F}} \Pr\left\{|R_{\mathrm{emp}}[f'] - R[f']| > \epsilon\right\}$$

- Testing over all functions in function class
  - Split error probability up among all functions
  - Take supremum over all terms

Multiple testing

- Our first generalization bound

$$\epsilon \leq L\sqrt{\frac{\log|\mathcal{F}| + \log 2/\delta}{2m}}$$

- Putting it all together

$$R[f^*] \leq \inf_{f \in \mathcal{F}} R_{\text{emp}}[f] + L\sqrt{\frac{\log|\mathcal{F}| + \log 2/\delta}{2m}}$$

Multiple testing

- Our first generalization bound

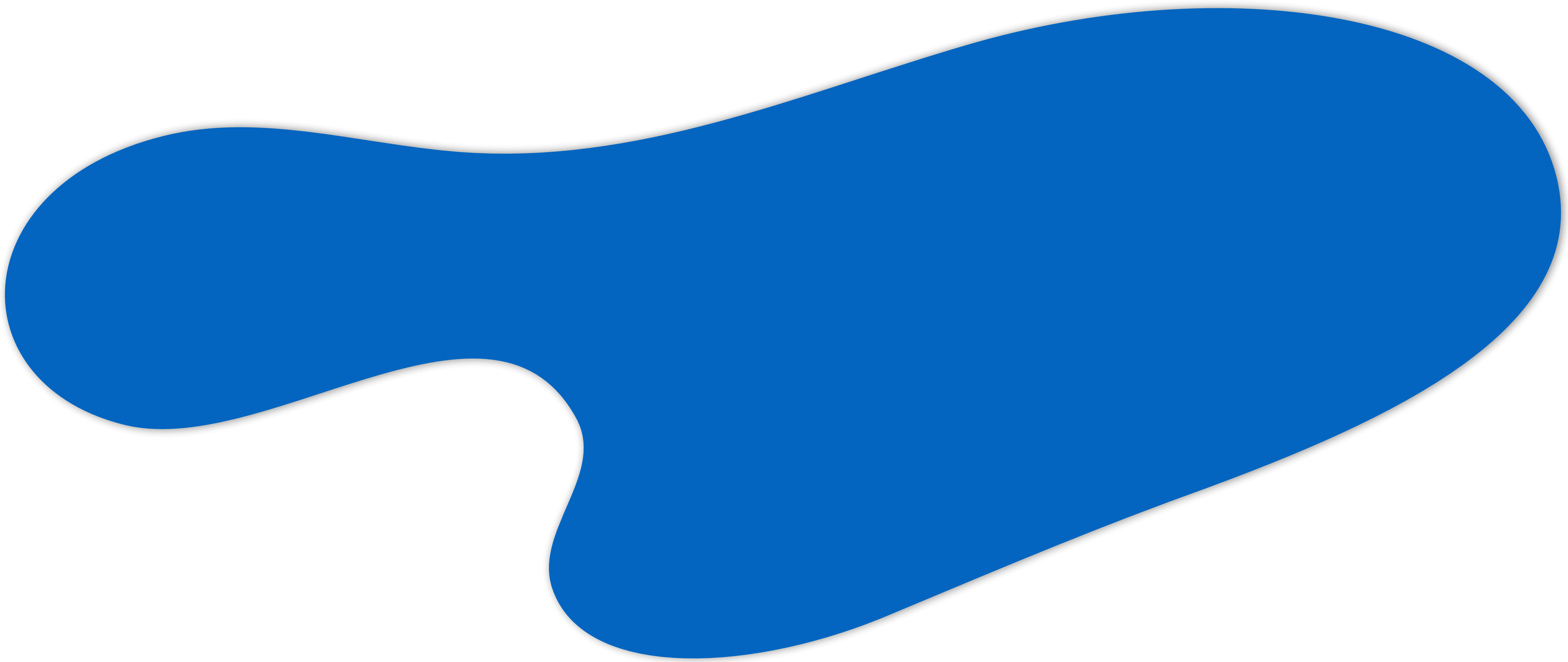
$$\epsilon \leq L\sqrt{\frac{\log|\mathcal{F}| + \log 2/\delta}{2m}}$$

- Putting it all together

$$R[f^*] \leq \inf_{f\in\mathcal{F}} R_{\mathrm{emp}}[f] + L\sqrt{\frac{\log|\mathcal{F}| + \log 2/\delta}{2m}}$$

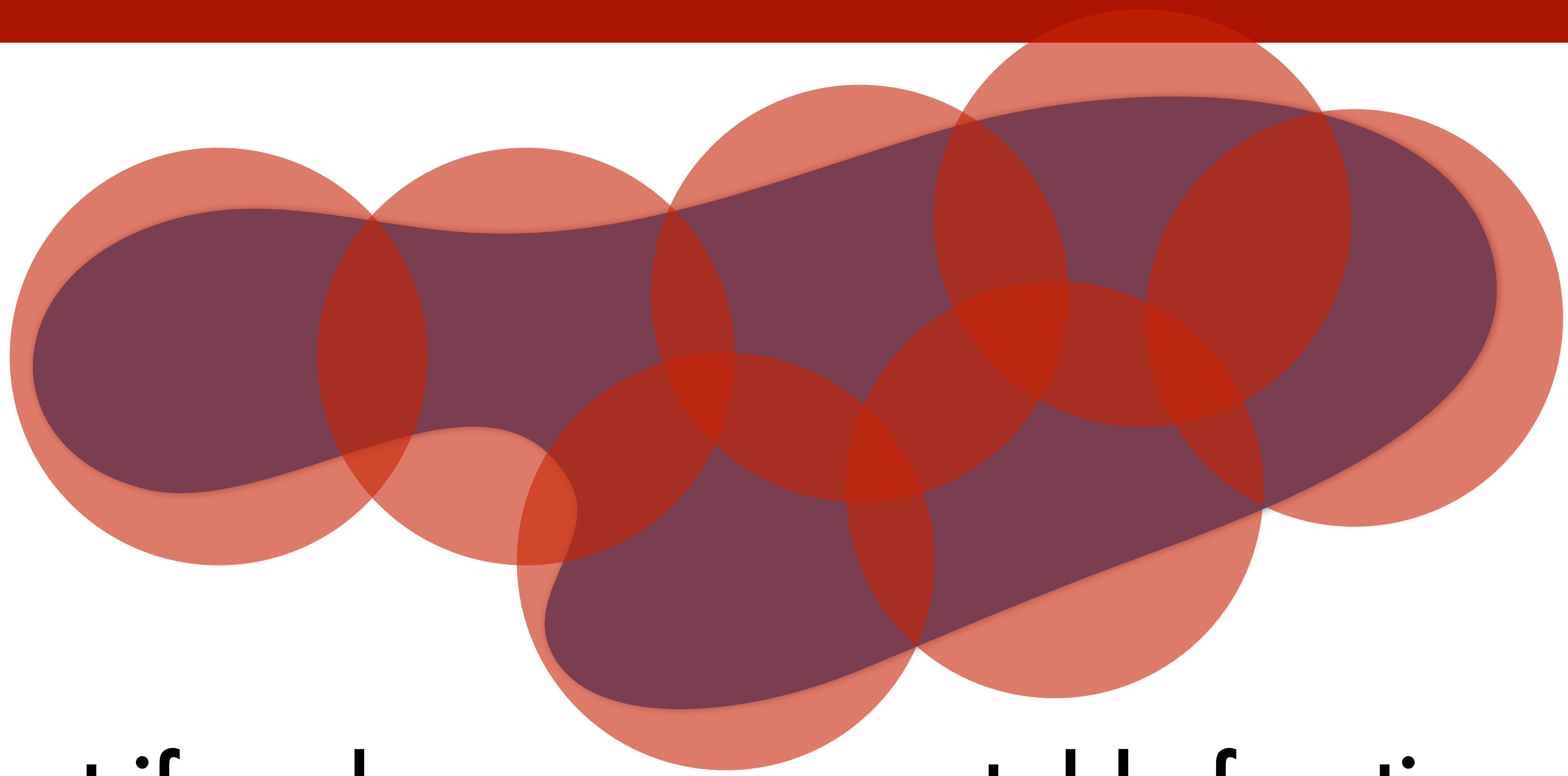
- What if function class is not discrete?
- What if we have binary loss

Covering Numbers



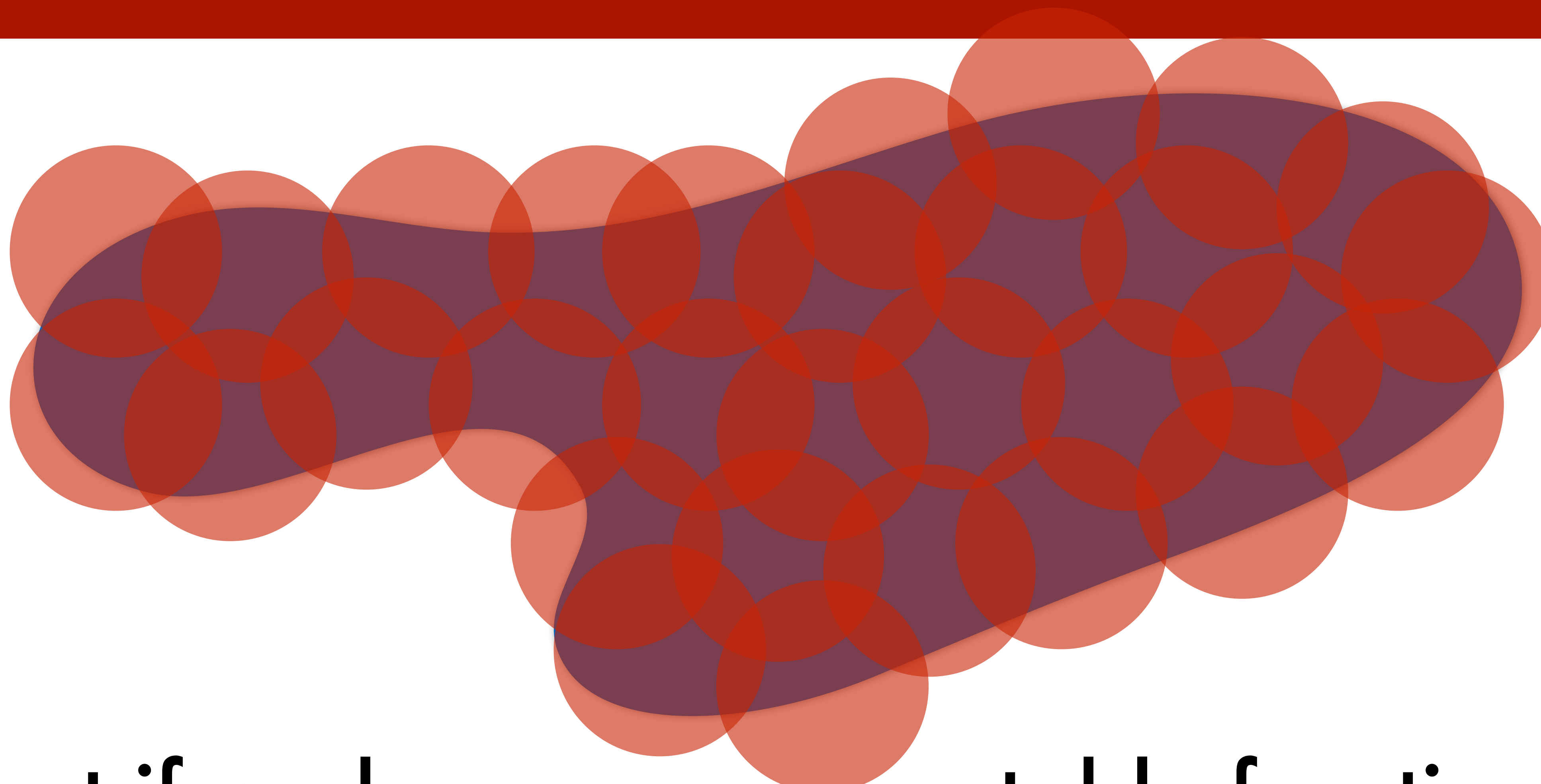
- What if we have an uncountable function class?
- Approximate by finite cover

Covering Numbers



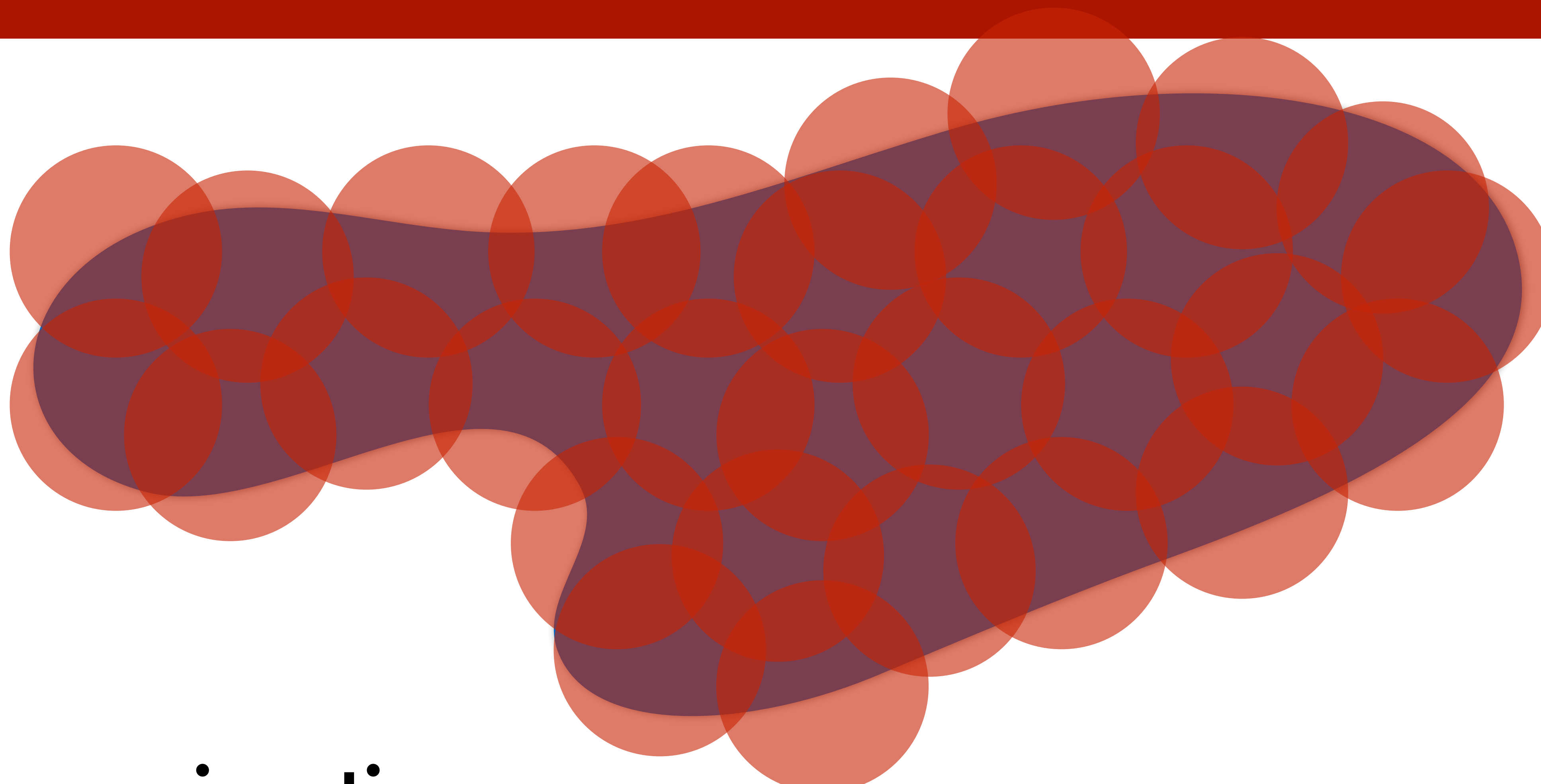
- What if we have an uncountable function class?
- Approximate by finite cover

Covering Numbers



- What if we have an uncountable function class?
- Approximate by finite cover
- Now bound depends on discretization, too

Covering Numbers

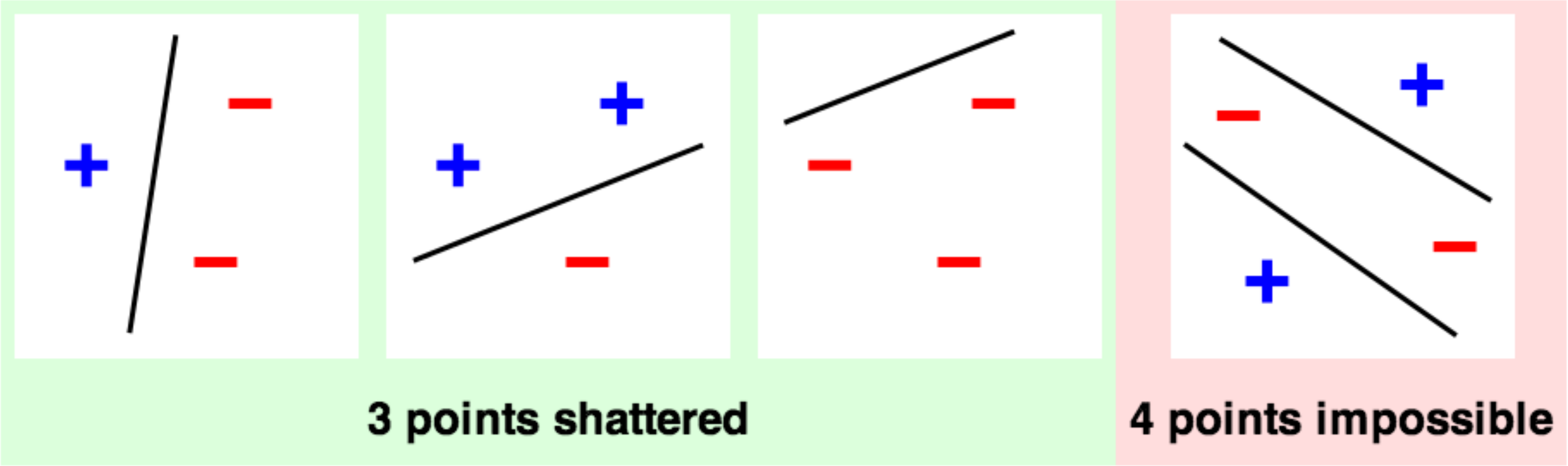


- Approximation error ϵ
- Covering number $N(\mathcal{F}, \epsilon)$ (actually need metric)

$$R[f^*] \leq \inf_{f \in \mathcal{F}} R_{\mathrm{emp}}[f] + L\sqrt{\frac{\log N(\mathcal{F}, \epsilon) + \log 2/\delta}{2m}} + L'\epsilon$$

VC Dimension

- Binary classification problem
- Given locations, enumerate all possible ways these points can be separated
- Example - linear separation



VC Dimension

- Binary classification problem
- Given locations, enumerate all possible ways these points can be separated
- Exponential growth to VCD, then polynomial

$$R[f^*] \leq \inf_{f \in \mathfrak{F}} R_{\mathrm{emp}}[f] + \sqrt{\frac{h(\log(2m/h) + 1) + \log 4/\delta}{m}}$$

- Examples
  - d-dimensional linear functions have h=d
  - $\sin(x/w)$ has infinite h

VC Dimension

- Binary classification problem
- Given locations, enumerate all possible ways these points can be separated
- Exponential growth to VCD, then polynomial

$$R[f^*] \leq \inf_{f \in \mathcal{F}} R_{\text{emp}}[f] + \sqrt{\frac{h(\log(2m/h) + 1) + \log 4/\delta}{m}}$$

polynomial growth

- Examples
  - d-dimensional linear functions have h=d
  - $\sin(x/w)$ has infinite h

Rademacher Averages

- Nontrivial bound (state of the art)
- Reasonably easy to compute
- Recall McDiarmid's inequality

$$\Pr\left(|f(x_1,\ldots,x_m) - \mathbf{E}_{X_1,\ldots,X_m}[f(x_1,\ldots,x_m)]| > \epsilon\right) \leq 2\exp\left(-2\epsilon^2 C^{-2}\right)$$

$$|f(x_1,\ldots,x_i,\ldots,x_m) - f(x_1,\ldots,x_i',\ldots,x_m)| \leq c_i$$

$$C^2 = \sum_{i=1}^m c_i^2$$

- Bound worst case deviation

$$\Pr\left\{\sup_{f\in\mathfrak{F}}\left|\frac{1}{m}\sum_{i=1}^m l(x_i,y_i,f(x_i)) - \mathbf{E}_{(x,y)}\left[l(x,y,f(x))\right]\right| > \epsilon\right\}$$

Rademacher Averages

- Worst case deviation

$$\Xi(X,Y) := \sup_{f\in\mathcal{F}} \left| \frac{1}{m}\sum_{i=1}^{m} l(x_i, y_i, f(x_i)) - \mathbf{E}_{(x,y)}\left[l(x,y,f(x))\right] \right|$$

- If we change single observation pair

$$\left|\Xi(X,Y) - \Xi(X^{-i}\cup\{x_i'\}, Y^{-i}\cup\{y_i'\})\right| \leq L/m$$

Rademacher Averages

- Worst case deviation

$$\Xi(X,Y) := \sup_{f\in\mathcal{F}} \left| \frac{1}{m}\sum_{i=1}^{m} l(x_i, y_i, f(x_i)) - \mathbf{E}_{(x,y)}\left[l(x,y,f(x))\right] \right|$$

- If we change single observation pair

$$\left|\Xi(X,Y) - \Xi(X^{-i}\cup\{x_i'\}, Y^{-i}\cup\{y_i'\})\right| \leq L/m$$

- Apply McDiarmid's bound to get

$$\Pr\left\{\left|\Xi(X,Y) > \mathbf{E}_{X,Y}[\Xi(X,Y)]\right| > \epsilon\right\} \leq 2\exp\left(-2m\epsilon^2 L^{-2}\right)$$

- Worst case deviation not far from typical case

Rademacher Averages

$$\mathbf{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m} l(x_i, y_i, f(x_i)) - \mathbf{E}_{(x,y)}\left[l(x,y,f(x))\right]\right|\right]$$

$$= \mathbf{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m} l(x_i, y_i, f(x_i)) - \mathbf{E}_{X',Y'}\frac{1}{m}\sum_{i=1}^{m}\left[l(x'_i, y'_i, f(x'_i))\right]\right|\right]$$

$$\leq \mathbf{E}_{X,Y,X',Y'}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m}\left[l(x_i, y_i, f(x_i)) - l(x'_i, y'_i, f(x'_i))\right]\right|\right]$$

$$= \mathbf{E}_{X,Y,X',Y'}\mathbf{E}_{\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left[l(x_i, y_i, f(x_i)) - l(x'_i, y'_i, f(x'_i))\right]\right|\right]$$

$$\leq \frac{2}{m}\mathbf{E}_{X,Y}\mathbf{E}_{\sigma}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i l(x_i, y_i, f(x_i))\right]$$

Rademacher Averages

- Putting it all together

$$R[f] \leq R_{\mathrm{emp}}[f] + 2\mathcal{R}[\mathcal{F}, m] + L\sqrt{\frac{\log 2/\delta}{2m}}$$

behavior for random labels

averaging

- Rademacher average can be bounded easily for linear function classes by solving a convex optimization problem.

Some Alternatives

- Validation set
  - Train on training set (e.g. 90% of the data)
  - Check performance on remaining 10%
  - Use only if dataset is huge and few tests
- Crossvalidation
  - Average over validation sets (e.g. 10 fold)
  - Nested cross-validation for model selection (e.g. 10-fold in each fold to find parameters)
- Bayesian statistics