Introduction to Machine Learning

Vapnik–Chervonenkis Theory

Barnabás Póczos

MACHINE LEARNING DEPARTMENT

Carnegie Mellon.
School of Computer Science

Empirical Risk and True Risk

Empirical Risk

For simplicity, let $L(x, y, f(x)) = L(y, f(x))$

Shorthand:

True risk of f (deterministic): $R(f) = R_{L,P}(f) = \mathbb{E}[L(Y, f(X))]$

Bayes risk: $R^* = R^*_{L,P} = \inf_{f:\mathcal{X}\to\mathbb{R}} R(f)$

We dont know P, and hence we don't know $R(f)$ either.

Let us use the empirical counter part:

Empirical risk: $\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i))$

Empirical Risk Minimization

$$R(f) = R_{L,P}(f) = \mathbb{E}[L(Y, f(X))] \qquad R^* = R^*_{L,P} = \inf_{f:\mathcal{X}\to\mathbb{R}} R(f)$$

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i))$$

Law of Large Numbers:

For each fixed f, $\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i)) \xrightarrow{n\to\infty} R(f)$

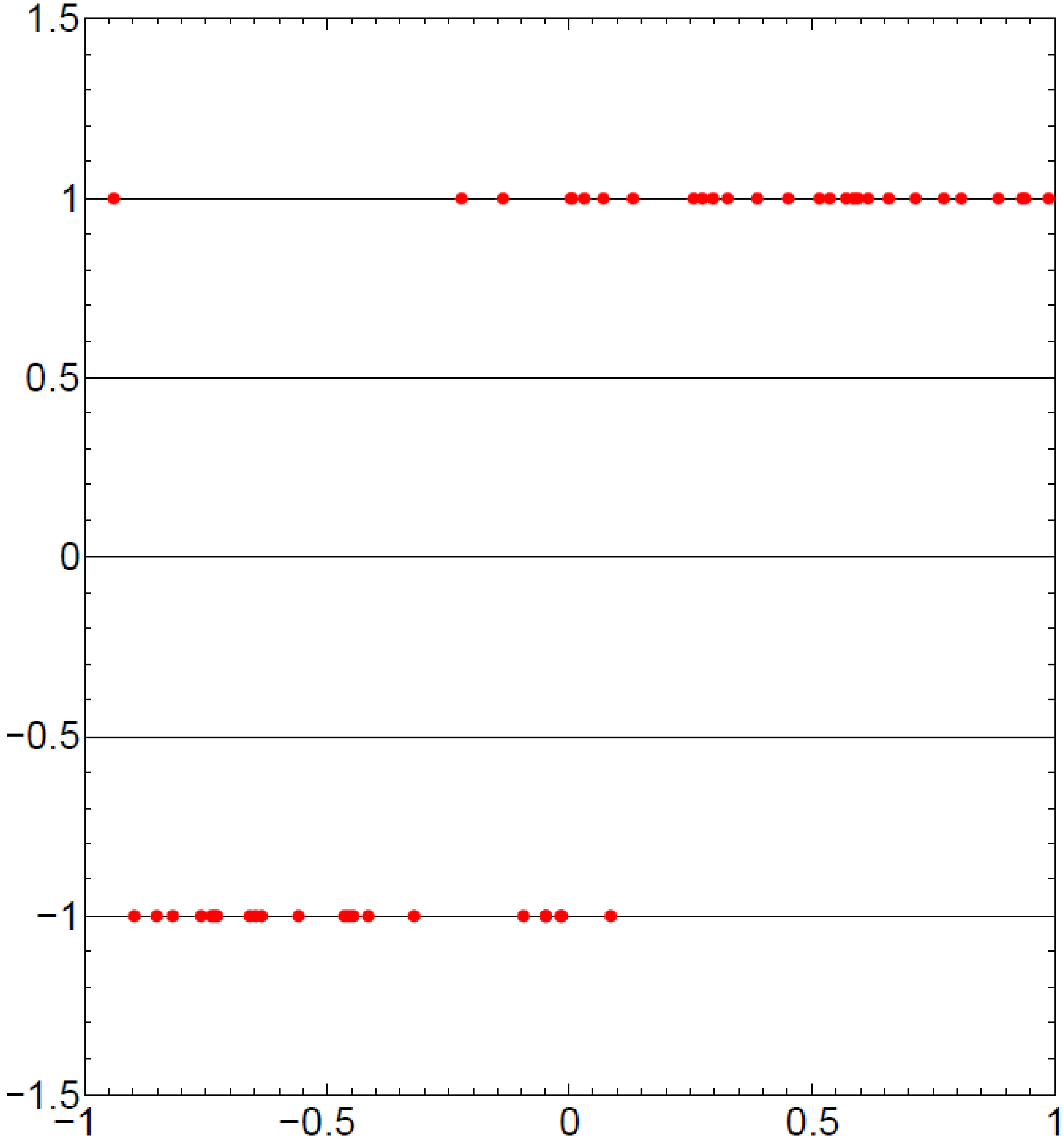
Empirical risk is converging to the true risk

We need $\inf_{f:\mathcal{X}\to\mathbb{R}} R(f)$, so let us calculate $\inf_{f:\mathcal{X}\to\mathbb{R}} \widehat{R}_n(f)$!

$$\inf_{f:\mathcal{X}\to\mathbb{R}} \widehat{R}_n(f) = \inf_{f:\mathcal{X}\to\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i))$$

This is a **terrible idea** to optimize over all possible $f : \mathcal{X} \to \mathbb{R}$ functions! [Extreme overfitting]

Overfitting in Classification with ERM



Picture from David Pal

Generative model:

$$X \sim U[-1,1]$$

$$\Pr(Y=1|X>0)=0.9$$
$$\Pr(Y=-1|X>0)=0.1$$

$$\Pr(Y=1|X<0)=0.1$$
$$\Pr(Y=-1|X<0)=0.9$$

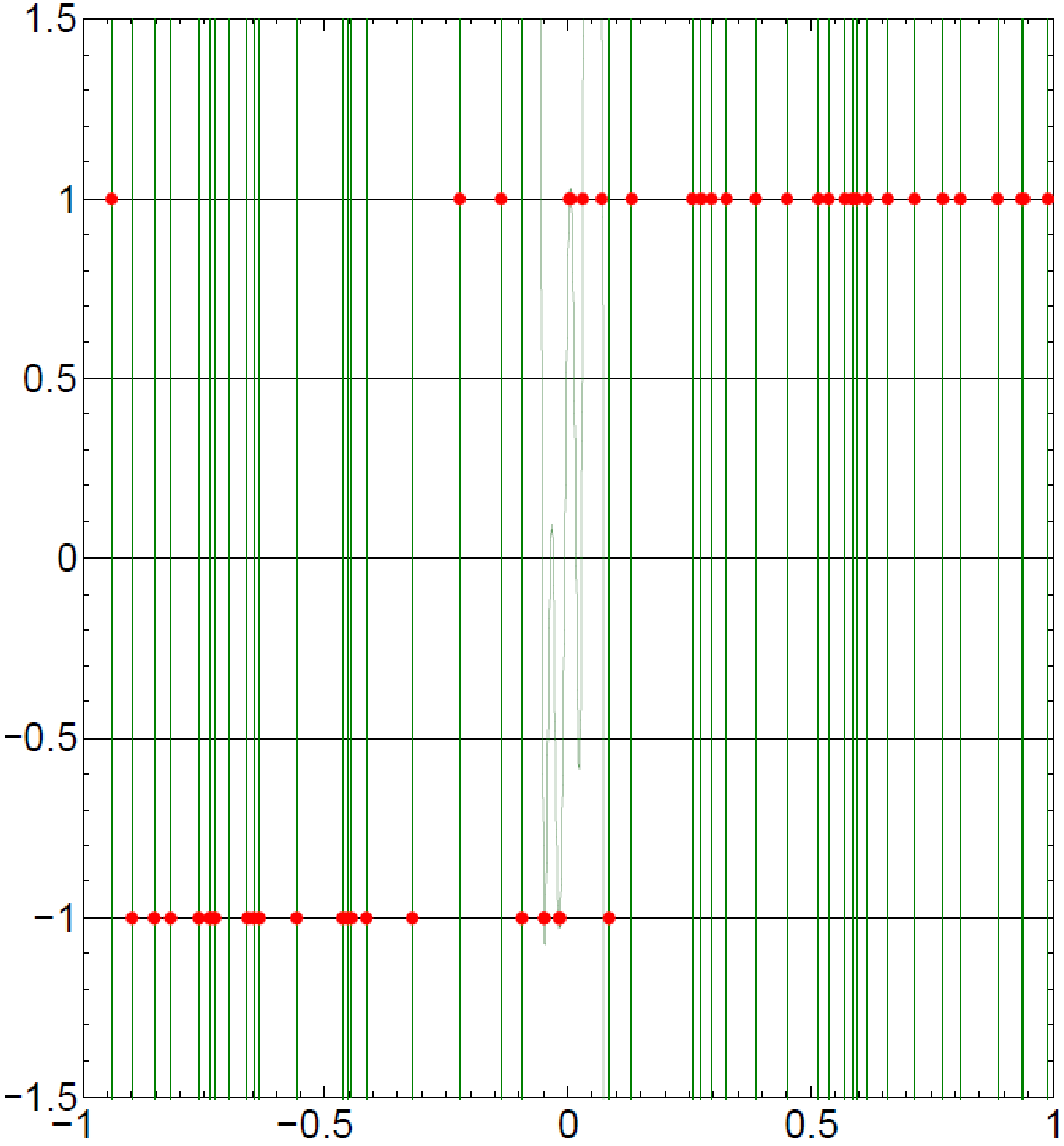
Bayes classifier:

$$f^* = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x \leq 0 \end{cases}$$

Bayes risk:

$$R^* = \Pr(Y \neq f^*(X)) = 0.1$$

Overfitting in Classification with ERM



Picture from David Pal

n-order thresholded polynomials

$$\mathcal{F} = \{f(x) = sign(\sum_{i=0}^{n} a_i x^i)\}$$

$$f_n^* = \arg\min_{f \in \mathcal{F}} \widehat{R}_n(f)$$

Empirical risk:

$$\widehat{R}_n(f_n^*) = 0$$

True risk of f_n^*=0.5

$$R(f_n^*) = \Pr(Y \neq f_n^*(X)) = 0.5$$

Bayes risk:

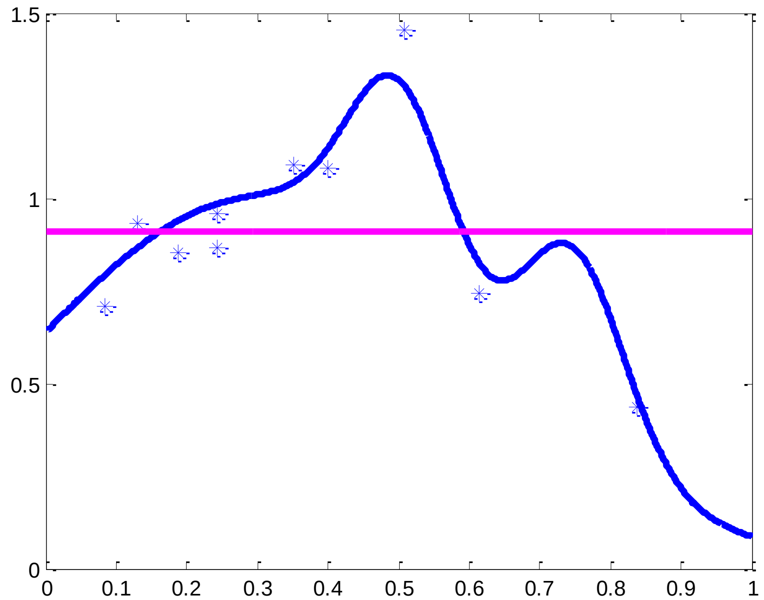
$$R^* = \Pr(Y \neq f^*(X)) = 0.1$$

Overfitting in Regression

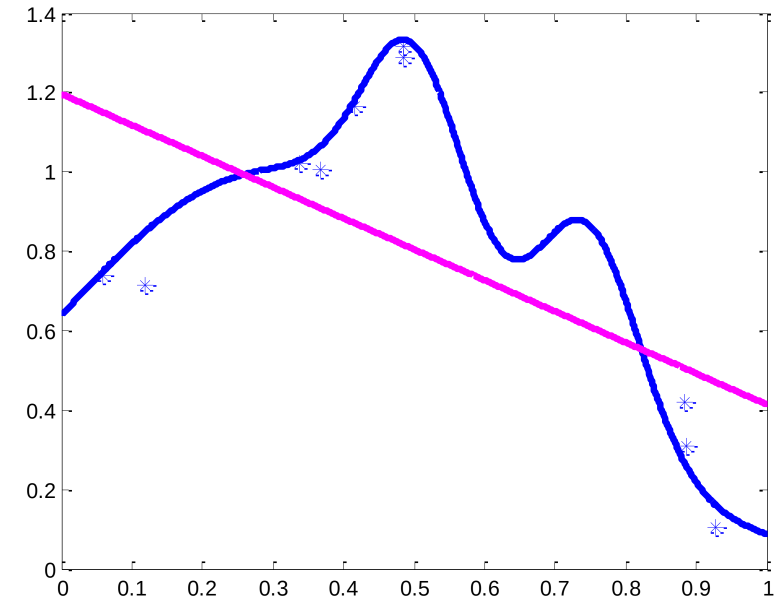
If we allow very complicated predictors, we could overfit the training data.

Examples: Regression (Polynomial of order k-1 – degree k)

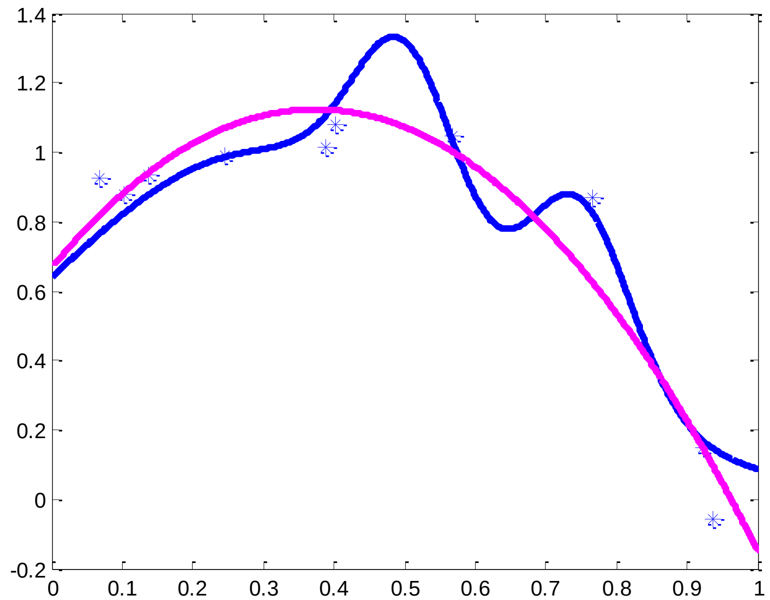
k=1
constant



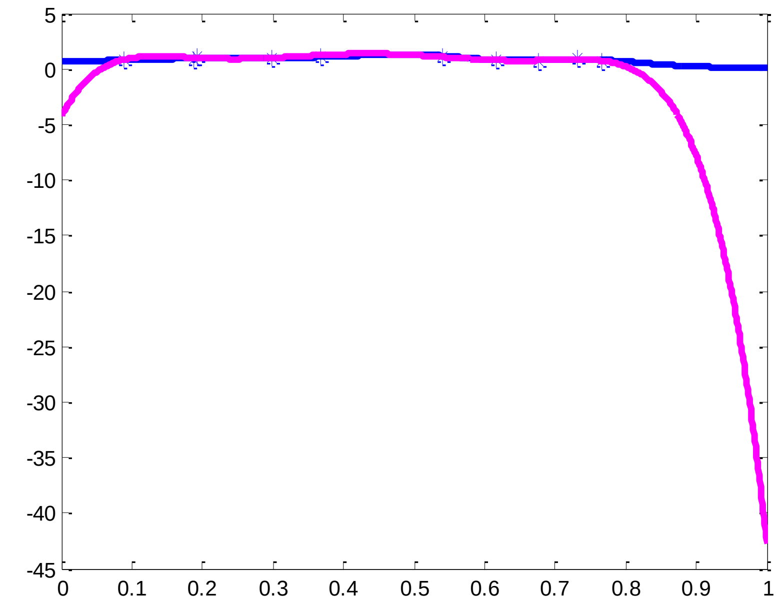
k=2
linear



k=3
quadratic



k=7
6th order



Solutions to Overfitting

Terrible idea to optimize over all possible $f : \mathcal{X} \to \mathbb{R}$ functions! [Extreme overfitting]

$\Rightarrow$ minimze over a smaller function set $\mathcal{F}$.

Empirical risk minimization over the function set $\mathcal{F}$.

$$\boxed{f_n^* = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i))}$$

Solutions to Overfitting
Structural Risk Minimization

Empirical risk minimization over the function set $\mathcal{F}$.

$$f_n^* = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i))$$

Notation: $R_{\mathcal{F}}^* = \inf_{f \in \mathcal{F}} \mathbb{E}[L(Y, f(X))]$ Risk

$\widehat{R}_{n,\mathcal{F}}^* = \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i))$ Empirical risk

Goal: $R_{\mathcal{F}}^* - R^* \geq 0$ needs to be small.

(Model error, Approximation error)

Risk in $\mathcal{F}$ - Bayes risk

Solution: **Structural Risk Minimzation (SRM)**

Let $\mathcal{F}_n$ increase with the smaple size n ($\mathcal{F}_{n+1} \supset \mathcal{F}_n$), and let $\mathcal{F}_{n+1}$ contain more complex functions than $\mathcal{F}_n$

Big Picture

Ultimate goal: $R(f_n^*) - R^* = 0$

ERM: $f_n^* = f_{n,\mathcal{F}}^* = \arg\min_{f \in \mathcal{F}} \widehat{R}_n(f) = \arg\min_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i))$

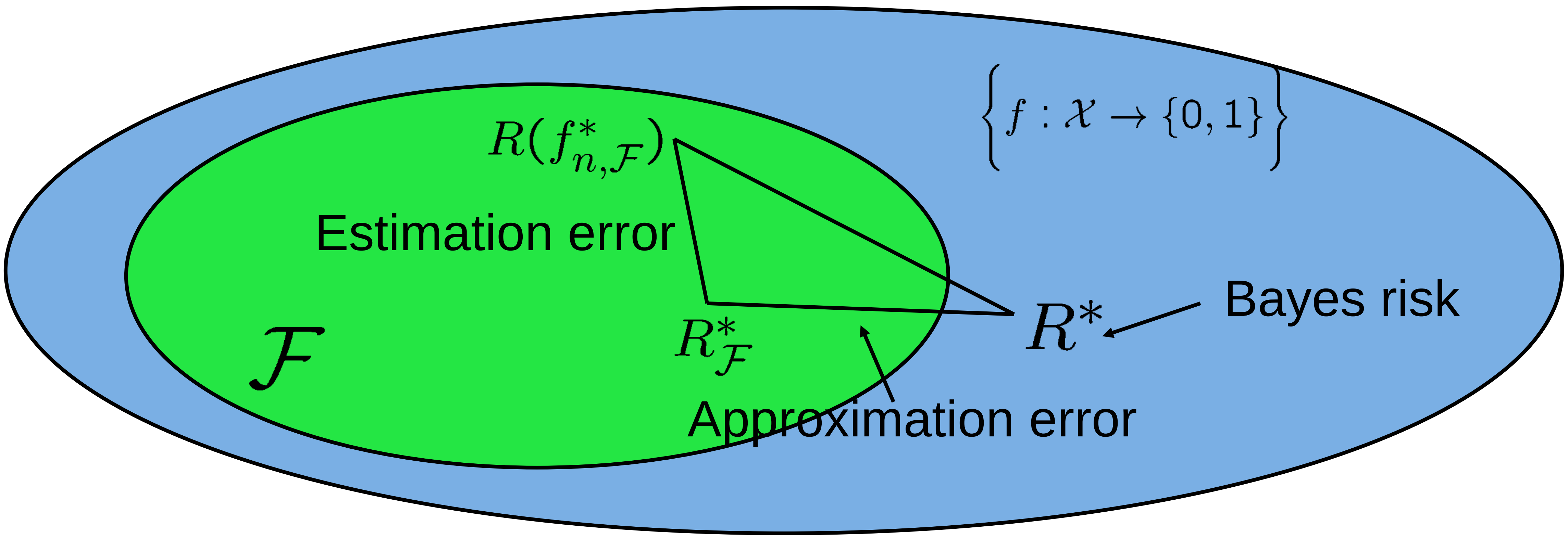
Risk of the classifier $f_{n,\mathcal{F}}^*$

$$R(f_{n,\mathcal{F}}^*) - R^* = \overbrace{R(f_{n,\mathcal{F}}^*) - R_{\mathcal{F}}^*}^{\text{Estimation error}} + \overbrace{R_{\mathcal{F}}^* - R^*}^{\text{Approximation error}}$$

Bayes risk

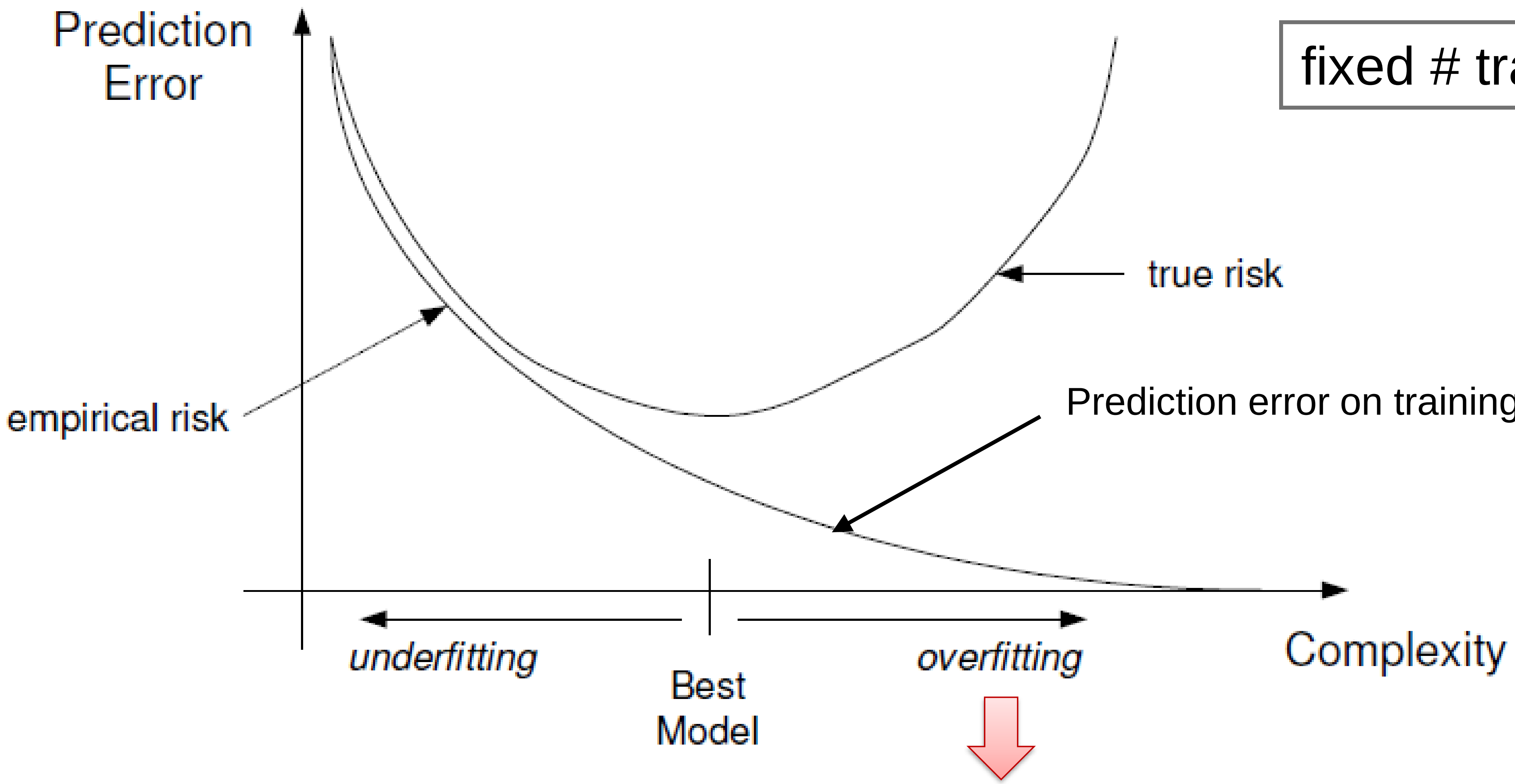
Bayes risk

$R_{\mathcal{F}}^* = \inf_{g \in \mathcal{F}} R(g)$ Best classifier in $\mathcal{F}$



Effect of Model Complexity

If we allow very complicated predictors, we could overfit the training data.



fixed # training data

Empirical risk is no longer a good indicator of true risk

Classification using the 0-1 loss

The Bayes Classifier

$$L(y, f(x)) = \begin{cases} 1 & y \neq f(x) \\ 0 & y = f(x) \end{cases}$$

$$\begin{aligned} R^* &= \inf_{f:\mathcal{X}\to\mathbb{R}} R(f) \\ &= \inf_{f:\mathcal{X}\to\mathbb{R}} \mathbb{E}[L(Y, f(X))] \\ &= \inf_{f:\mathcal{X}\to\mathbb{R}} \Pr(Y \neq f(X)) \end{aligned}$$

$$\begin{aligned} f^* &= \arg \inf_{f:\mathcal{X}\to\mathbb{R}} R(f) \\ &= \arg \inf_{f:\mathcal{X}\to\mathbb{R}} \mathbb{E}[L(Y, f(X))] \\ &= \arg \inf_{f:\mathcal{X}\to\mathbb{R}} \Pr(Y \neq f(X)) \end{aligned}$$

Lemma I: $\Pr(Y \neq f^*(X)) \leq \Pr(Y \neq f(X)) \quad \forall f$

Lemma II: $f^* = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \eta(x) \leq 1/2 \end{cases}$ $\qquad \eta(x) = \mathbb{E}[Y = 1|x]$

Proofs: Lemma I: Trivial from definition

Lemma II: Surprisingly long calculation

The Bayes Classifier

$R(f) = \Pr[Y \neq f(X)]$	$R^* = R(f^*) = \inf_{f:\mathcal{X}\to\mathbb{R}} R(f)$	$f^* = \arg\min_{f:\mathcal{X}\to\mathbb{R}} R(f)$
	$R^*_{\mathcal{F}} = R(f^*_{\mathcal{F}}) = \inf_{f\in\mathcal{F}} R(f)$	$f^*_{\mathcal{F}} = \arg\min_{f\in\mathcal{F}} R(f)$
$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \neq f(X_i)\}}$	$\widehat{R}^*_{n,\mathcal{F}} = \inf_{f\in\mathcal{F}} \widehat{R}_n(f)$	$f^*_{n,\mathcal{F}} = \arg\min_{f\in\mathcal{F}} \widehat{R}_n(f)$

This is what the learning algorithm produces

We will need these definitions, please copy it!

$R(f) =$ Risk

$R^* =$ Bayes risk

$\widehat{R}_n(f) =$ Emprical risk

$f^* =$ Bayes classifier

$f^*_n = f^*_{n,\mathcal{F}} =$ the classifier that the learning algorithm produces

The Bayes Classifier

$R(f) = \Pr[Y \neq f(X)]$	$R^* = R(f^*) = \inf_{f:\mathcal{X}\to\mathbb{R}} R(f)$	$f^* = \arg\min_{f:\mathcal{X}\to\mathbb{R}} R(f)$
	$R^*_{\mathcal{F}} = R(f^*_{\mathcal{F}}) = \inf_{f\in\mathcal{F}} R(f)$	$f^*_{\mathcal{F}} = \arg\min_{f\in\mathcal{F}} R(f)$
$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i\neq f(X_i)\}}$	$\widehat{R}^*_{n,\mathcal{F}} = \inf_{f\in\mathcal{F}} \widehat{R}_n(f)$	$f^*_{n,\mathcal{F}} = \arg\min_{f\in\mathcal{F}} \widehat{R}_n(f)$

This is what the learning algorithm produces

Theorem I: Bound on the Estimation error

The true risk of what the learning algorithm produces

$$|R(f^*_{n,\mathcal{F}}) - R^*_{\mathcal{F}}| \leq 2 \sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$$

How far $f^*_{n,\mathcal{F}}$ is from the optimal in $\mathcal{F}$

$\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$ can be used to get an upper bound for this

Proof of Theorem 1

Theorem I: Bound on the Estimation error

The true risk of what the learning algorithm produces

$$|R(f^*_{n,\mathcal{F}}) - R^*_{\mathcal{F}}| \leq 2 \sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)|$$

How far $f^*_{n,\mathcal{F}}$ is from the optimal in $\mathcal{F}$

Proof:

The Bayes Classifier

$R(f) = \Pr[Y \neq f(X)]$	$R^* = R(f^*) = \inf_{f:\mathcal{X}\to\mathbb{R}} R(f)$	$f^* = \arg \inf_{f:\mathcal{X}\to\mathbb{R}} R(f)$
	$R^*_{\mathcal{F}} = R(f^*_{\mathcal{F}}) = \inf_{f\in\mathcal{F}} R(f)$	$f^*_{\mathcal{F}} = \arg \inf_{f\in\mathcal{F}} R(f)$
$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \neq f(X_i)\}}$	$\widehat{R}^*_{n,\mathcal{F}} = \inf_{f\in\mathcal{F}} \widehat{R}_n(f)$	$f^*_{n,\mathcal{F}} = \arg \min_{f\in\mathcal{F}} \widehat{R}_n(f)$

Theorem II:

This is what the learning algorithm produces

$$|\widehat{R}_n(f^*_{n,\mathcal{F}}) - R(f^*_{n,\mathcal{F}})| \leq \sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$$

How far the empirical risk of $f^*_{n,\mathcal{F}}$ is from its true risk.

Proof: Trivial

$\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$ can be used to get an upper bound for this

Corollary

Corollary:

$$|\widehat{R}_n(f^*_{n,\mathcal{F}}) - R^*_{\mathcal{F}}| \leq 3 \sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)|$$

True risk of the best possible classifer in $\mathcal{F}$ (unknown)

Empirical risk of the learned classifier $f^*_{n,\mathcal{F}}$ (known)

Main message:

It's enough to derive upper bounds for

$\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)|$

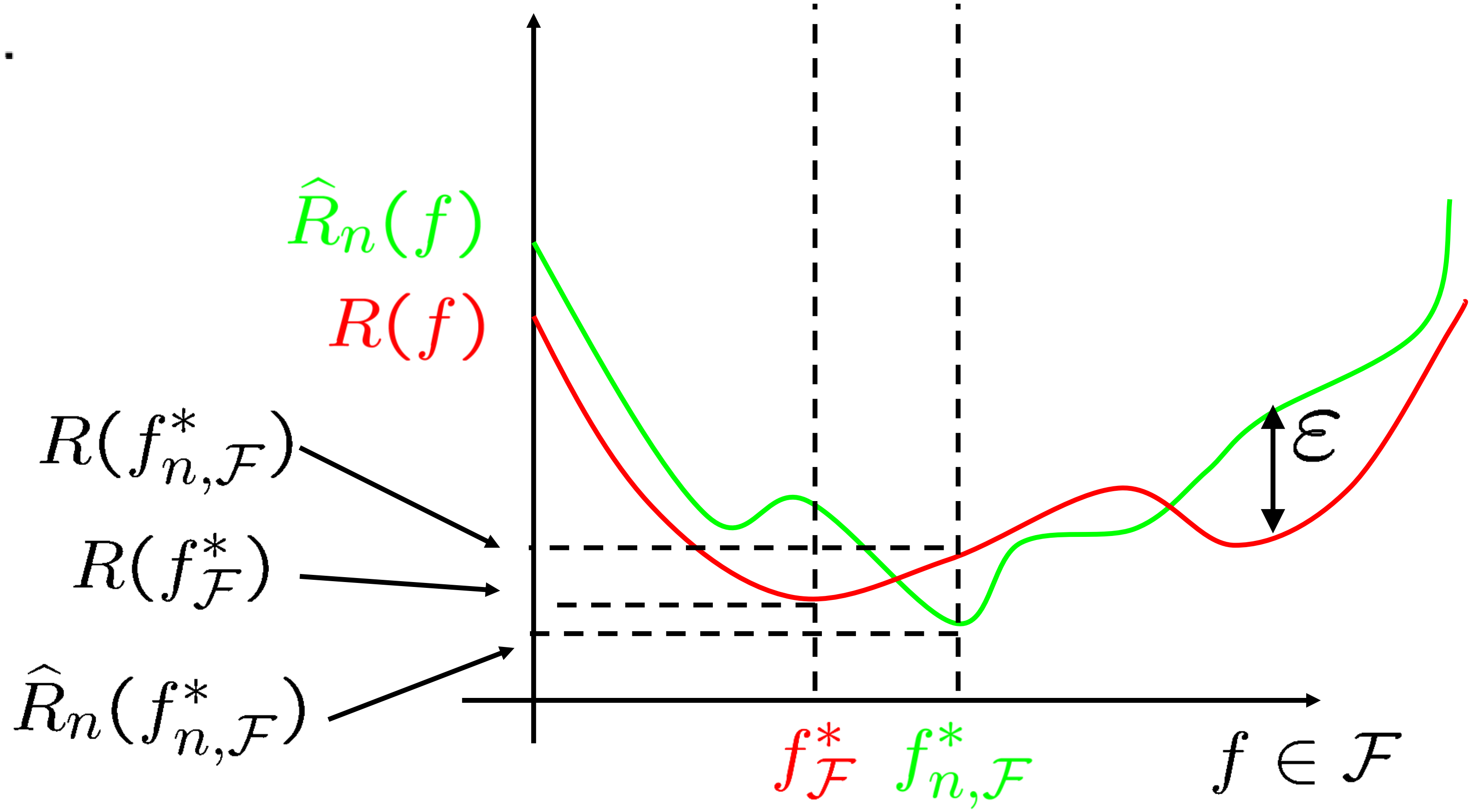
Illustration of the Risks

$$|\widehat{R}_n(f^*_{n,\mathcal{F}}) - R(f^*_{n,\mathcal{F}})| \leq \sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| = \varepsilon$$

$$|R(f^*_{n,\mathcal{F}}) - R(f^*_{\mathcal{F}})| \leq 2 \sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| = 2\varepsilon$$

$$|\widehat{R}_n(f^*_{n,\mathcal{F}}) - R(f^*_{\mathcal{F}})| \leq 3 \sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| = 3\varepsilon$$

$\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)|$ can be used to get an upper bound for these.



It's enough to derive upper bounds for

$$\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)|$$

It is a random variable that we need to bound!
We will bound it with tail bounds!

Hoeffding's inequality (1963)

$$\left.\begin{array}{r} Z_1, \ldots, Z_n \text{ independent} \\ Z_i \in [a_i, b_i] \\ \varepsilon > 0 \end{array}\right\} \Rightarrow$$

$$\Rightarrow \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])\right| > \varepsilon\right) \leq 2\exp\left(\frac{-2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Special case

$$Z_i \text{ is Bernoulli}(p) \Rightarrow \sum_{i=1}^{n} Z_i \text{ is Binomial}(n, p)$$

$$\Rightarrow \Pr\left(\left|\sum_{i=1}^{n}\frac{1}{n}(Z_i - \mathbb{E}[Z_i])\right| > \varepsilon\right) \leq 2\exp\left(\frac{-2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(1-0)^2}\right) = 2\exp\left(-2n\varepsilon^2\right)$$

Binomial distributions

Our goal is to bound $\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \neq f(X_i)\}} \Rightarrow n\widehat{R}_n(f) = \sum_{i=1}^{n} 1_{\{Y_i \neq f(X_i)\}} \sim Binom(n,p)$$

Bernoulli(p)

where $p = \mathbb{E}[1_{\{Y\neq f(X)\}}] = \Pr(Y \neq f(X)) = R(f)$

Let $Z_i = 1_{\{Y_i \neq f(X_i)\}} \sim \text{Bernoulli}(p)$

$$\Rightarrow |\widehat{R}_n(f) - R(f)| = |\frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \neq f(X_i)\}} - p| = |\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i]|$$

Therefore, from Hoeffding we have:

$$\boxed{\Pr(|\widehat{R}_n(f) - R(f)| > \varepsilon) \leq 2\exp\left(-2n\varepsilon^2\right)}$$

Yuppie!!!

Inversion

From Hoeffding we have:

$$\boxed{\Pr(|\widehat{R}_n(f) - R(f)| \geq \varepsilon) \leq 2\exp\left(-2n\varepsilon^2\right)}$$

$$\text{Let } 2\exp\left(-2n\varepsilon^2\right) \leq \delta$$

$$-2n\varepsilon^2 \leq \log(\delta/2)$$

$$\varepsilon^2 \geq \frac{\log(2/\delta)}{2n}$$

Therefore,

$$\Pr\left(|\widehat{R}_n(f) - R(f)| \geq \sqrt{\frac{\log(2/\delta)}{2n}}\right) \leq \delta$$

$$\Pr\left(|\widehat{R}_n(f) - R(f)| < \sqrt{\frac{\log(2/\delta)}{2n}}\right) \geq 1 - \delta$$

Usually $\delta = 0.05$ (5%), and $1 - \delta = 0.95$ (95%)

Union Bound

Our goal is to bound: $\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

We already know: $\Pr(|\widehat{R}_n(f) - R(f)| > \varepsilon) \leq 2\exp\left(-2n\varepsilon^2\right)$

Theorem: [tail bound on the 'deviation' in the worst case]

Let $\mathcal{F} = \{f : \mathcal{X} \to \{0,1\}\}$, and $\boxed{|\mathcal{F}| \leq N}$

$$\Rightarrow \Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| > \varepsilon\right) \leq 2N\exp\left(-2n\varepsilon^2\right)$$

Worst case error

This is not the worst classifier in terms of classification accuracy!

Worst case means that the empirical risk of classifier *f* is the furthest from its true risk!

Proof: $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$

$$\Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| > \varepsilon\right) = \Pr\left(\bigcup_{f\in\mathcal{F}} \left\{|\widehat{R}_n(f) - R(f)| > \varepsilon\right\}\right)$$

$$\Pr\left(\bigcup_{f\in\mathcal{F}} \left\{|\widehat{R}_n(f) - R(f)| > \varepsilon\right\}\right) \leq \sum_{f\in\mathcal{F}} \Pr\left(|\widehat{R}_n(f) - R(f)| > \varepsilon\right)$$

Inversion of Union Bound

We already know: Let $\mathcal{F} = \{f : \mathcal{X} \to \{0,1\}\}$, and $|\mathcal{F}| \le N$

$$\Rightarrow \Pr\left(\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| > \varepsilon\right) \le 2N \exp\left(-2n\varepsilon^2\right)$$

Let $2N \exp\left(-2n\varepsilon^2\right) \le \delta \;\Rightarrow\; -2n\varepsilon^2 \le \log(\delta/(2N)) \;\Rightarrow\; \varepsilon^2 \ge \frac{\log(2N/\delta)}{2n}$

Therefore,

$$\Pr\left(\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| \ge \sqrt{\frac{\log(N) + \log(2/\delta)}{2n}}\right) \le \delta$$

$$\Pr\left(\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| < \sqrt{\frac{\log(N) + \log(2/\delta)}{2n}}\right) \ge 1 - \delta$$

Inversion of Union Bound

$$\Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| \geq \sqrt{\frac{\log(N) + \log(2/\delta)}{2n}}\right) \leq \delta$$

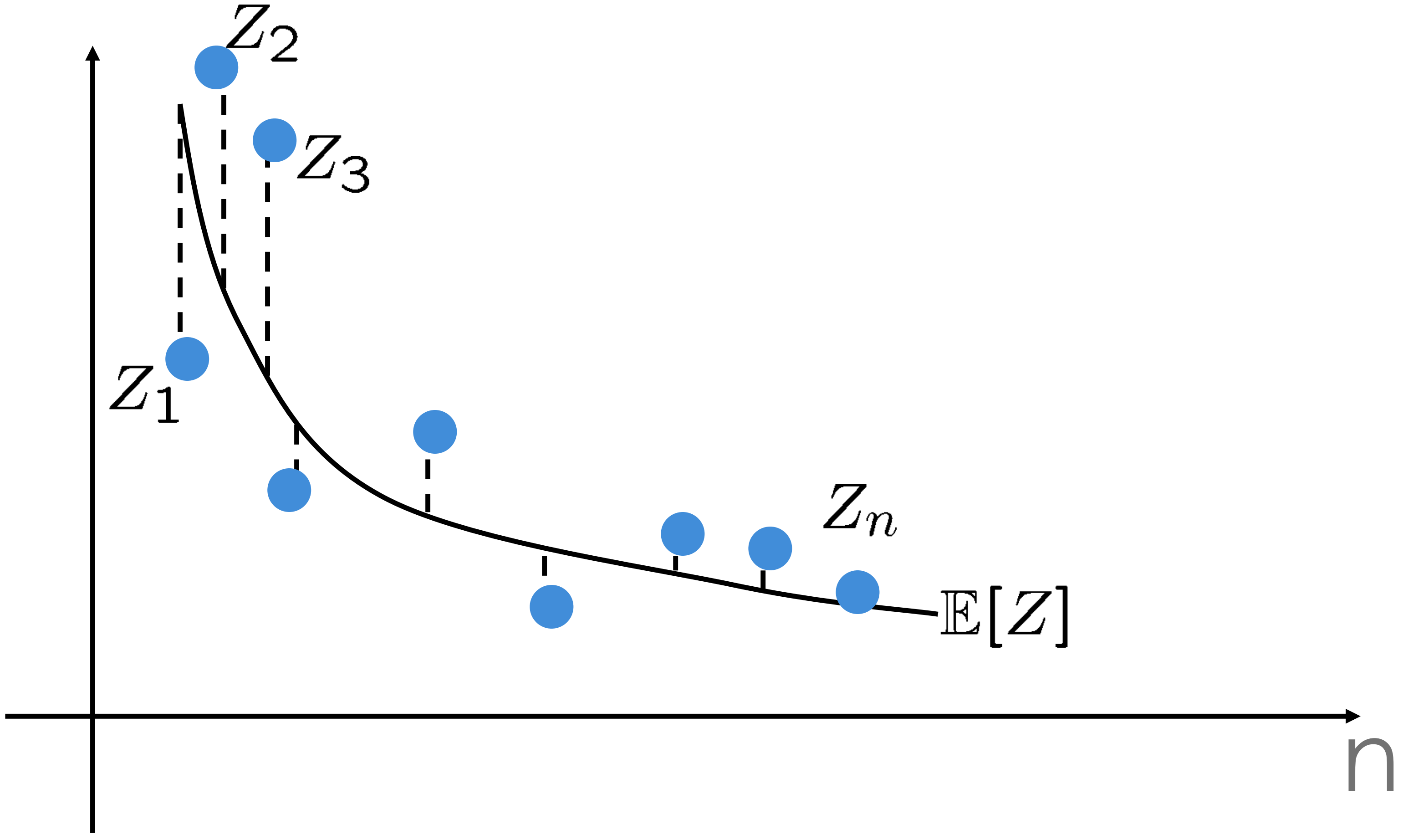
$$\Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| < \sqrt{\frac{\log(N) + \log(2/\delta)}{2n}}\right) \geq 1 - \delta$$

- The larger the *N*, the looser the bound
- This results is distribution free: True for all P(X,Y) distributions
- It is useless if N is big, or infinite… (e.g. all possible hyperplanes)

It can be fixed with McDiarmid inequality and VC dimension…

Concentration and Expected Value

$$Z_n = \sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)|$$



The Expected Error

Our goal is to bound: $\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

We already know: $\Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| > \varepsilon\right) \le 2N \exp\left(-2n\varepsilon^2\right)$

(Tail bound, Concentration inequality)

Theorem: [Expected 'deviation' in the worst case]

Let $\mathcal{F} = \{f : \mathcal{X} \to \{0,1\}\}$, and $|\mathcal{F}| \le N$

$$\Rightarrow \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|\right] \le \sqrt{\frac{\log(2N)}{2n}}$$

Worst case deviation

This is not the worst classifier in terms of classification accuracy!
Worst case means that the empirical risk of classifier *f* is the furthest from its true risk!

Proof: we already know a tail bound. If $Y \ge 0$, then $\mathbb{E}[Y] = \int_0^\infty \Pr(Y \ge z)dz$

(From that actually we get a bit weaker inequality… oh well)

Function classes with infinite many elements

McDiarmid's Bounded Difference Inequality

Suppose $X_1, X_2, \ldots, X_n$ are independent and assume that

$$\sup_{x_1, x_2, \ldots, x_n, \widehat{x}_i} |f(x_1, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_{i-1}, \widehat{x}_i, x_{i+1}, \ldots, x_n)| \le c_i$$

$$\text{for } 1 \le i \le n$$

(**Bounded Difference Assumption**: replacing the i-th coordinate x_i changes the value of f by at most c_i.)

It follows that

$$\Pr\{f(X_1, X_2, \ldots, X_n) - E[f(X_1, X_2, \ldots, X_n)] \ge \varepsilon\} \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

$$\Pr\{E[f(X_1, X_2, \ldots, X_n)] - f(X_1, X_2, \ldots, X_n) \ge \varepsilon\} \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

$$\Pr\{|E[f(X_1, X_2, \ldots, X_n)] - f(X_1, X_2, \ldots, X_n)| \ge \varepsilon\} \le 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Bounded Difference Condition

Our main goal is to bound $\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

Lemma:

The **"bounded difference condition"** is satisfied for $\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

Proof:

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{f(X_i)\neq Y_i\}}$$

Let *g* denote the following function:

$$g(Z_1,\ldots,Z_n) = g((X_1,Y_1),\ldots,(X_n,Y_n)) = \sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$$

Observation:

If we change $Z_i = (X_i, Y_i)$, then g can change $c_i = 1/n$ at most.

(Look at how much $\widehat{R}_n(f)$ can change if we change either X_i or Y_i!)

=> McDiarmid can be applied for *g!*

Bounded Difference Condition

The **"bounded difference condition"** is satisfied for $\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

Corollary:

$$\Pr\{g - \mathbb{E}[g] \geq \varepsilon\} \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

for $g = \sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

$c_i = 1/n$

$$\Pr\left\{\left|\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| - \mathbb{E}[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|]\right| \geq \varepsilon\right\} \leq 2\exp\left(-2\varepsilon^2 n\right)$$

$\Rightarrow$ $\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$ is concentrated around its mean!

Therefore, it is enough to study how $\mathbb{E}[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|]$ behaves.

The Vapnik-Chervonenkis inequality does that with the ***shatter coefficient*** (and ***VC dimension)!***

Vapnik-Chervonenkis inequality

Our main goal is to bound $\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

We already know:

$$\Pr\left\{\left|\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| - \mathbb{E}[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|]\right| \geq \varepsilon\right\} \leq 2\exp\left(-2\varepsilon^2 n\right)$$

Vapnik-Chervonenkis inequality:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|\right] \leq 2\sqrt{\frac{\log(2S_{\mathcal{F}}(n))}{n}}$$

Corollary: Vapnik-Chervonenkis theorem:

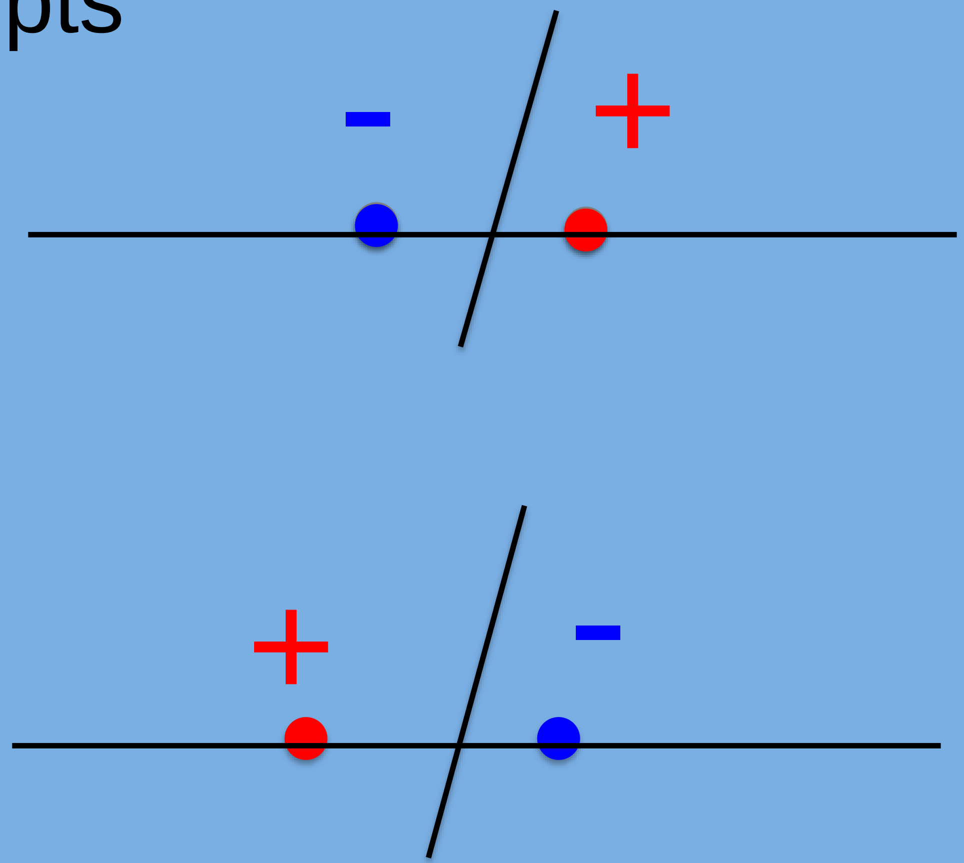
$$\Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| > t\right) \leq 4S_{\mathcal{F}}^2(n)\exp(-nt^2/8)$$

We will define $S_{\mathcal{F}}(n)$ later.

Shattering

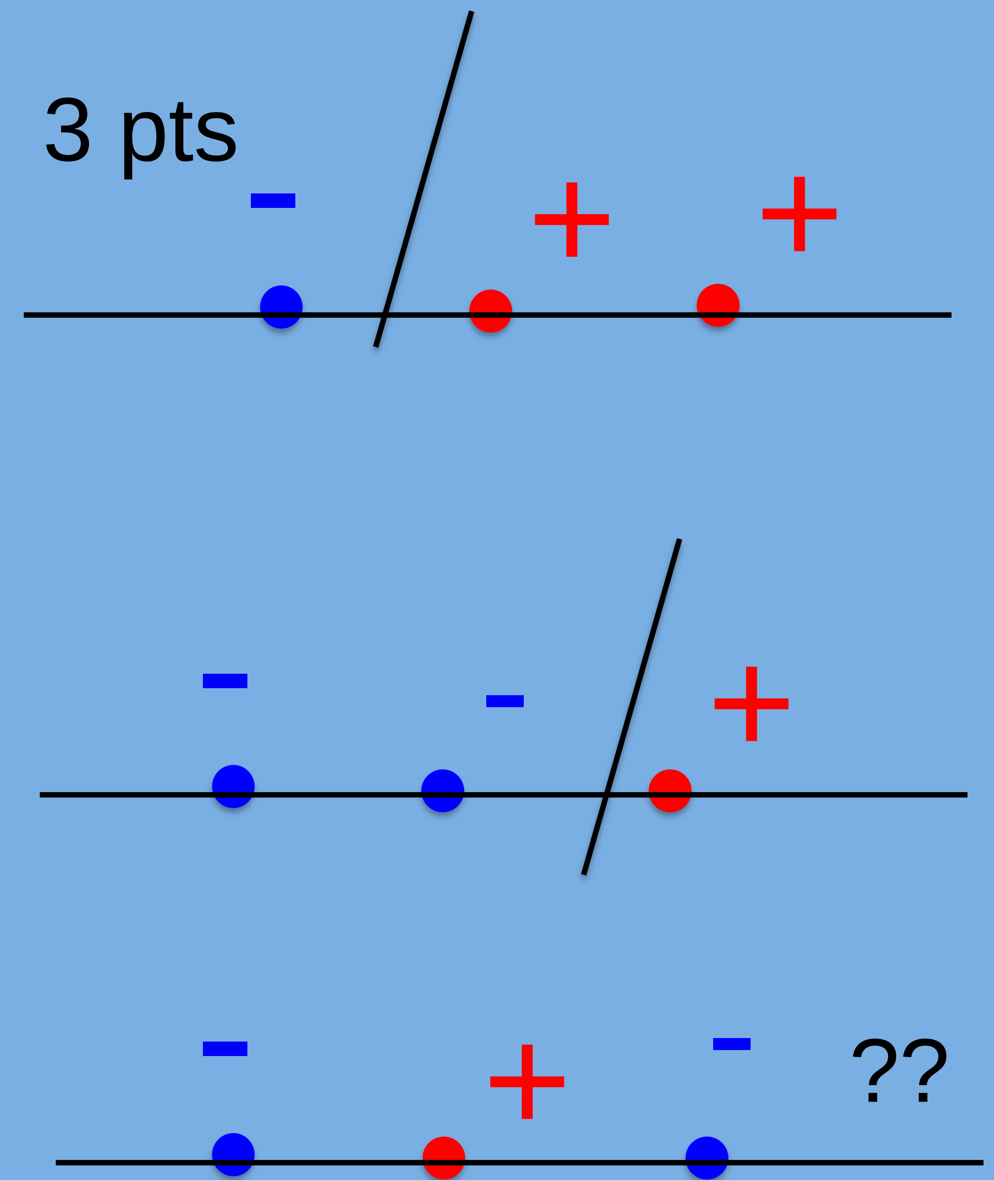
How many points can a linear boundary classify exactly in 1D?

2 pts



There exists placement s.t. all labelings can be classified

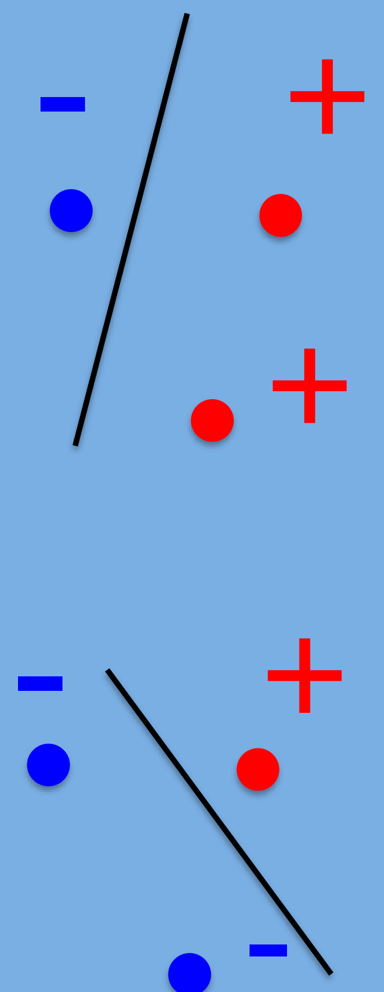
3 pts



The answer is 2

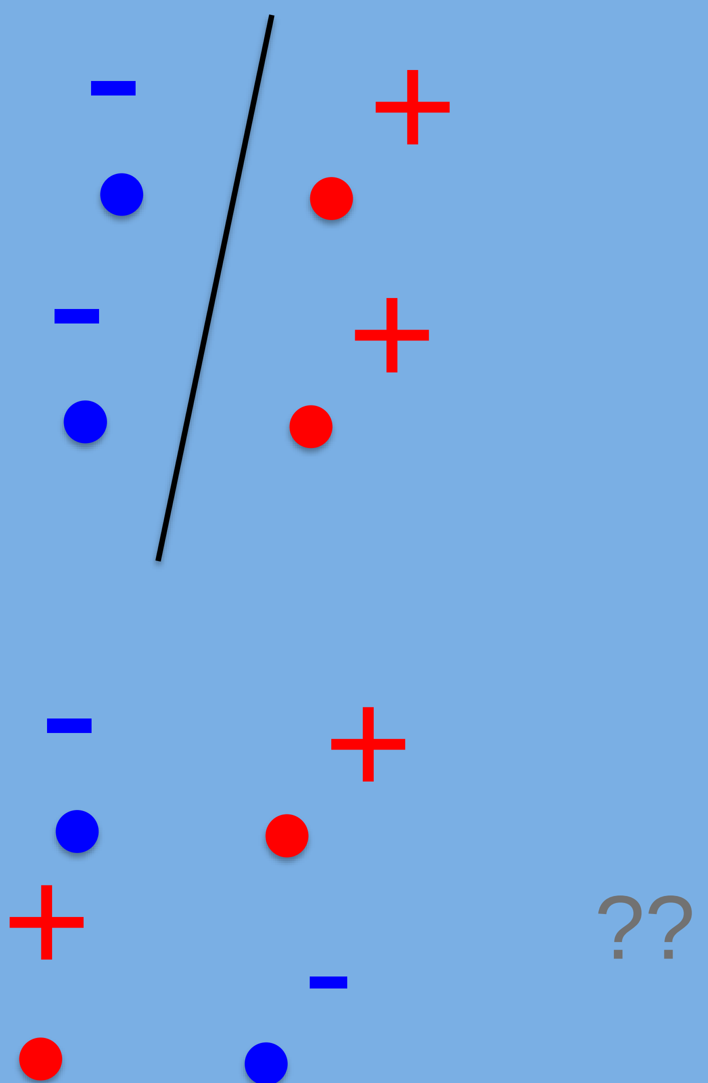
How many points can a linear boundary classify exactly in 2D?

3 pts



There exists a placement s.t. all labelings can be classified

4 pts

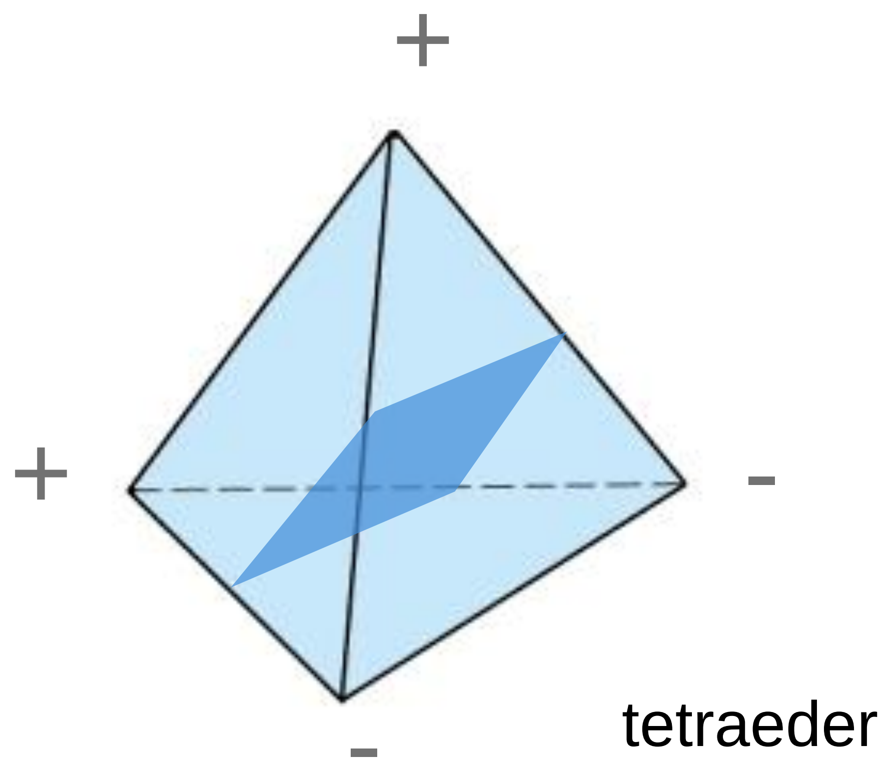


The answer is 3

No matter how we place 4 points, there is a labeling that cannot be classified

How many points can a linear boundary classify exactly in 3D?

The answer is 4



How many points can a linear boundary classify exactly in d-dim?

The answer is d+1

Growth function, Shatter coefficient

Let $\mathcal{F} = \mathcal{X} \rightarrow \{0, 1\}$

How many different behaviour can we get with $[f(x_1), \ldots, f(x_n)]$, $f \in \mathcal{F}$?

$|\mathcal{F}| = 7$

	x_1	x_2	x_3
f_1	0	0	0
f_2	0	1	0
f_3	1	1	1
f_4	1	0	0
f_5	0	1	1
f_6	0	1	0
f_7	1	1	1

Definition

$$S_{\mathcal{F}}(x_1, \ldots, x_n) = |\{f(x_1), \ldots, f(x_n)\}; f \in \mathcal{F}|$$

(=5 in this example)

Growth function, Shatter coefficient

$$S_{\mathcal{F}}(n) = \max_{x_1, \ldots, x_n} |\{f(x_1), \ldots, f(x_n)\}; f \in \mathcal{F}|$$

maximum number of behaviors on *n* points

Growth function, Shatter coefficient

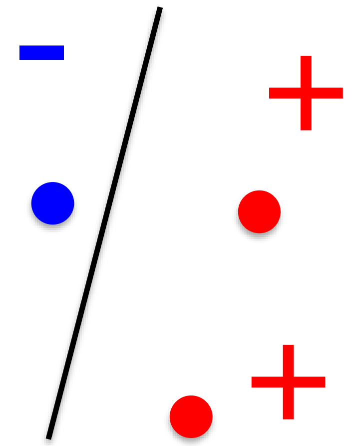
Definition

$$S_{\mathcal{F}}(x_1, \ldots, x_n) = |\{f(x_1), \ldots, f(x_n)\}; f \in \mathcal{F}|$$

Growth function, Shatter coefficient

$$S_{\mathcal{F}}(n) = \max_{x_1, \ldots, x_n} |\{f(x_1), \ldots, f(x_n)\}; f \in \mathcal{F}|$$

maximum number of behaviors on *n* points

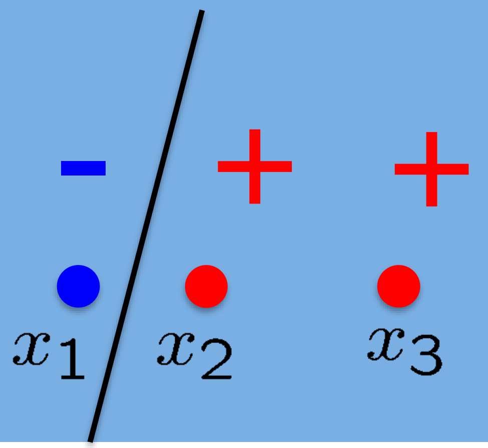


Example: Half spaces in 2D $\Rightarrow S_{\mathcal{F}}(3) = 2^3 = 8$

(Although $\exists x_1, x_2, x_3$ such that $S_{\mathcal{F}}(x_1, x_2, x_3) = 6 < 8$)

$\{\emptyset\}, \{x_1\}, \{x_3\}, \{x_1, x_2\}, \{x_2, x_3\}, \{x_1, x_2, x_3\}$

We can't get $\{x_2\}$ and $\{x_1, x_3\}$



VC-dimension

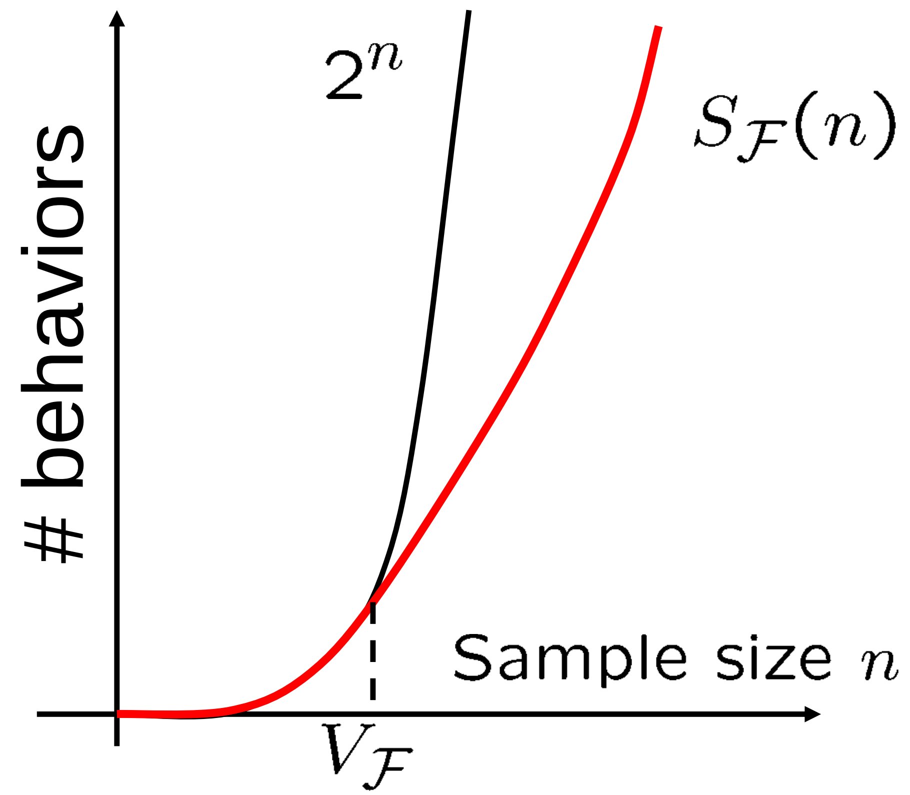
Definition

$$S_{\mathcal{F}}(x_1, \ldots, x_n) = |\{f(x_1), \ldots, f(x_n)\}; f \in \mathcal{F}|$$

Growth function, Shatter coefficient

$$S_{\mathcal{F}}(n) = \max_{x_1, \ldots, x_n} |\{f(x_1), \ldots, f(x_n)\}; f \in \mathcal{F}|$$

maximum number of behaviors on *n* points



Definition: VC-dimension

$$V_{\mathcal{F}} = \max\{n : S_{\mathcal{F}}(n) = 2^n\}$$

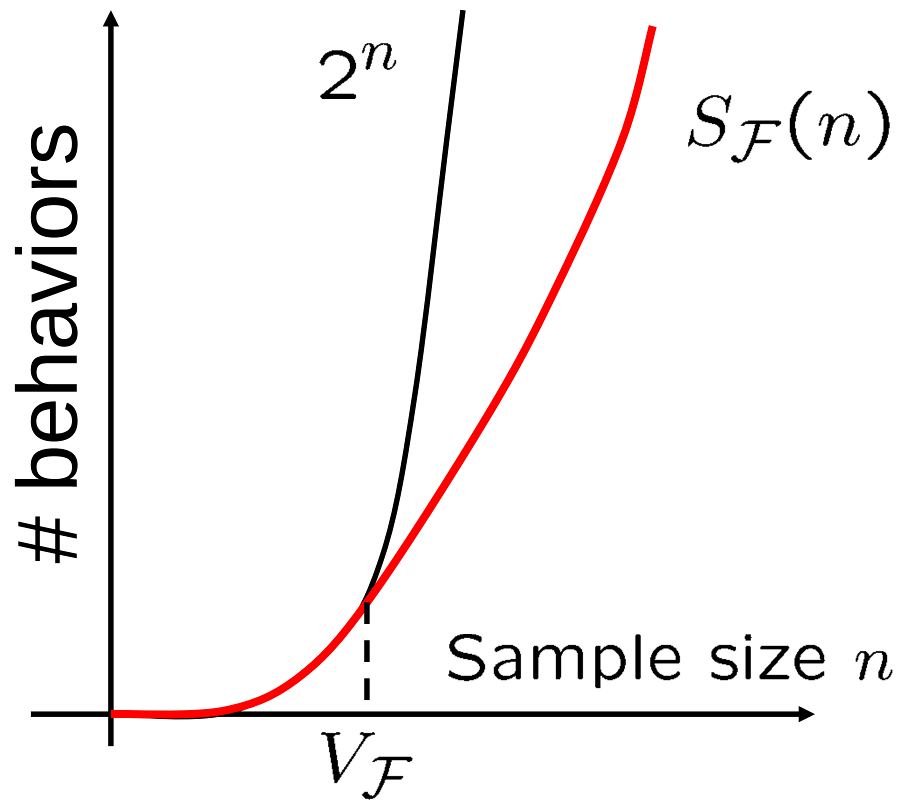
Definition: Shattering

$\mathcal{F}$ shatters the sample $x_1, \ldots, x_n$ iff $\mathcal{F}$ has all the 2^n behaviors on the sample.

Note: $V_{\mathcal{F}}$ is the size of largest shattered sample

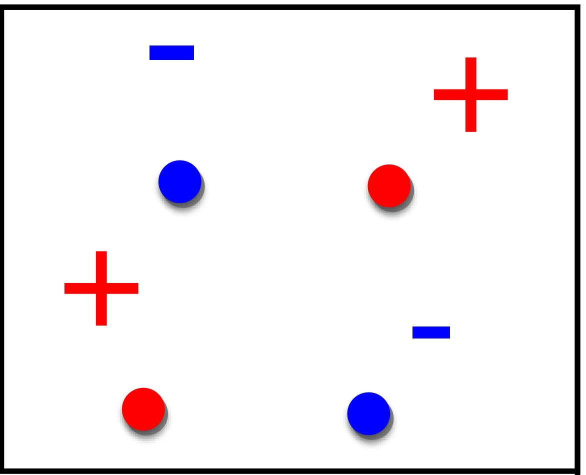
VC-dimension

Definition $V_{\mathcal{F}} = \max\{n : S_{\mathcal{F}}(n) = 2^n\}$



- If the VC dimension is n, then we can find n points that can be shattered, i.e. show 2^n behaviours.
- $n+1$ points never show 2^{n+1} behaviours.

VC-dimension



- You pick set of points $x_1, \ldots, x_n$
  (such that you want to maximize the # of different behaviors)
- Adversary assigns labels $y_1, \ldots, y_n$
- If $VC_{\mathcal{F}} \geq n$, then you find a hypothesis f in $\mathcal{F}$ consistent with the labels, i.e. $f(x_i) = y_i \quad (1 \leq i \leq n)$
- If $VC_{\mathcal{F}} = n$, then for any $n+1$ points, there exists a labeling that cannot be shattered (can't find a hypothesis f in $\mathcal{F}$ consistent with it)

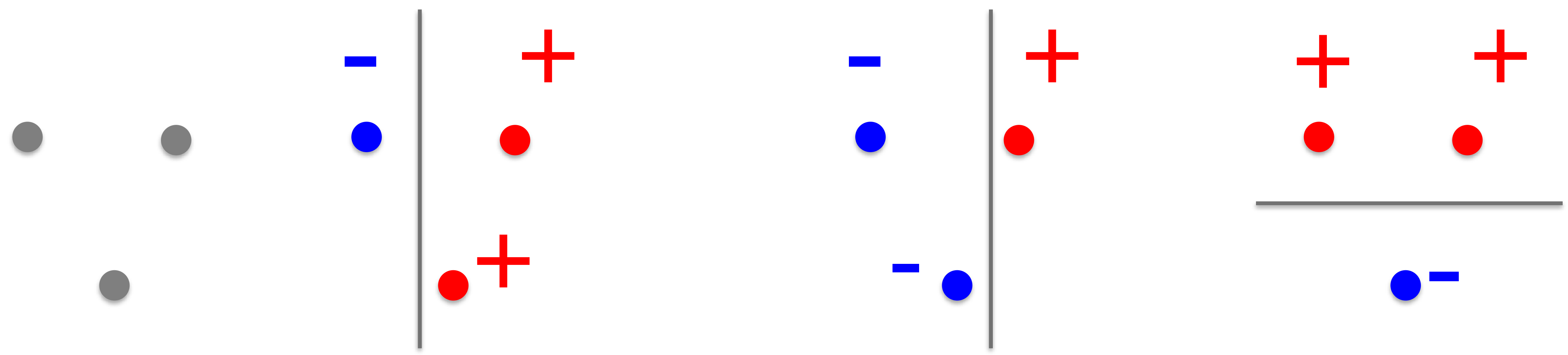
The VC dimension measures how rich $\mathcal{F}$ is.

If the VC dimension is high, e.g. ∞, then it is easy to overfit!

Examples

VC dim of decision stumps (axis aligned linear separator) in 2d

What's the VC dim. of decision stumps in 2d?



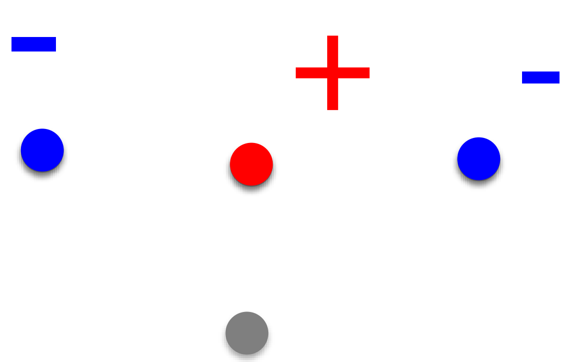
There is a placement of 3 pts that can be shattered => VC dim ≥ 3

VC dim of decision stumps (axis aligned linear separator) in 2d

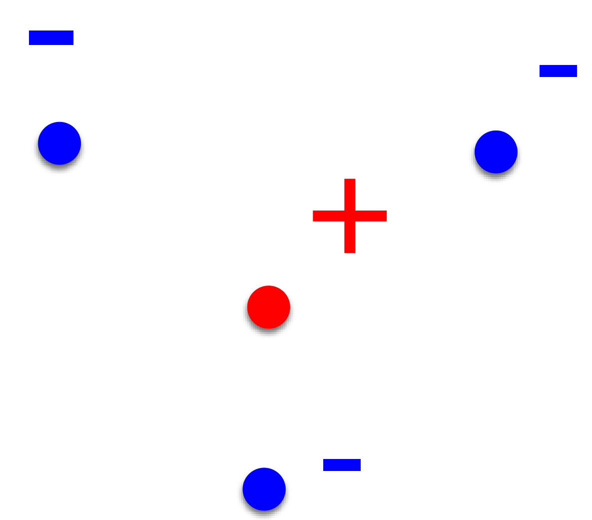
What's the VC dim. of decision stumps in 2d?

If VC dim = 3, then for all placements of 4 pts, there exists a labeling that can't be shattered

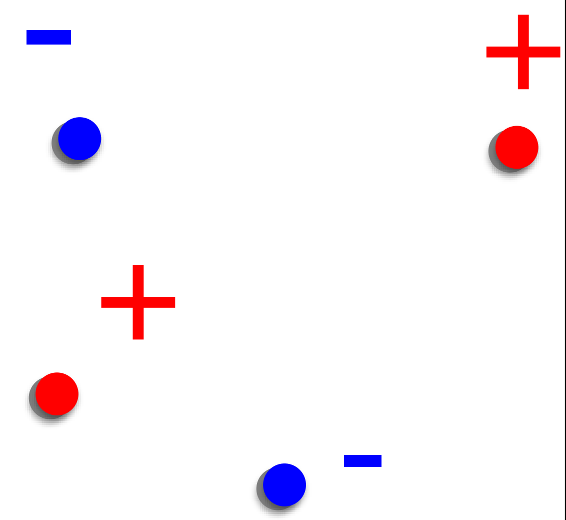
3 collinear



1 in convex hull of other 3



quadrilateral

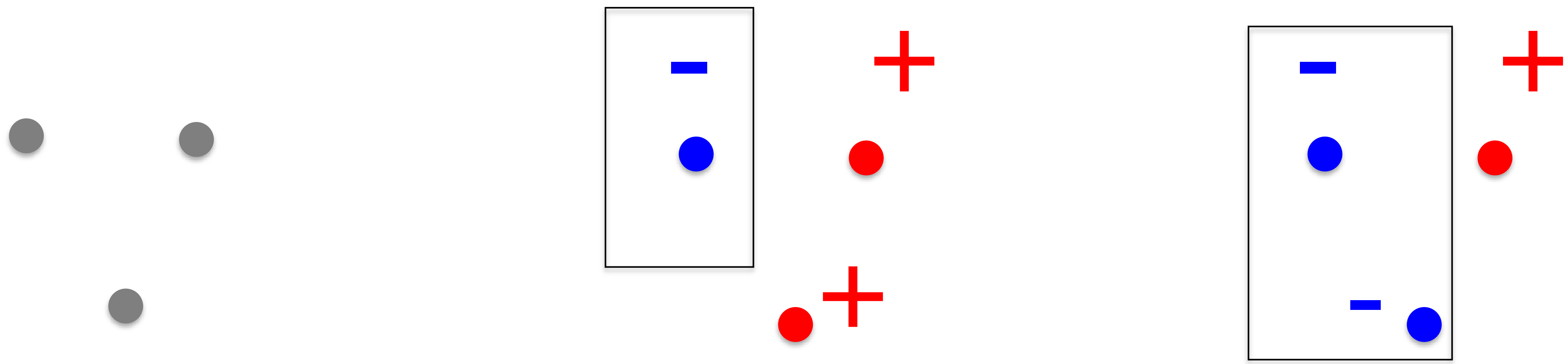


=> VC dim = 3

VC dim. of axis parallel rectangles in 2d

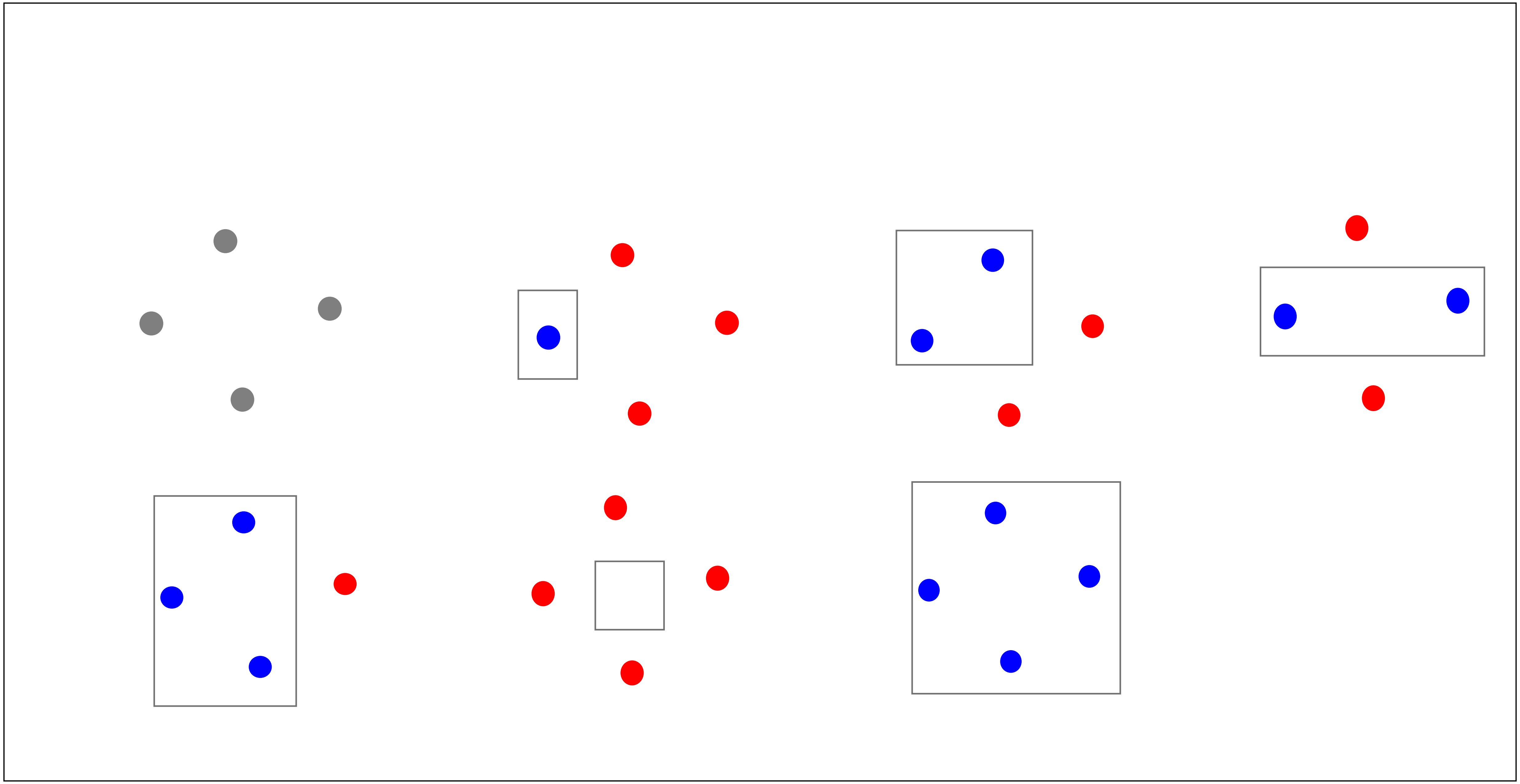
What's the VC dim. of axis parallel rectangles in 2d?

$$f(x) = \text{sign}(1 - 2 \cdot 1_{\{x \in \text{ rectangle}\}})$$



There is a placement of 3 pts that can be shattered => VC dim ≥ 3

VC dim. of axis parallel rectangles in 2d



There is a placement of 4 pts that can be shattered $\Rightarrow$ VC dim ≥ 4

VC dim. of axis parallel rectangles in 2d

What's the VC dim. of axis parallel rectangles in 2d?

$$f(x) = \text{sign}(1 - 2 \cdot 1_{\{x \in \text{ rectangle}\}})$$

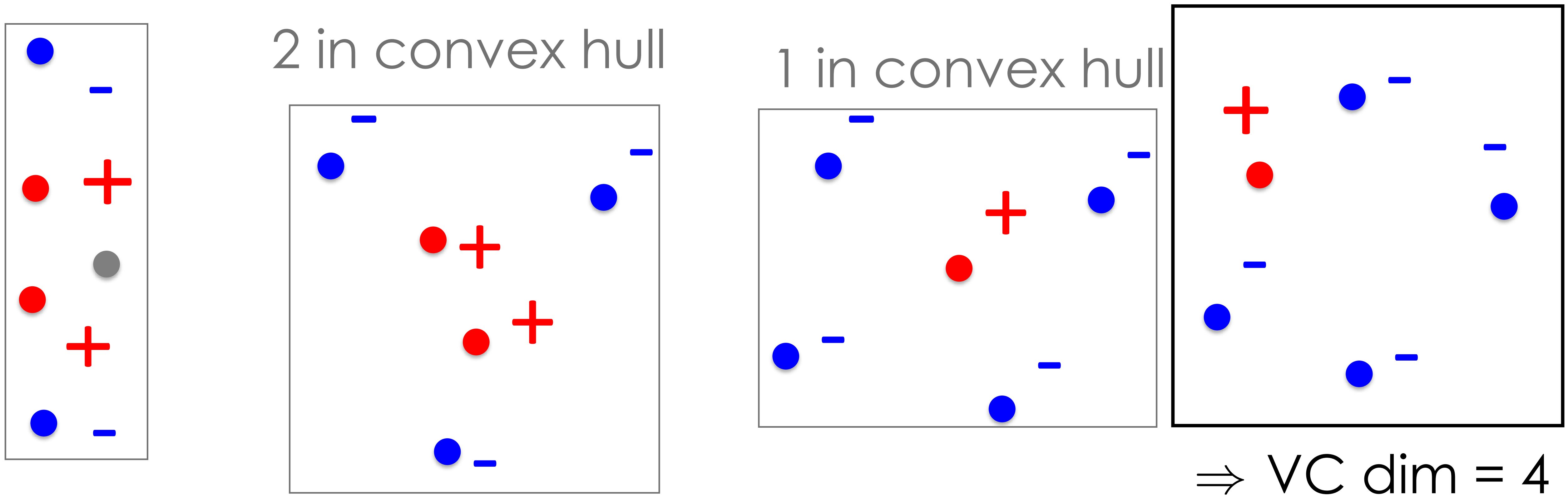
If VC dim = 4, then for all placements of 5 pts, there exists a labeling that can't be shattered

4 collinear

2 in convex hull

1 in convex hull

pentagon



$\Rightarrow$ VC dim = 4

Sauer's Lemma

We already know that $S_{\mathcal{F}}(n) \leq 2^n$ [Exponential in *n*]

Sauer's lemma:

$$S_{\mathcal{F}}(n) \leq \sum_{k=0}^{VC_{\mathcal{F}}} \binom{n}{k}$$

The VC dimension can be used to upper bound the shattering coefficient.

Corollary: $S_{\mathcal{F}(n)} \leq (n+1)^{VC_{\mathcal{F}}}$ [Polynomial in *n*]

$$S_{\mathcal{F}}(n) \leq \left(\frac{ne}{VC_{\mathcal{F}}}\right)^{VC_{\mathcal{F}}}$$

Vapnik-Chervonenkis inequality

When $|\mathcal{F}| = N < \infty$, we already know $\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|\right] \le \sqrt{\frac{\log(2N)}{2n}}$

Vapnik-Chervonenkis inequality: [We don't prove this]

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|\right] \le 2\sqrt{\frac{\log(2S_{\mathcal{F}}(n))}{n}}$$

From Sauer's lemma:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|\right] \le 2\sqrt{\frac{\log(2S_{\mathcal{F}}(n))}{n}} \le 2\sqrt{\frac{VC_{\mathcal{F}}\log(n+1) + \log 2}{n}}$$

Since $|R(f_n^*) - R(f_{\mathcal{F}}^*)| \le 2\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

Therefore, $\mathbb{E}[|R(f_n^*) - R(f_{\mathcal{F}}^*)|] \le 4\sqrt{\frac{VC_{\mathcal{F}}\log(n+1) + \log 2}{n}}$

Estimation error

Yayyy! ☺

Linear (hyperplane) classifiers

We already know that

$$\mathbb{E}[|R(f_n^*) - R(f_{\mathcal{F}}^*)|] \leq 4\sqrt{\frac{VC_{\mathcal{F}} \log(n+1) + \log 2}{n}}$$

Estimation error

For linear classifiers in dimension when $\mathcal{X} = \mathbb{R}^d$: $VC_{\mathcal{F}} = d + 1$.

$$\Rightarrow \mathbb{E}[|R(f_n^*) - R(f_{\mathcal{F}}^*)|] \leq 4\sqrt{\frac{(d+1) \log(n+1) + \log 2}{n}}$$

Estimation error

If we do feature map first, $x = \phi(x) \in \mathbb{R}^{d'}$, then linear separation (SVM) $\Rightarrow VC_{\mathcal{F}} = d' + 1$.

Estimation error

$$\Rightarrow \mathbb{E}[|R(f_n^*) - R(f_{\mathcal{F}}^*)|] \leq 4\sqrt{\frac{(d'+1) \log(n+1) + \log 2}{n}}$$

Vapnik-Chervonenkis Theorem

We already know from McDiarmid:

$$\Pr\left\{\left|\sup_{f\in\mathcal{F}}|\widehat{R}_n(f)-R(f)|-\mathbb{E}[\sup_{f\in\mathcal{F}}|\widehat{R}_n(f)-R(f)|]\right|\geq\varepsilon\right\}\leq 2\exp\left(-2\varepsilon^2 n\right)$$

Vapnik-Chervonenkis inequality: $\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\widehat{R}_n(f)-R(f)|\right]\leq 2\sqrt{\frac{\log(2S_{\mathcal{F}}(n))}{n}}$

Corollary: Vapnik-Chervonenkis theorem: [We don't prove them]

$$\Pr\left(\sup_{f\in\mathcal{F}}|\widehat{R}_n(f)-R(f)|>t\right)\leq 4S_{\mathcal{F}}(2n)\exp(-nt^2/8)$$

$$\Pr\left(\sup_{f\in\mathcal{F}}|\widehat{R}_n(f)-R(f)|>t\right)\leq 8S_{\mathcal{F}}(n)\exp(-nt^2/32)$$

We already know: Hoeffding + Union bound for finite function class:

When $|\mathcal{F}|=N<\infty$, $\Rightarrow \Pr\left(\sup_{f\in\mathcal{F}}|\widehat{R}_n(f)-R(f)|>t\right)\leq 2N\exp\left(-2nt^2\right)$

PAC Bound for the Estimation Error

VC theorem:

$$\Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| > t\right) \le 8S_{\mathcal{F}}(n)\exp(-nt^2/32)$$

Inversion:

$$8S_{\mathcal{F}}(n)\exp(-nt^2/32) \le \delta \qquad \Rightarrow t^2 \ge \frac{32}{n}\log\left(\frac{8S_{\mathcal{F}}(n)}{\delta}\right)$$

$$\Rightarrow \Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| \le 8\sqrt{\frac{\log(S_{\mathcal{F}}(n)) + \log\left(\frac{8}{\delta}\right)}{2n}}\right) \ge 1-\delta$$

$$S_{\mathcal{F}}(n) \le \left(\frac{ne}{VC_{\mathcal{F}}}\right)^{VC_{\mathcal{F}}} \qquad \Rightarrow \Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| \le 8\sqrt{\frac{VC_{\mathcal{F}}\log\left(\frac{ne}{VC_{\mathcal{F}}}\right) + \log\left(\frac{8}{\delta}\right)}{2n}}\right) \ge 1-\delta$$

Don't forget that $|R(f_n^*) - R(f_{\mathcal{F}}^*)| \le 2\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

Estimation error

$$\Rightarrow \Pr\left(|R(f_n^*) - R(f_{\mathcal{F}}^*)| \le 16\sqrt{\frac{\log(VC_{\mathcal{F}}\log\left(\frac{ne}{VC_{\mathcal{F}}}\right) + \log\left(\frac{8}{\delta}\right)}{2n}}\right) \ge 1-\delta$$

What you need to know

Complexity of the classifier depends on number of points that can be classified exactly

Finite case – Number of hypothesis

Infinite case – Shattering coefficient, VC dimension

Thanks for your attention ☺

Attic

Proof of Sauer's Lemma

Write all different behaviors on a sample $(x_1, x_2, \ldots x_n)$ in a matrix:

$f_i(x_j)$:

$|\mathcal{F}| = 7$

	x_1	x_2	x_3
f_1	0	0	0
f_2	0	1	0
f_3	1	1	1
f_4	1	0	0
f_5	0	1	0
f_6	1	1	1
f_7	0	1	1

$\rightarrow$

$|\mathcal{F}| = 7$

	x_1	x_2	x_3
f_1	0	0	0
f_2	0	1	0
f_3	1	1	1
f_4	1	0	0
f_7	0	1	1

VC dim =2

Proof of Sauer's Lemma

$$
|\mathcal{F}| = 7 \quad
\begin{array}{c|ccc|}
 & x_1 & x_2 & x_3 \\
\hline
f_1 & 0 & 0 & 0 \\
f_2 & 0 & 1 & 0 \\
f_3 & 1 & 1 & 1 \\
f_4 & 1 & 0 & 0 \\
f_7 & 0 & 1 & 1 \\
\hline
\end{array}
= A
$$

Shattered subsets of columns:

$$\{\emptyset\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}$$

We will prove that

$$S_{\mathcal{F}}(x_1, \ldots, x_n) = \#\ \text{rows}(A) \leq \#\ \text{shattered subsets of columns of } A \leq \sum_{k=0}^{VC_{\mathcal{F}}} \binom{n}{k}$$

Therefore,

$$S_{\mathcal{F}(n)} = \max_{x_1,\ldots,x_n} S_{\mathcal{F}}(x_1, \ldots, x_n) \leq \sum_{k=0}^{VC_{\mathcal{F}}} \binom{n}{k}$$

In this example: 5· 1+3+3=7, since VC=2, n=3

Proof of Sauer's Lemma

$$
\begin{array}{c|ccc}
|\mathcal{F}| = 7 & x_1 & x_2 & x_3 \\
\begin{array}{c} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_7 \end{array} & \multicolumn{3}{c}{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}}
\end{array} = A
$$

Shattered subsets of columns:

$$\{\emptyset\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}$$

Lemma 1 $\#$ shattered subsets of columns of $A \leq \sum_{k=0}^{VC_{\mathcal{F}}} \binom{n}{k}$

In this example: 6· 1+3+3=7

Lemma 2 $\#$ rows$(A) \leq \#$ shattered subsets of columns of A

for any binary matrix with no repeated rows.

In this example: 5· 6

Proof of Lemma 1

$$|\mathcal{F}| = 7 \quad \begin{array}{c|ccc} & x_1 & x_2 & x_3 \\ \hline f_1 & 0 & 0 & 0 \\ f_2 & 0 & 1 & 0 \\ f_3 & 1 & 1 & 1 \\ f_4 & 1 & 0 & 0 \\ f_7 & 0 & 1 & 1 \end{array} = A$$

Shattered subsets of columns:

$$\{\emptyset\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}$$

In this example: 6· 1+3+3=7

Lemma 1 # shattered subsets of columns of $A \leq \sum_{k=0}^{VC_{\mathcal{F}}} \binom{n}{k}$

Proof

$VC_{\mathcal{F}}$ is the size of largest imaginable shattered sample. $VC_{\mathcal{F}} = \max\{n : S_{\mathcal{F}}(n) = 2^n\}$

If a shattered subsets of columns has d elements, then $VC_{\mathcal{F}} \geq d$

For example if $\{x_1, x_3\}$ are shattered in A, then $VC_{\mathcal{F}} \geq 2$.

$$|\{\emptyset\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}| \leq |\{\emptyset\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}|$$

Q.E.D.

Proof of Lemma 2

Lemma 2 $\quad \#\ \mathsf{rows}(A) \leq \#\ \mathsf{shattered\ subsets\ of\ columns\ of}\ A$

for any binary matrix with no repeated rows.

Proof: Induction on the number of columns

Base case: A has one column. There are three cases:

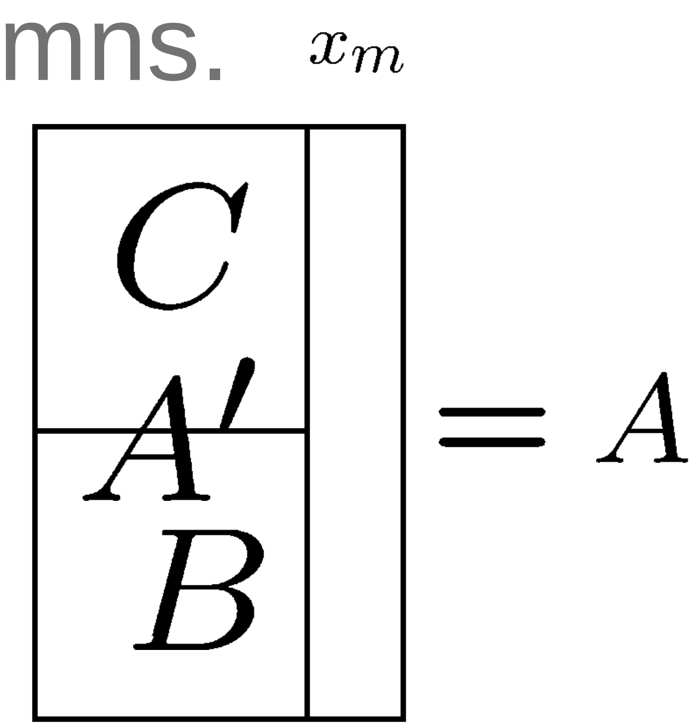
$A = (0)$ => $1 \cdot 1$ shattered subsets of columns: $\{\emptyset\}$

$A = (1)$ => $1 \cdot 1$ shattered subsets of columns: $\{\emptyset\}$

$A = \binom{0}{1}$ => $2 \cdot 2$ shattered subsets of columns: $\{\emptyset\}, \{x_1\}$

Proof of Lemma 2

Inductive case: A has at least two columns.



Let A' be A minus its last column x_m removed

In A' each row can occure once or twice.

If "twice" $\Rightarrow$ move one of them to B the other to C

If "once" $\Rightarrow$ move them to C

We have,

0	0	0
0	1	0
1	1	1
1	0	0
0	1	1

$$\# \text{ rows}(A) = \# \text{ rows}(B) + \# \text{ rows}(C)$$

$$\leq \# \text{ shattered subsets of columns of } (B)$$
$$+ \# \text{ shattered subsets of columns of } (C)$$

By induction (less columns)

Proof of Lemma 2

$\{\emptyset\}$ $\quad\quad$ $\{\emptyset\}, \{x_1\}, \{x_2\}\{x_1, x_2\}$

$\#$ shattered subsets of columns of (B) + $\#$ shattered subsets of columns of (C)

$\leq$ $\#$ shattered subsets of columns of (A)

$\{\emptyset\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}$

because

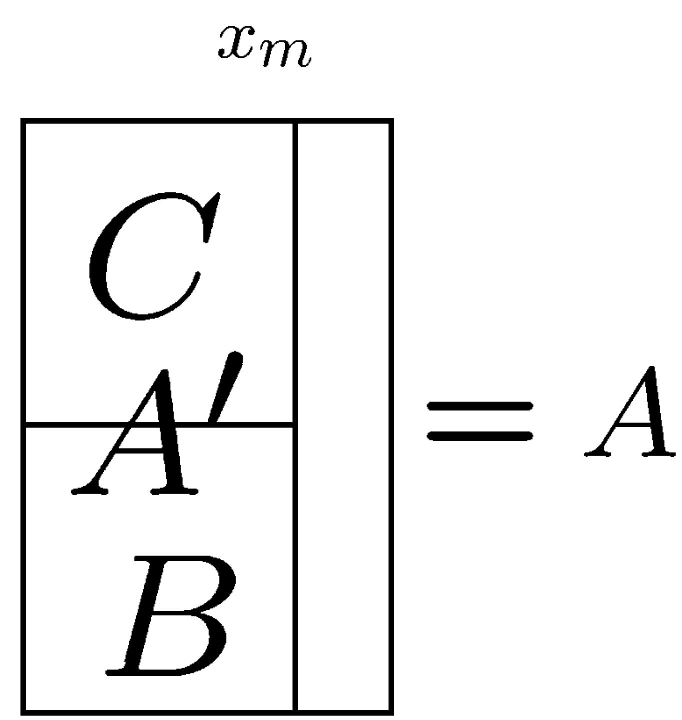
"once" $\Rightarrow$ move them to C

Therefore, if C shatters S e.g. $\{x_1, x_2\}$, then A shatters S.

"twice" $\Rightarrow$ move one of them to B the other to C

Therefore, if B shatters S, then A shatters $S \cup x_m$.

Q.E.D.



0	0	0
0	1	0
1	1	1
1	0	0
0	1	1

Solution to Overfitting

$$R^*_{\mathcal{F}} = \inf_{f \in \mathcal{F}} R(f) = \inf_{f \in \mathcal{F}} \mathbb{E}[L(Y, f(X))]$$

ERM on $\mathcal{F}$: $$\widehat{R}^*_{n,\mathcal{F}} = \inf_{f \in \mathcal{F}} \widehat{R}_n(f) = \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i))$$

2nd issue: $$\widehat{R}^*_{n,\mathcal{F}} = \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i))$$

$\inf_{f \in \mathcal{F}} \widehat{R}_n(f)$ might be a very difficult optimization problem in f

It might be not even convex in f

Solution:

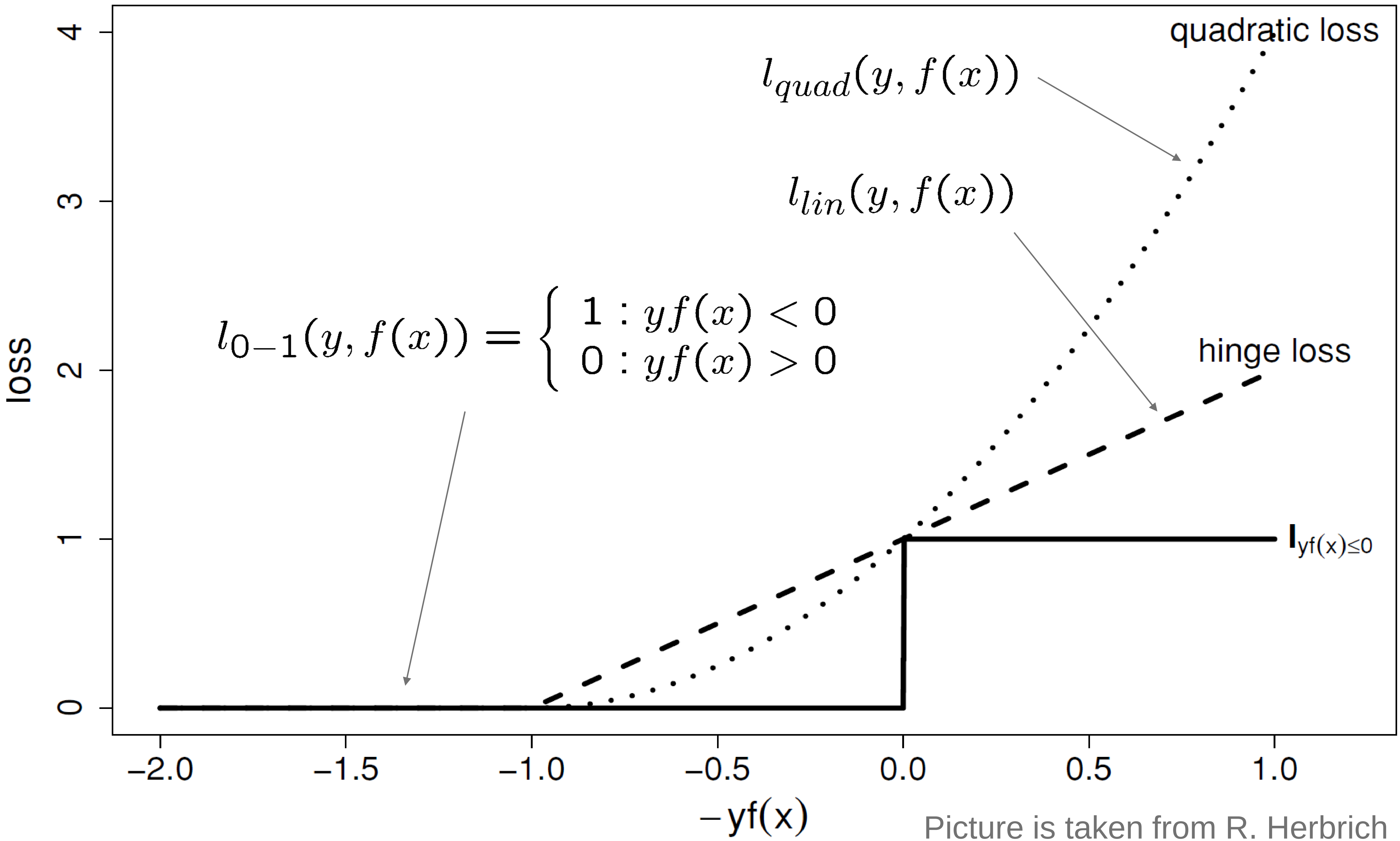
Choose loss function L such that $\widehat{R}_n(f)$ will be convex in f

$$L(y, f(x)) = \begin{cases} 1 & y \neq f(x) \\ 0 & y = f(x) \end{cases} \Rightarrow \text{not convex } \widehat{R}_n(f)$$

Hinge loss $\Rightarrow$ convex $\widehat{R}_n(f)$

Quadratic loss $\Rightarrow$ convex $\widehat{R}_n(f)$

Approximation with the Hinge loss and quadratic loss

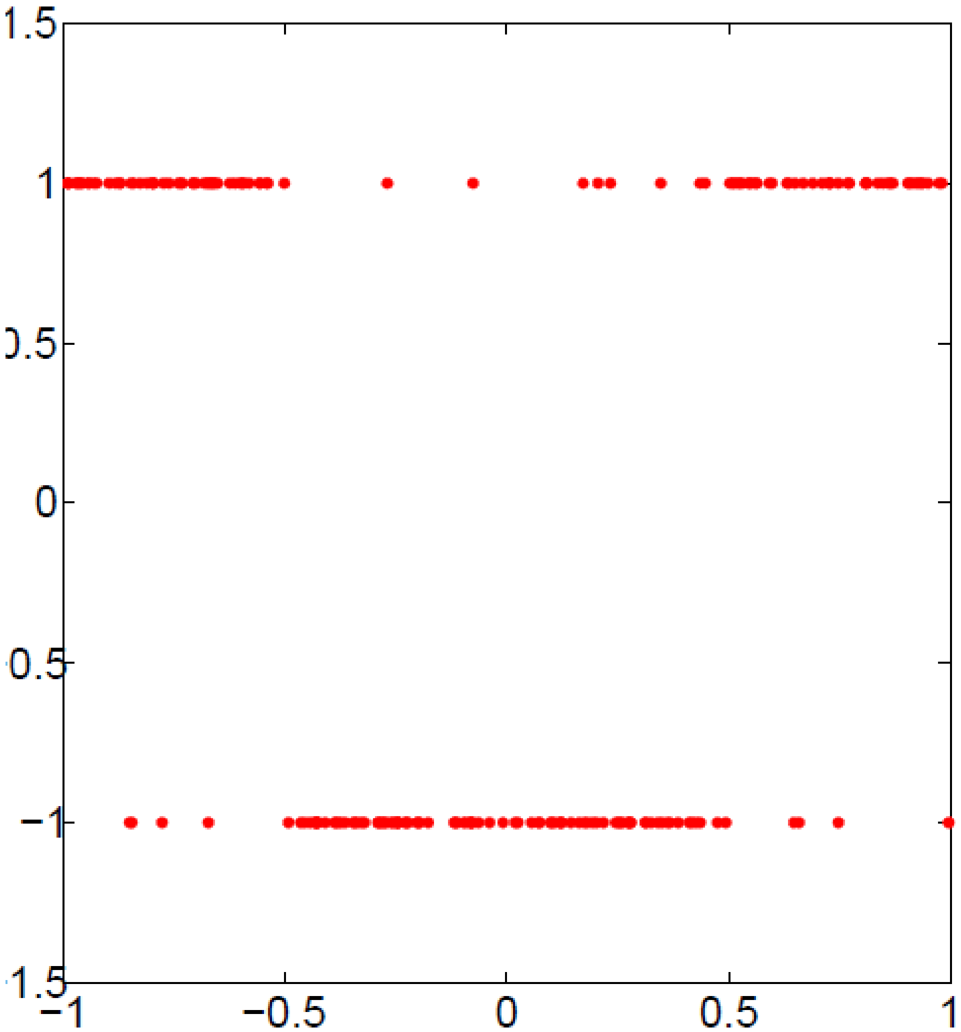


Picture is taken from R. Herbrich

Underfitting

Let $\mathcal{F}$ be the class of thresholded polynomials of degree at most one.

$$\mathcal{F} = \{f : f(x) = \text{sign}(ax + b), a, b \in \mathbb{R}\}$$



$$X \sim U[-1, 1]$$

$$\Pr(Y = +1 | X \in (-0.5, 0.5)) = 0.9$$
$$\Pr(Y = -1 | X \in (-0.5, 0.5)) = 0.1$$

$$\Pr(Y = +1 | X \notin (-0.5, 0.5)) = 0.1$$
$$\Pr(Y = -1 | X \notin (-0.5, 0.5)) = 0.9$$

$$f^*(x) = \begin{cases} 1 & \text{if } x \notin (-0.5, 0.5) \\ -1 & \text{if } x \in (-0.5, 0.5) \end{cases}$$

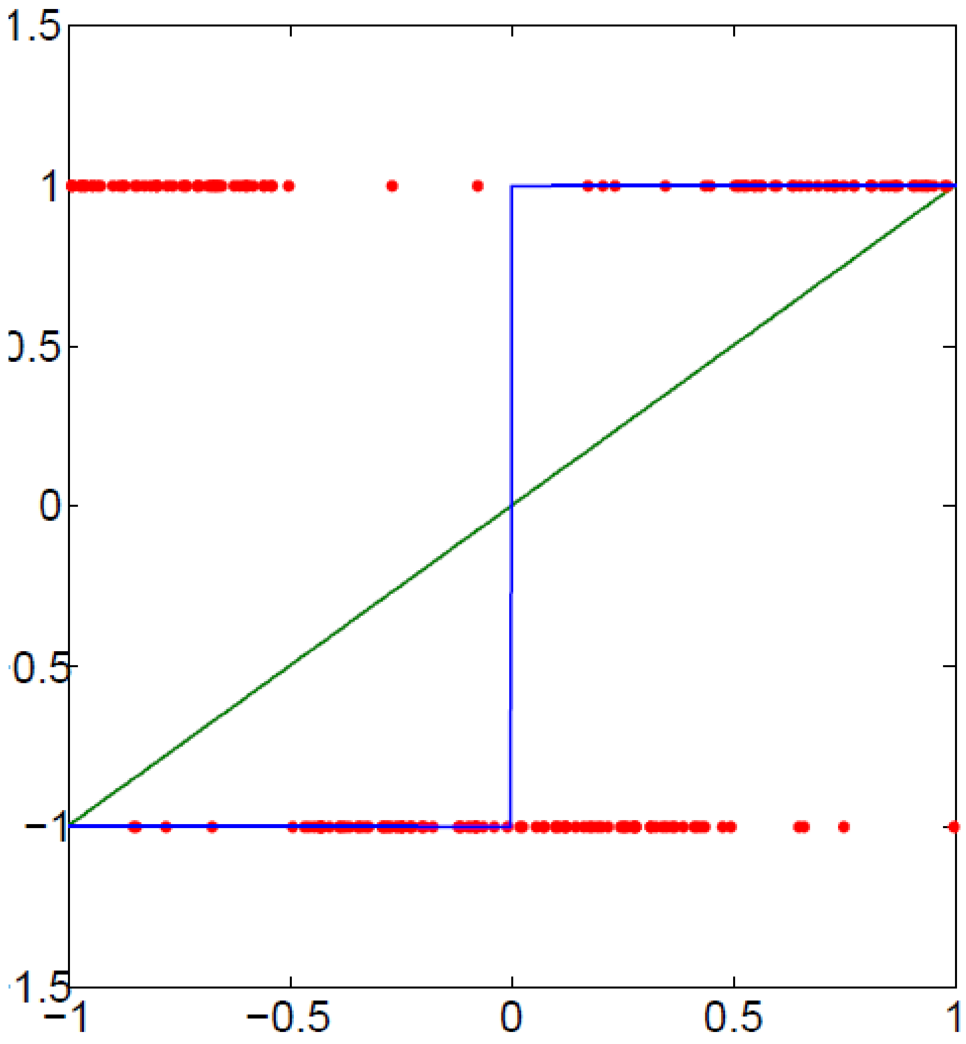
$$R^*_{\mathcal{F}} = \inf_{f \in \mathcal{F}} R(f) = \inf_{f \in \mathcal{F}} \mathbb{E}[L(Y, f(X))]$$

Bayes risk = 0.1

Underfitting

$$\mathcal{F} = \{f : f(x) = \text{sign}(ax + b), a, b \in \mathbb{R}\}$$

Best linear classifier:



$$R^*_{\mathcal{F}} = R(f^*_{\mathcal{F}}) = \inf_{f \in \mathcal{F}} \Pr[Y \neq f(X)]$$

$$= \frac{1}{4} \times 0.9 + \frac{1}{4} \times 0.1 + \frac{1}{4} \times 0.9 + \frac{1}{4} \times 0.1 = 0.5$$

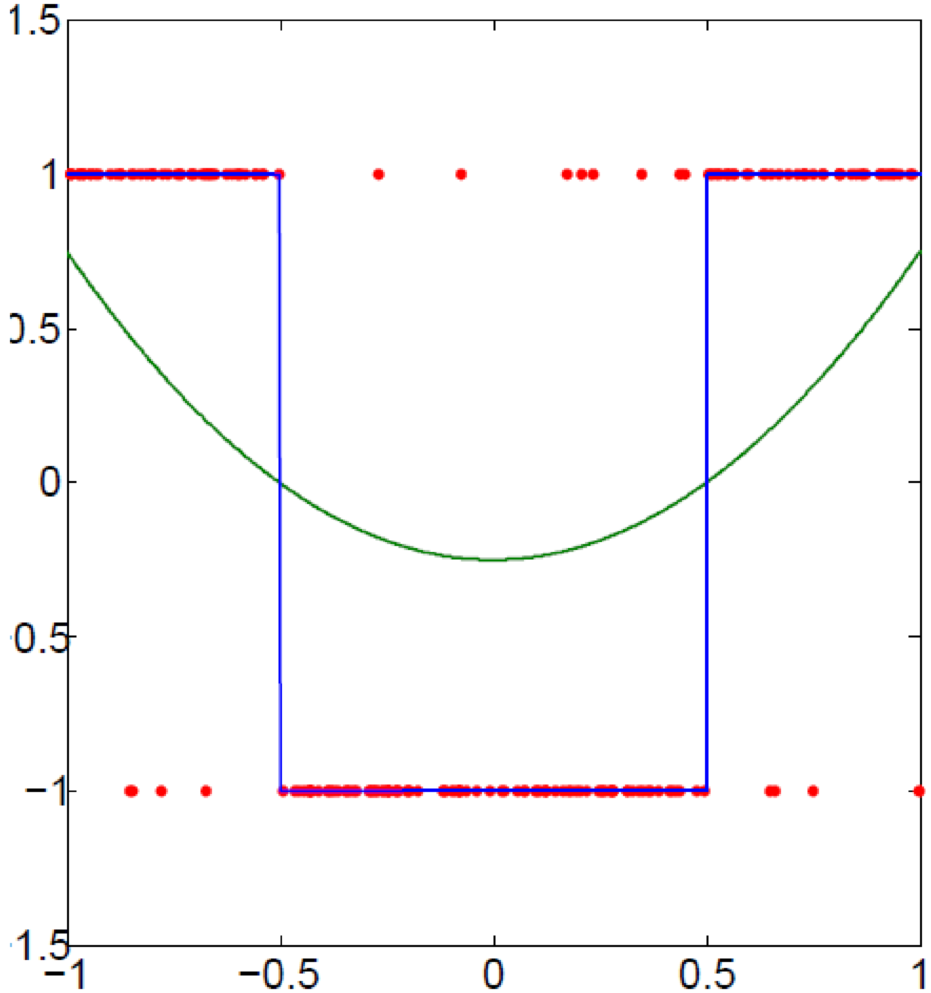
The empirical risk of the best linear classifier:

$$\widehat{R}_n(f^*_{\mathcal{F}}) \approx 0.5$$

Underfitting

$$\mathcal{F} = \{f : f(x) = \text{sign}(ax^2 + bx + c), a, b, c \in \mathbb{R}\}$$

Best quadratic classifier:



$$f^*_{\mathcal{F}} = \text{sign}((x - 0.5)(x + 0.5))$$

$$R^*_{\mathcal{F}} = R(f^*_{\mathcal{F}}) = \inf_{f \in \mathcal{F}} \Pr[Y \neq f(X)]$$
$$= \frac{1}{4} \times 0.1 + \frac{1}{4} \times 0.1 + \frac{1}{4} \times 0.1 + \frac{1}{4} \times 0.1 = 0.1$$

Same as the Bayes risk) good fit!

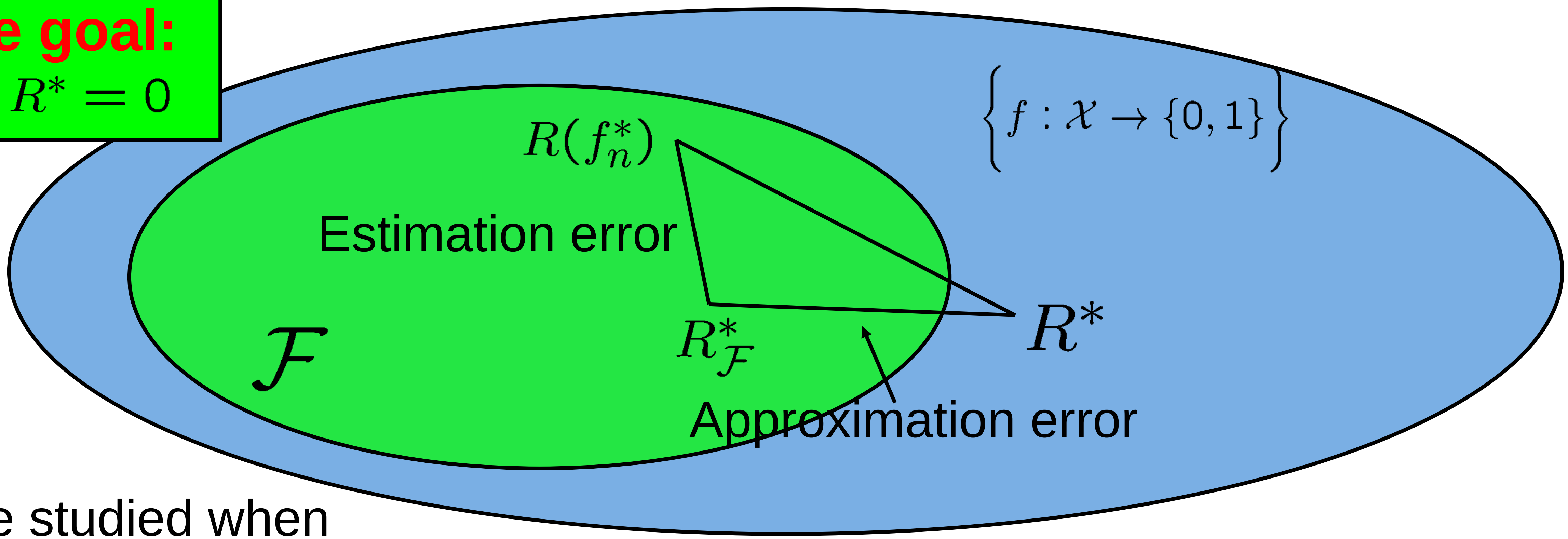
Structural Risk Minimization

Risk of the classifier f_n^*

$$R(f_n^*) - R^* = \overbrace{R(f_n^*) - R^*_{\mathcal{F}}}^{\text{Estimation error}} + \overbrace{R^*_{\mathcal{F}} - R^*}^{\text{Approximation error}}$$

Bayes risk

Ultimate goal:
$R(f_n^*) - R^* = 0$



So far we studied when
estimation error ! 0, but we also want approximation error ! 0

Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F}_n \subseteq \ldots$ such that $VC_{\mathcal{F}_1} \leq VC_{\mathcal{F}_2} \leq \ldots \leq VC_{\mathcal{F}_n} \leq \ldots$

Many different variants…
penalize too complex models to avoid overfitting