

# Introduction to Machine Learning

## Risk Minimization

Barnabás Póczos

# What have we seen so far?

**Several classification & regression algorithms seem to work fine on training datasets:**

- Linear regression
- Gaussian Processes
- Naïve Bayes classifier
- Support Vector Machines

- ☐ How good are these algorithms on unknown test sets?
- ☐ How many training samples do we need to achieve small error?
- ☐ What is the smallest possible error we can achieve?

**=> Learning Theory**

# Outline

- Risk and loss
  - Loss functions
  - Risk
  - Empirical risk vs True risk
  - Empirical Risk minimization
- Underfitting and Overfitting
- Classification
- Regression

# Supervised Learning Setup

$\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$  training data

$\{(X_{n+1}, Y_{n+1}), \dots, (X_m, Y_m)\}$  test data

Features:  $X \in \mathcal{X} \subset \mathbb{R}^d$

Labels:  $Y \in \mathcal{Y} \subset \mathbb{R}$

**Generative model of the data:**  $X \sim \mu, \mu(A) = \Pr(X \in A)$   
(train and test data)  $Y \sim p(\cdot|X)$

**Regression:** Labels:  $\mathcal{Y} = [a, b] \subset \mathbb{R}$ , or  $\mathcal{Y} = \mathbb{R}$

**Classification:** Labels:  $\mathcal{Y} = \{0, 1\}$

# Loss

$\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$  training data

$\{(X_{n+1}, Y_{n+1}), \dots, (X_m, Y_m)\}$  test data

Loss function:  $L(x, y, f(x))$

where  $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow [0, \infty]$

It measures how good we are on a particular (x,y) pair.

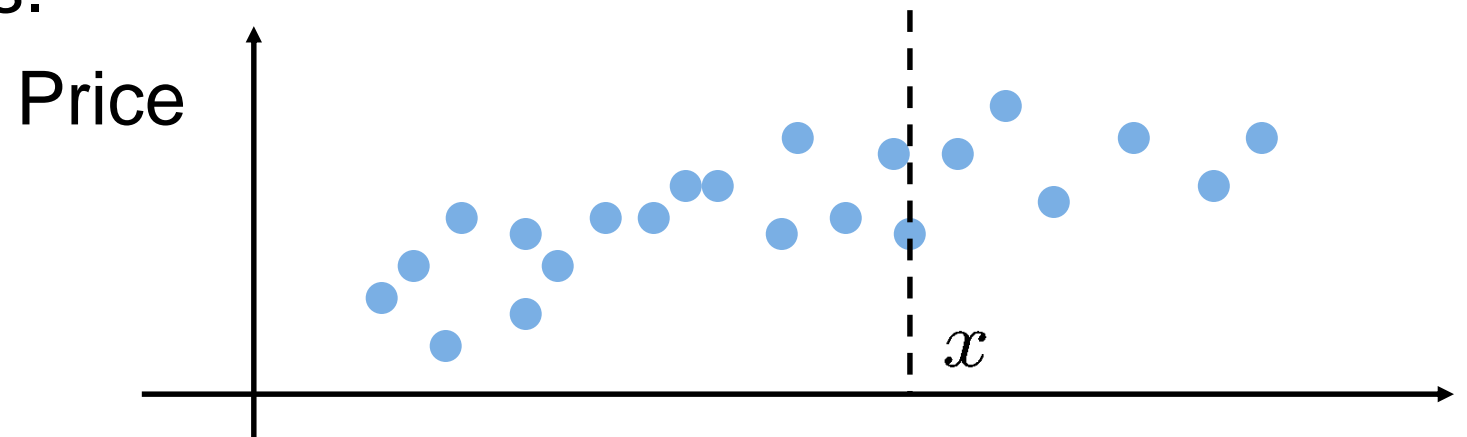
We want the loss  $L(X_t, Y_t, f(X_t))$  to be small for many  $(X_t, Y_t)$  pairs in the test data.

# Loss Examples

Classification loss:

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

**Regression:** Predict house prices.

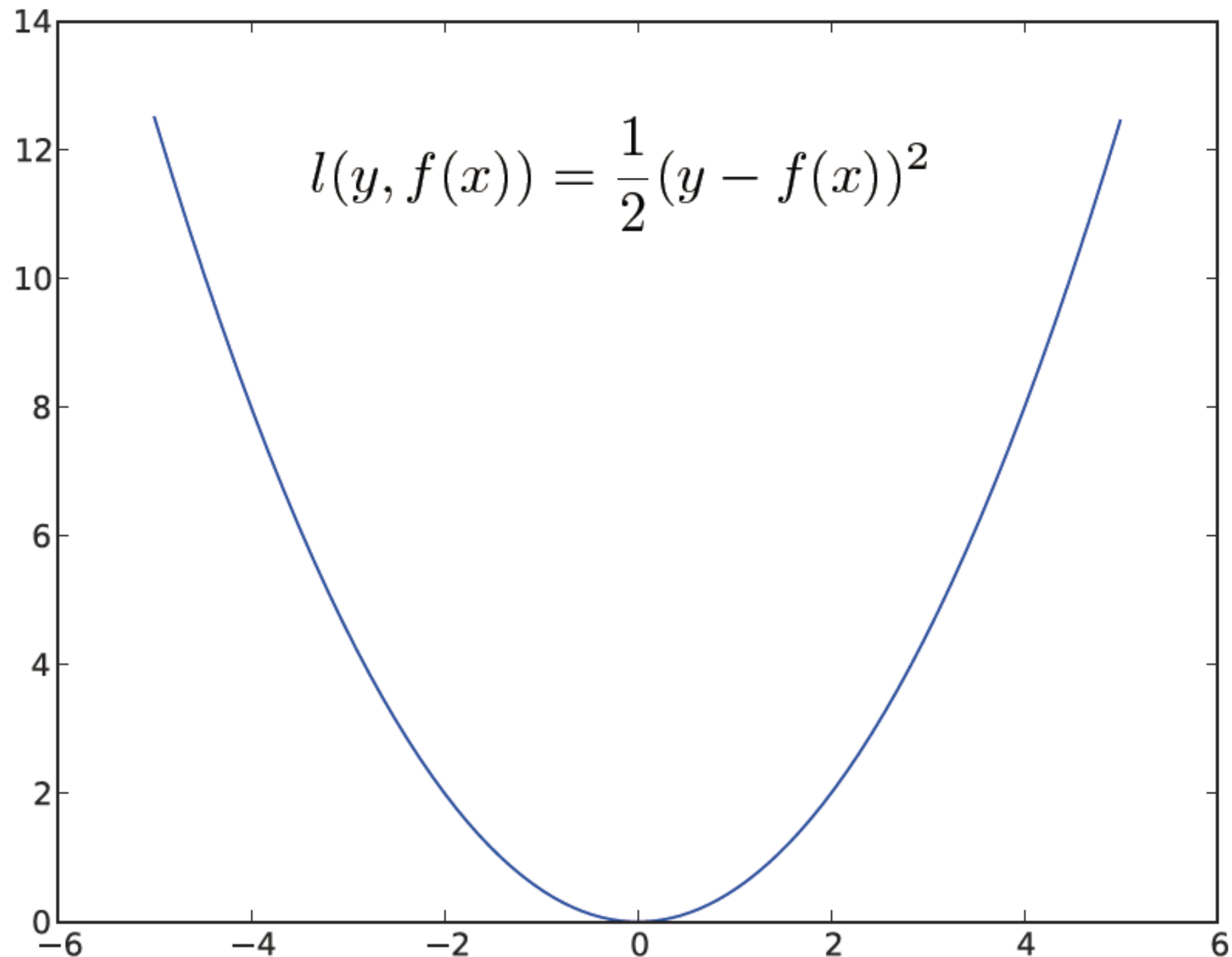


The price of house with feature  $x$  is  $p(\cdot|x)$

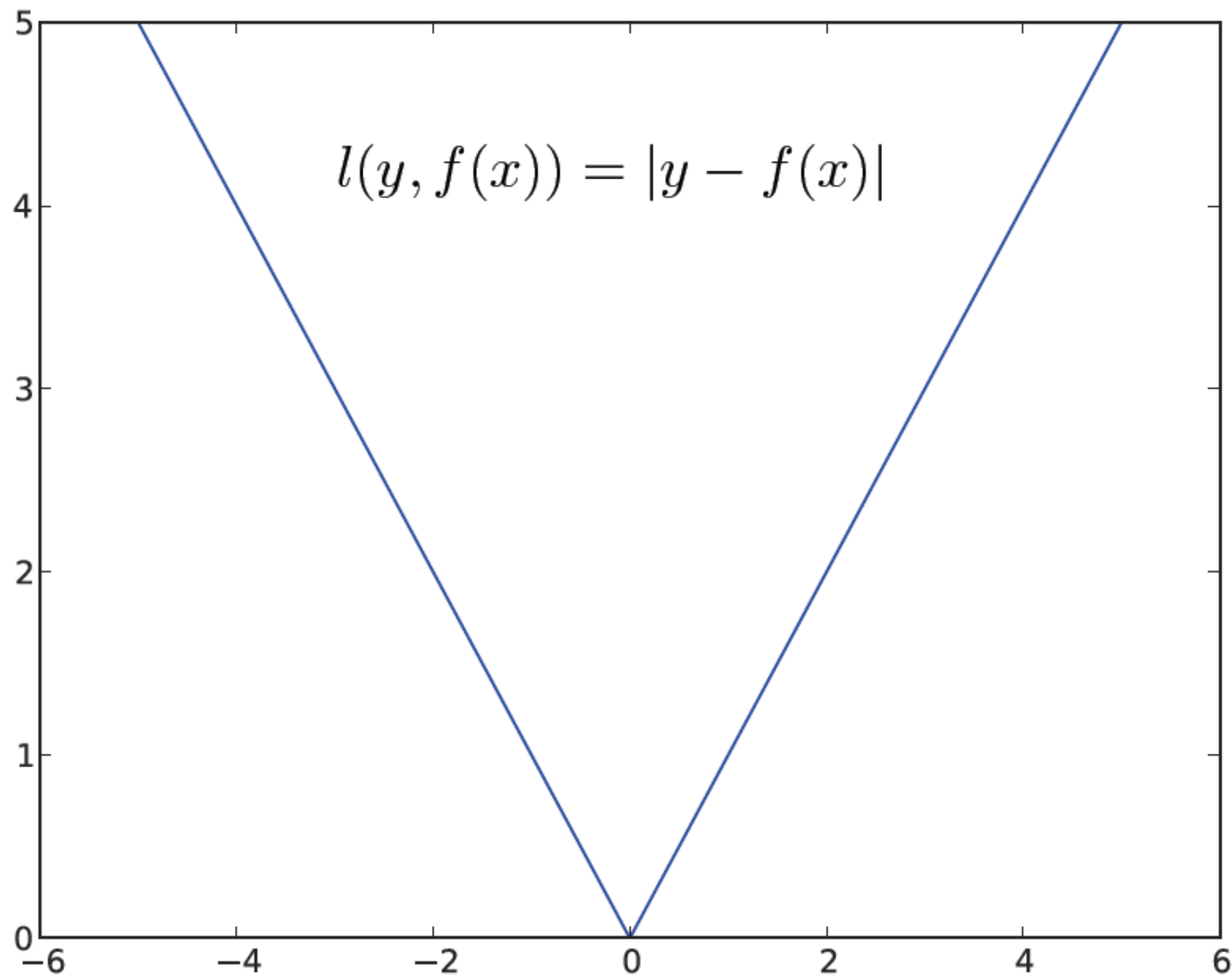
**$L_2$  loss for regression:**  $L(x, y, f(x)) = (y - f(x))^2$

**$L_1$  loss for regression:**  $L(x, y, f(x)) = |y - f(x)|$

# Squared loss, $L_2$ loss

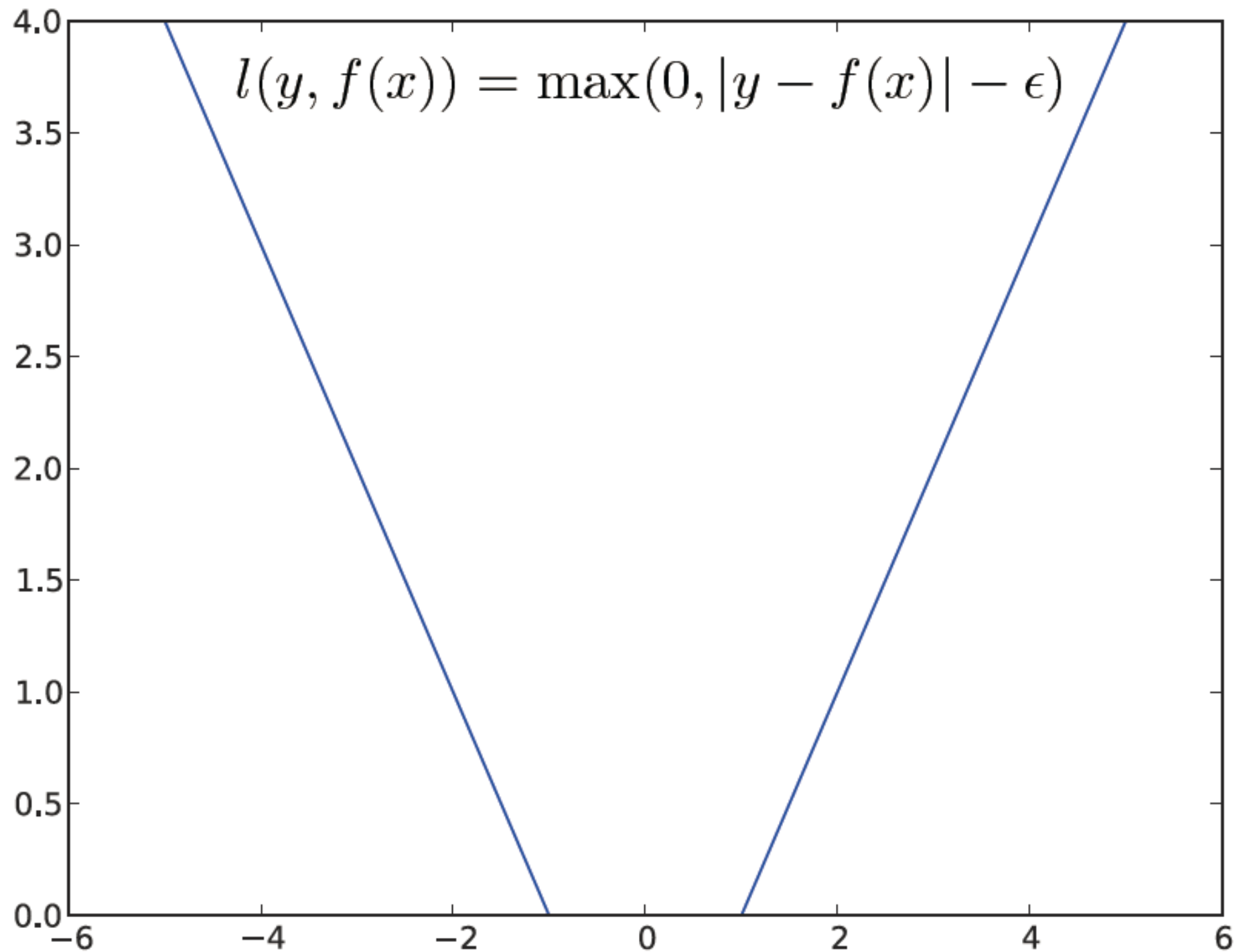


# $L_1$ loss



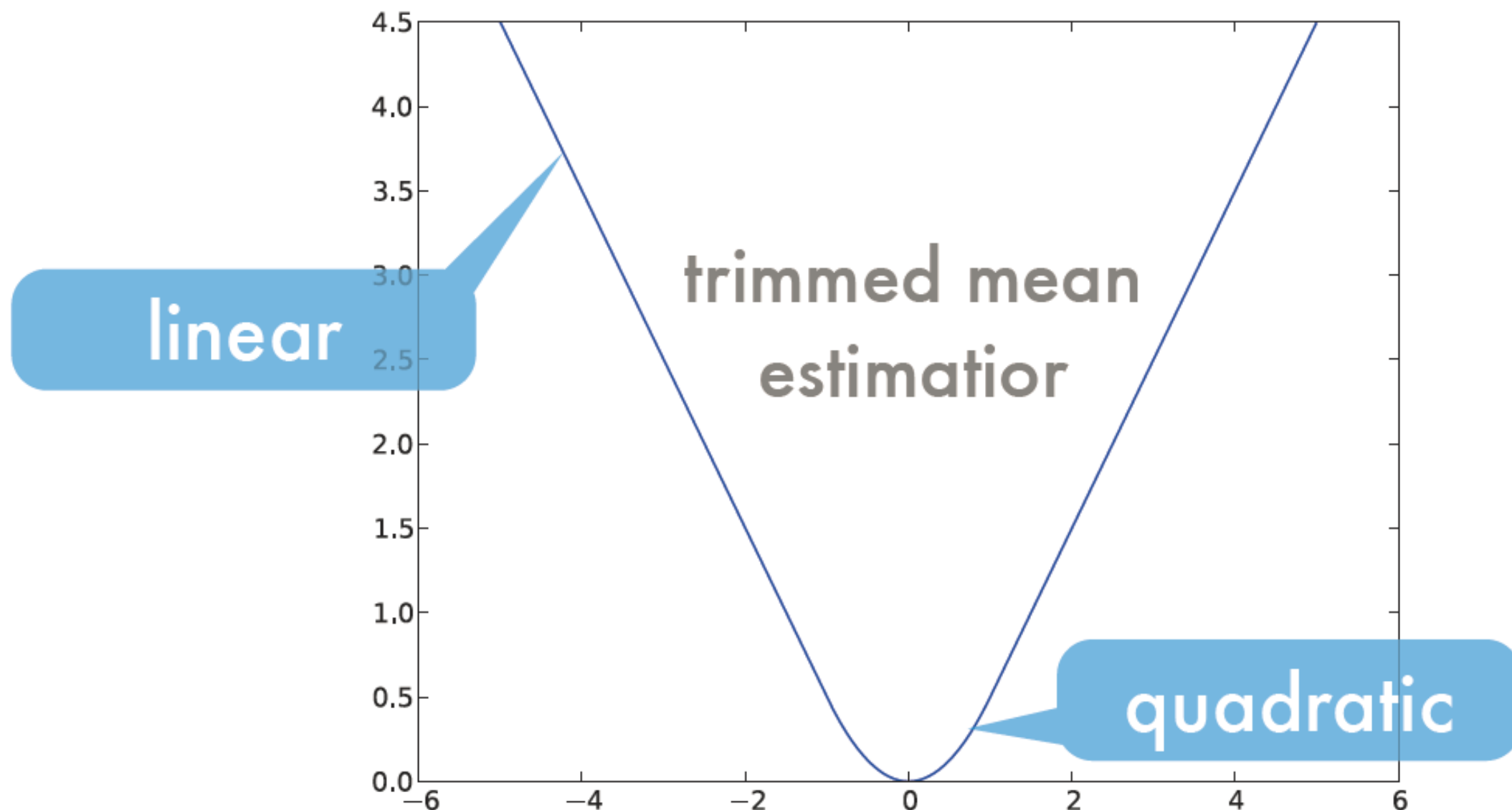


# $\epsilon$ -insensitive loss



# Huber's robust loss

$$l(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{if } |y - f(x)| < 1 \\ |y - f(x)| - \frac{1}{2} & \text{otherwise} \end{cases}$$




# Risk

Risk of  $f$  classification/regression function:

$$R_{L,P}(f) = \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) dP(x, y) = \text{The expected loss}$$
$$= \mathbb{E}[L(X, Y, f(X))]$$

$p(y, x) dy dx$



$L(x, y, f(x))$ : Loss function

$P(x, y)$ : Distribution of the data.

Why do we care about this?

# Why do we care about risk?

Risk of  $f$  classification/regression function:

$$\begin{aligned} R_{L,P}(f) &= \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) dP(x, y) && = \text{The expected loss} \\ &= \mathbb{E}[L(X, Y, f(X))] \end{aligned}$$

$\nwarrow$   
 $p(y, x) dy dx$

Our true goal is to minimize the loss of the test points!

$$f^* = \arg \min_f \frac{1}{m - n} \sum_{i=n+1}^m L(X_i, Y_i, f(X_i))$$

Usually we don't know the test points and their labels in advance..., but

$$\frac{1}{m - n} \sum_{i=n+1}^m L(X_i, Y_i, f(X_i)) \xrightarrow{m \rightarrow \infty} R_{L,P}(f) \quad (\text{LLN})$$

That is why our goal is to minimize the risk.

# Risk Examples

**Risk:**  $R_{L,P}(f) = \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) dP(x, y)$       The expected loss

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

**Risk of classification loss:**

$$R_{L,P}(f) = \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) dP(x, y) = \mathbb{E}[\mathbf{1}_{\{f(X) \neq Y\}}] = \Pr(f(X) \neq Y)$$

**L<sub>2</sub> loss for regression:**  $L(x, y, f(x)) = (y - f(x))^2$

**Risk of L<sub>2</sub> loss:**  $R_{L,P}(f) = \mathbb{E}[(Y - f(X))^2]$

# Bayes Risk

$$R_{L,P}(f) = \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) dP(x, y) \quad \text{The expected loss}$$

## Definition: Bayes Risk

$$R_{L,P}^* = \inf_{f: \mathcal{X} \rightarrow \mathbb{R}} \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) dP(x, y)$$

↙  
We consider all possible function  $f$  here

We don't know  $P$ , but we have i.i.d. training data sampled from  $P$ !

## Goal of Learning:

Build a function  $f_D$  (using data  $D$ ) whose risk  $R_{L,P}(f_D)$  will be close to the Bayes risk  $R_{L,P}^*$

The learning algorithm constructs this function  $f_D$  from the training data.

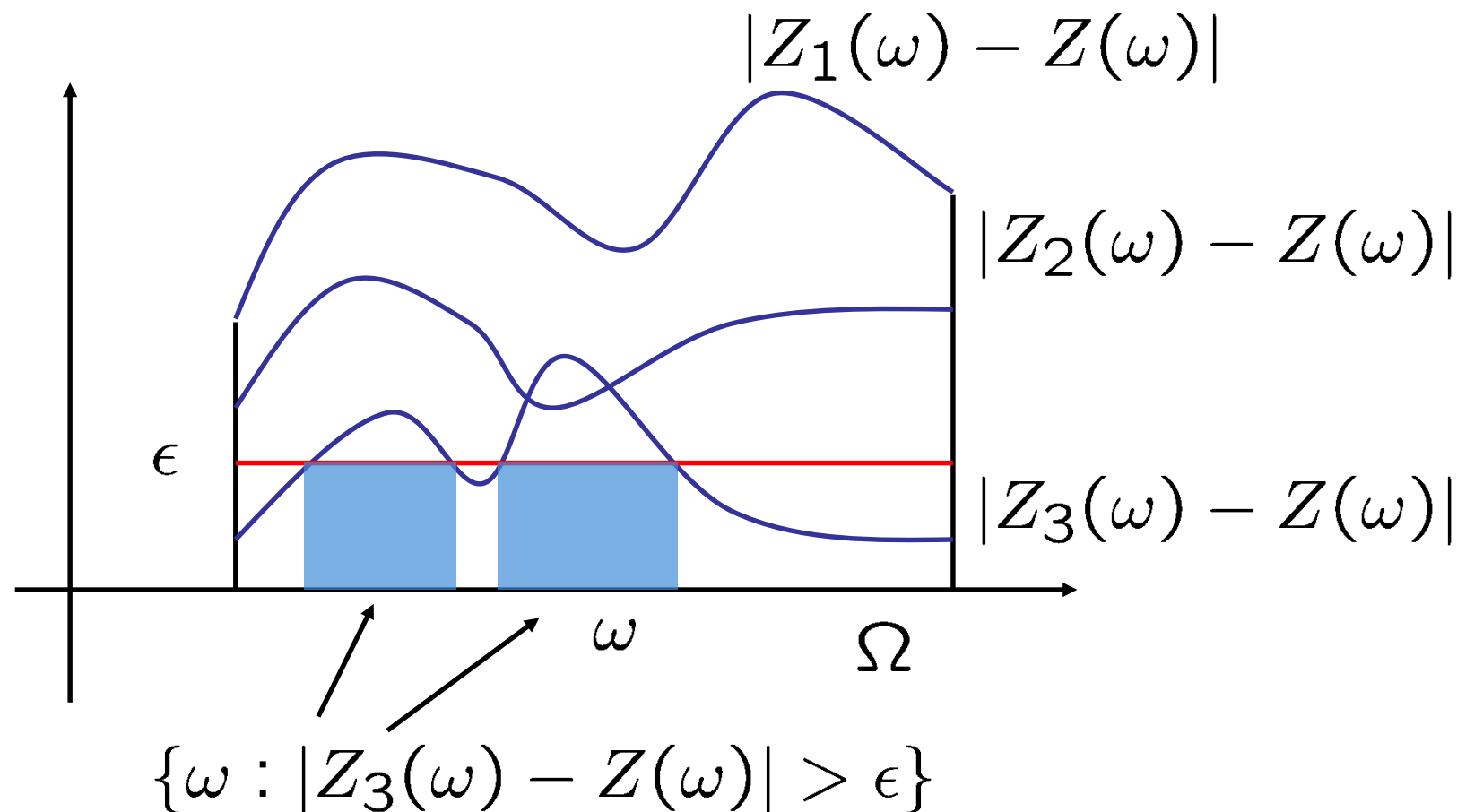
# Convergence in Probability

# Convergence in Probability

Notation:  $Z_n \xrightarrow{p} Z$

Definition:  $\forall \varepsilon > 0 \lim_{n \rightarrow \infty} \Pr(|Z_n - Z| \geq \varepsilon) = 0.$

$\forall \varepsilon > 0 \lim_{n \rightarrow \infty} \Pr(|Z_n - Z| < \varepsilon) = 1.$



This indeed measures how far the values of  $Z_n(\omega)$  and  $Z(\omega)$  are from each other.



# Consistency of learning methods

Risk is a random variable:  $R_{L,P}(f_D) = \mathbb{E}[L(X, Y, f_D(X)|D)]$

## Definition:

A learning method is **universally consistent** if **for all**  $P(X, Y)$  distributions the risk converges to the Bayes risk when we increase the sample size

$$R_{L,P}(f_D) \xrightarrow{p} R_{L,P}^* \text{ as } n \rightarrow \infty.$$

**Stone's theorem 1977:** Many classification, regression algorithms are universally consistent for certain loss functions under certain conditions: kNN, Parzen kernel regression, SVM,...

Yayyy!!! 😊

Wait! This doesn't tell us anything about the rates...

# No Free Lunch!

**Devroy 1982:** For every consistent learning method and for every fixed convergence rate  $a_n$ , there exists  $P(X,Y)$  distribution such that the convergence rate of this learning method on  $P(X,Y)$  distributed data is slower than  $a_n$

$$R_{L,P}(f_D) \xrightarrow{p} R_{L,P}^* \text{ as } n \rightarrow \infty \text{ with slower rate than } a_n$$



What can we do now?

# 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:*  $R^* = R_{L,P}^* = \inf_{f:\mathcal{X}\rightarrow\mathbb{R}} R(f)$

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

Let us use the empirical counter part:

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

# Empirical Risk Minimization

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

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

## Law of Large Numbers:

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

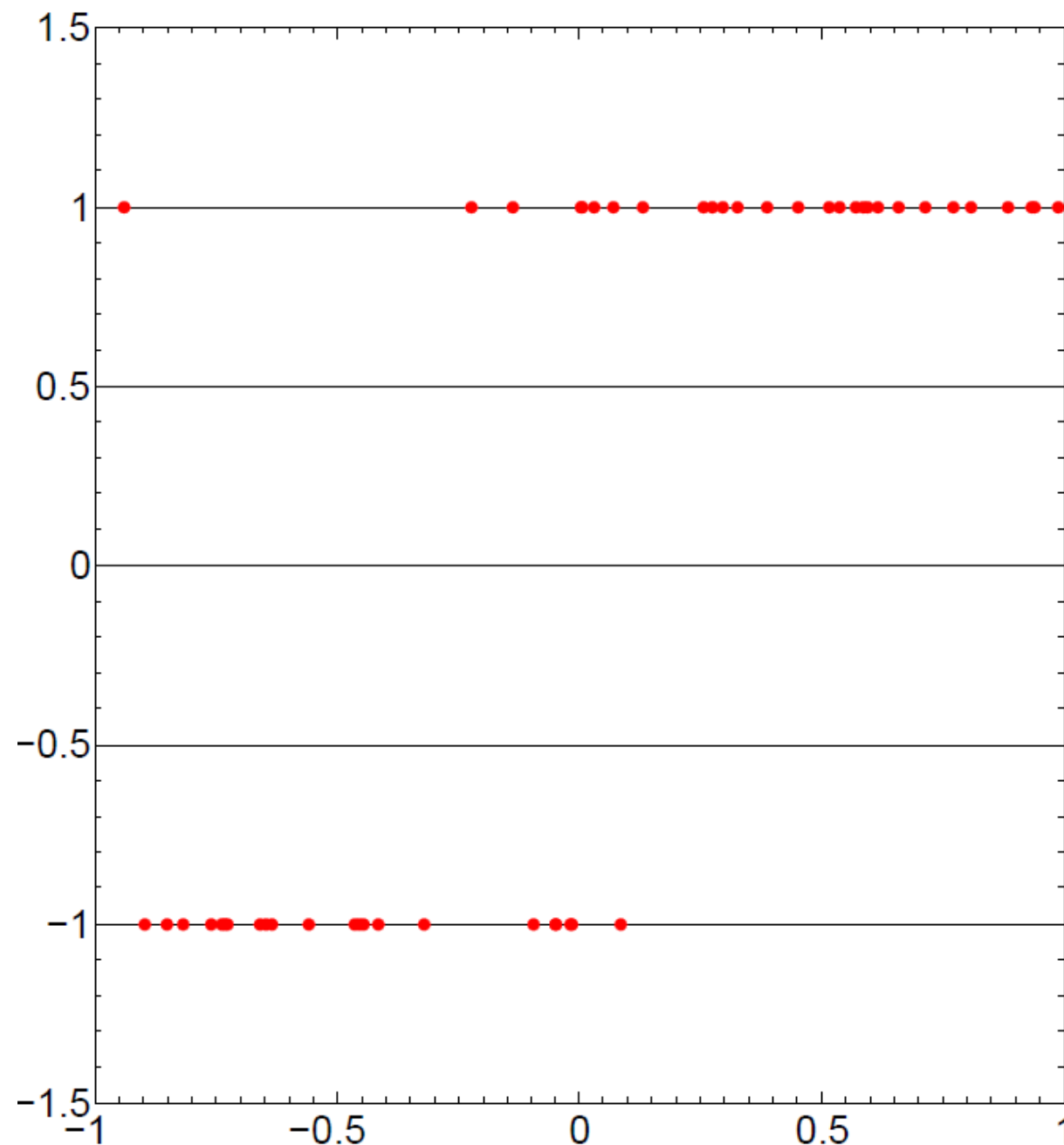
Empirical risk is converging to the Bayes risk

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

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

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

# Overfitting in Classification with ERM



Picture from David Pal

Generative model:

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

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

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

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

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

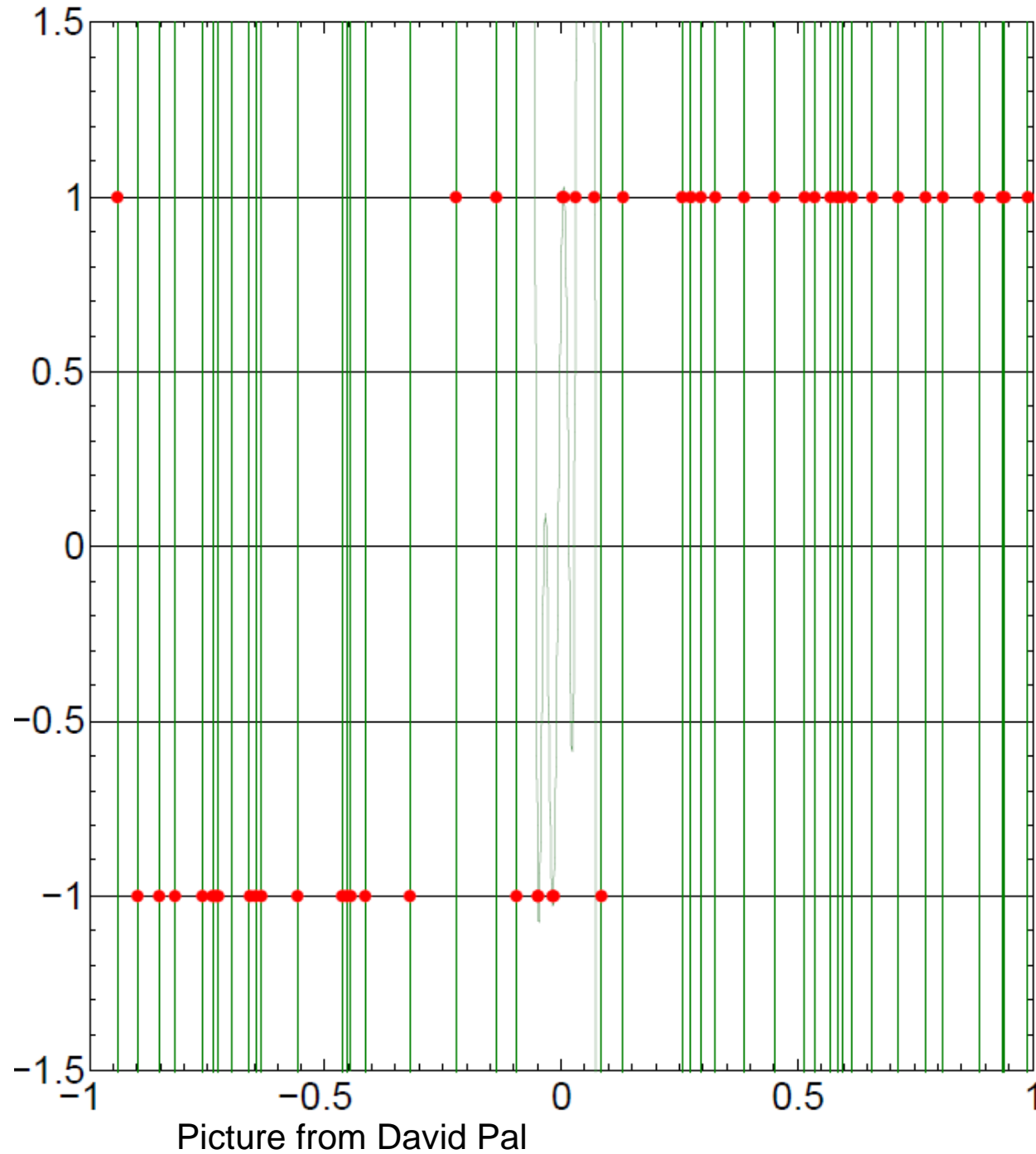
Bayes classifier:

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

Bayes risk:

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

# Overfitting in Classification with ERM



n-order thresholded polynomials

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

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

Empirical risk:

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

True risk of  $f_n^* = 0.5$

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

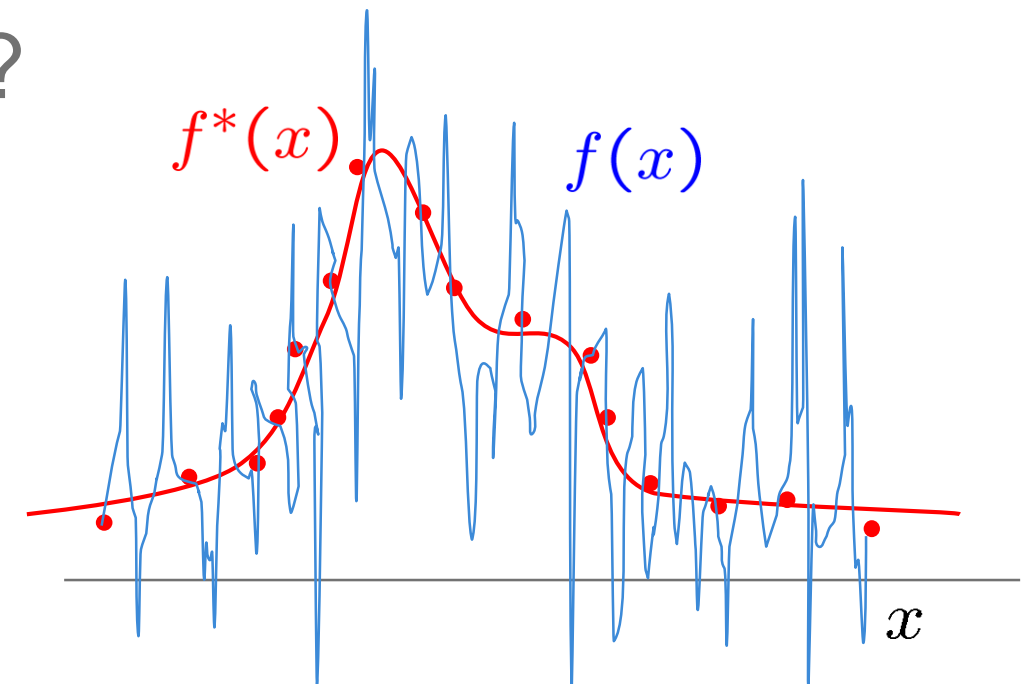
Bayes risk:

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

# Overfitting in Regression with ERM

Is the following predictor a good one?

$$f(x) = \begin{cases} Y_i, & x = X_i \text{ for } i = 1, \dots, n \\ \text{any value,} & \text{otherwise} \end{cases}$$



What is its empirical risk? (performance on training data)

zero !

What about true risk?

> zero

Will predict very poorly on new random test point:

Large generalization error !

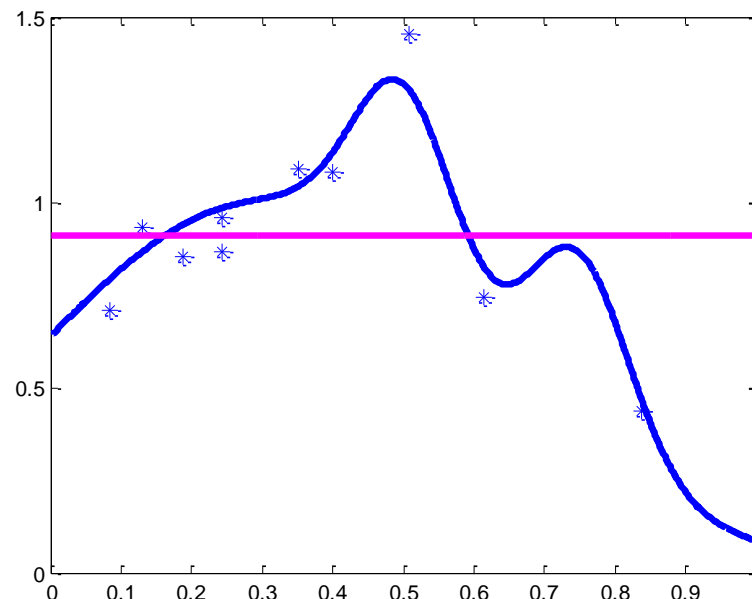


# Overfitting in Regression

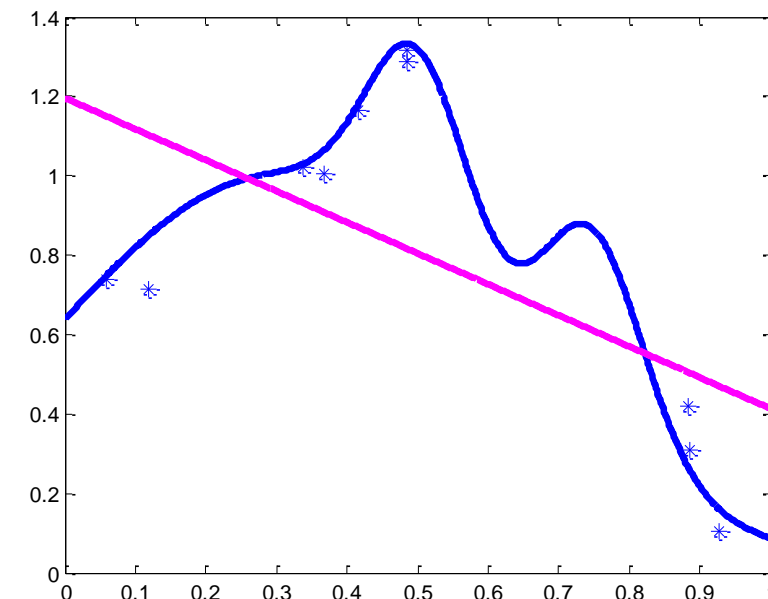
If we allow very complicated predictors, we could overfit the training data.

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

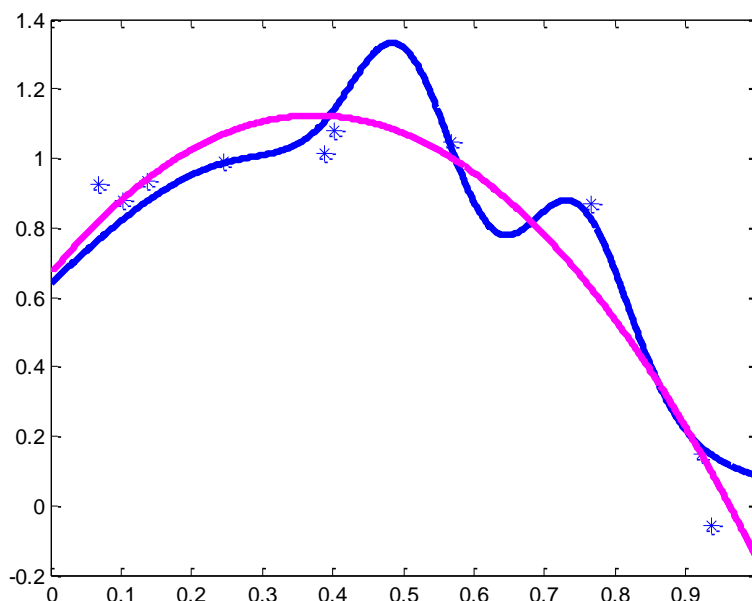
$k=1$   
constant



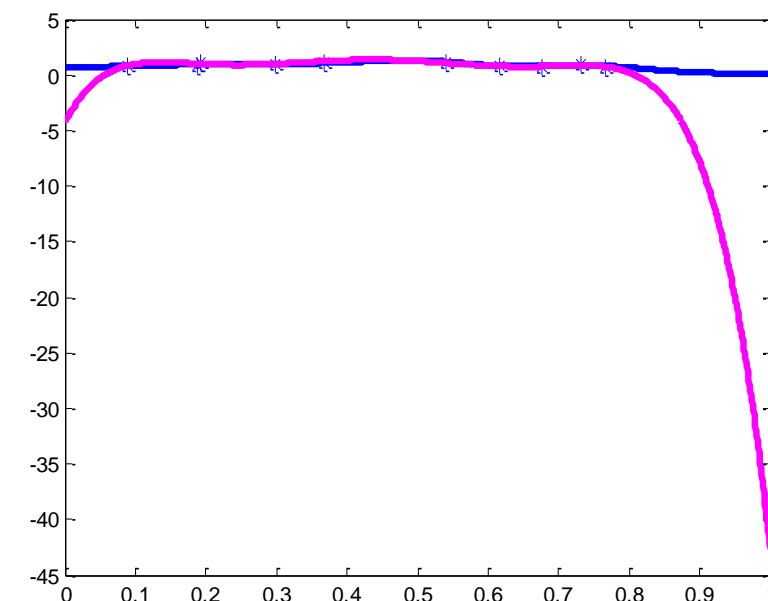
$k=2$   
linear



$k=3$   
quadratic



$k=7$   
6<sup>th</sup> order



# Solutions to Overfitting

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

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

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

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

# 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:  $R_{\mathcal{F}}^* = \inf_{f \in \mathcal{F}} \mathbb{E}[L(Y, f(X))]$   $\hat{R}_{n, \mathcal{F}}^* = \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n L(Y_i, f(X_i))$

Risk Empirical risk

**1<sup>st</sup> issue:**  $R_{\mathcal{F}}^* - R^* \geq 0$  needs to be small.  
 (Model error, Approximation error)  
 Risk in  $\mathcal{F}$  - Bayes risk

## Solution: Structural Risk Minimization (SRM)

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

# Approximation error, Estimation error, PAC framework

Risk of the classifier  $f$

$$R(f) - R(f^*) = \underbrace{R(f) - \inf_{f \in \mathcal{F}} R(f)}_{\text{Estimation error}} + \underbrace{\inf_{f \in \mathcal{F}} R(f) - R(f^*)}_{\text{Approximation error}}$$

$R^*$  (Best classifier in  $\mathcal{F}$ ) points to  $R(f)$  and  $\inf_{f \in \mathcal{F}} R(f)$ .  
 $R^*$  (Bayes risk) points to  $R(f^*)$ .  
 $R_{\mathcal{F}}^*$  (Best classifier in  $\mathcal{F}$ ) points to  $\inf_{f \in \mathcal{F}} R(f)$ .

## Probably Approximately Correct (PAC) learning framework

Learning algorithm produces  $f_n^* = f_{n,\mathcal{F}}^*$  classifier. For each  $\varepsilon, \delta > 0$

we want to find  $n$  large enough such that

$$\Pr(\underbrace{R(f_n^*) - \inf_{f \in \mathcal{F}} R(f)}_{\text{Estimation error}} > \varepsilon) < \delta$$

# Big Picture

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

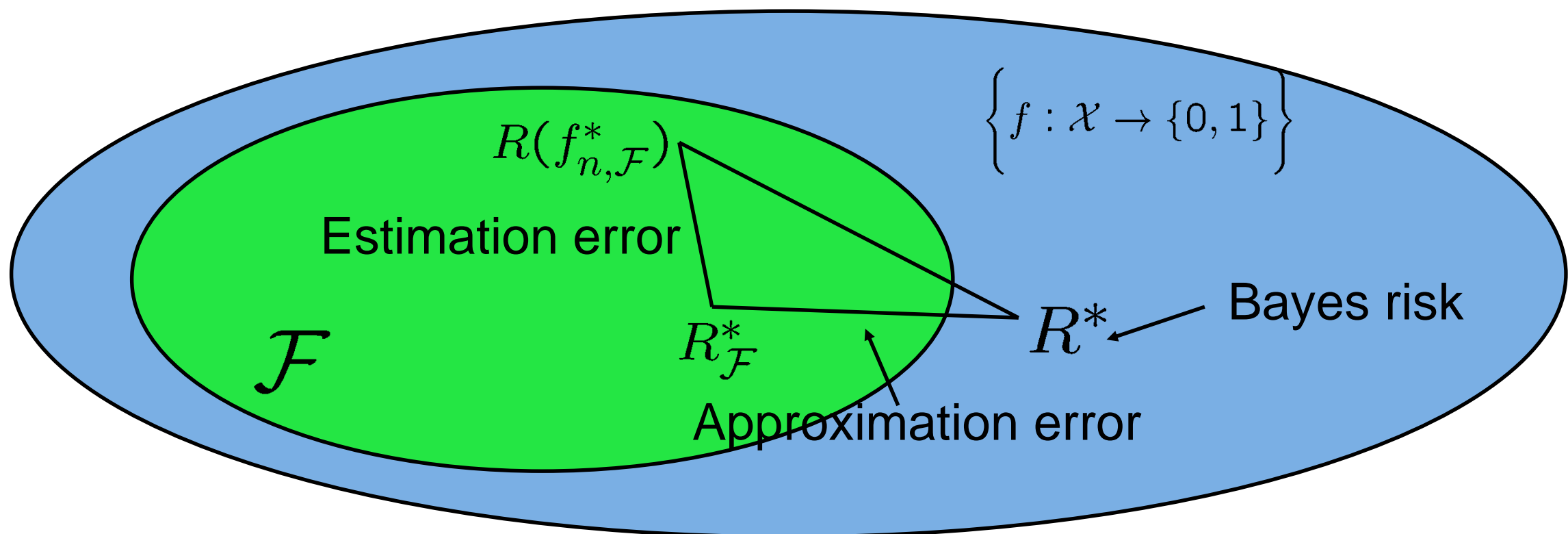
ERM:  $f_n^* = f_{n,\mathcal{F}}^* = \arg \min_{f \in \mathcal{F}} \hat{R}_n(f) = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n L(Y_i, f(X_i))$

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

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

Bayes risk      Bayes risk

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



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

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

$\inf_{f \in \mathcal{F}} \hat{R}_n(f)$  might be a very difficult optimization problem in  $f$   
It might be not even convex in  $f$

## Solution:

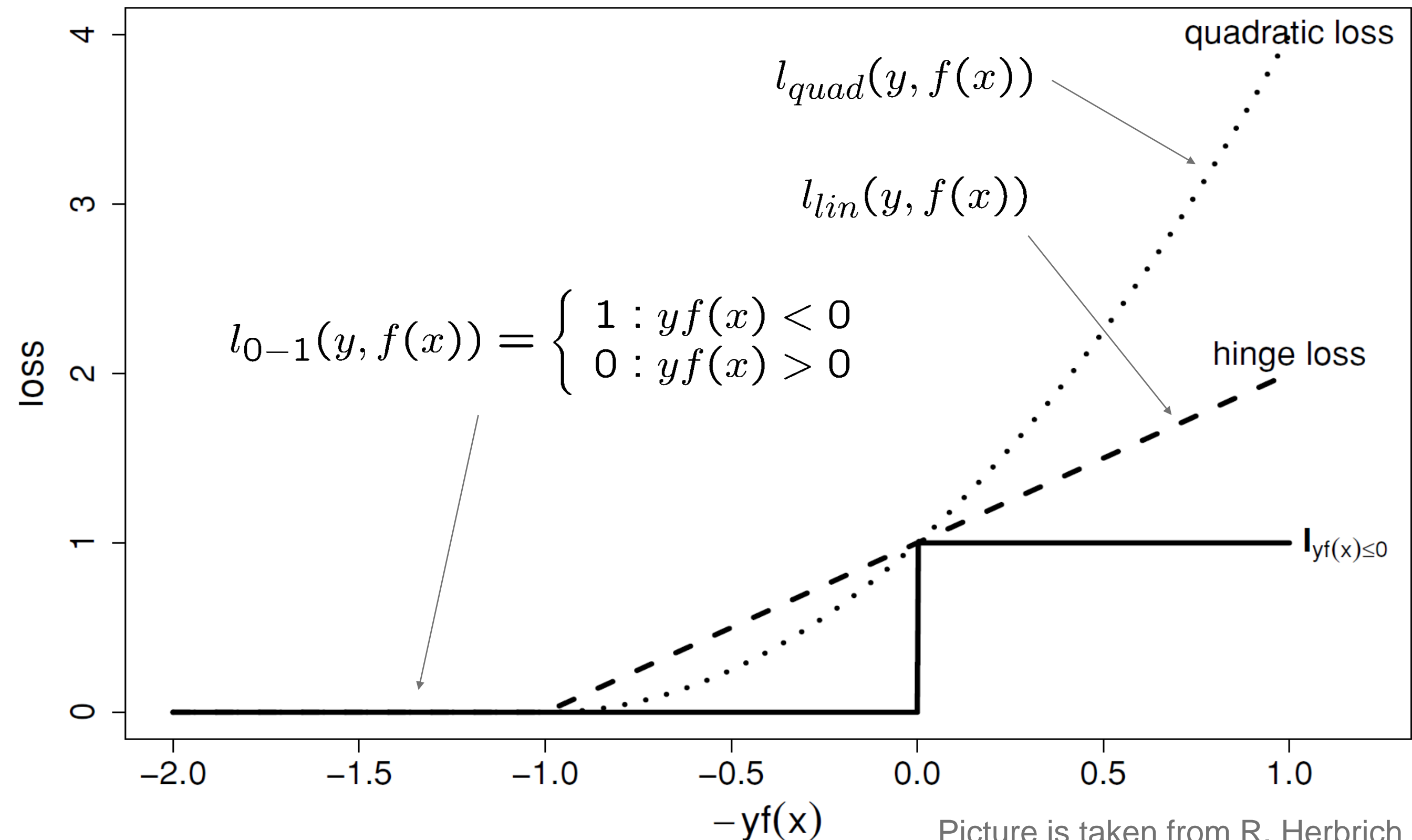
Choose loss function  $L$  such that  $\hat{R}_n(f)$  will be convex in  $f$

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

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

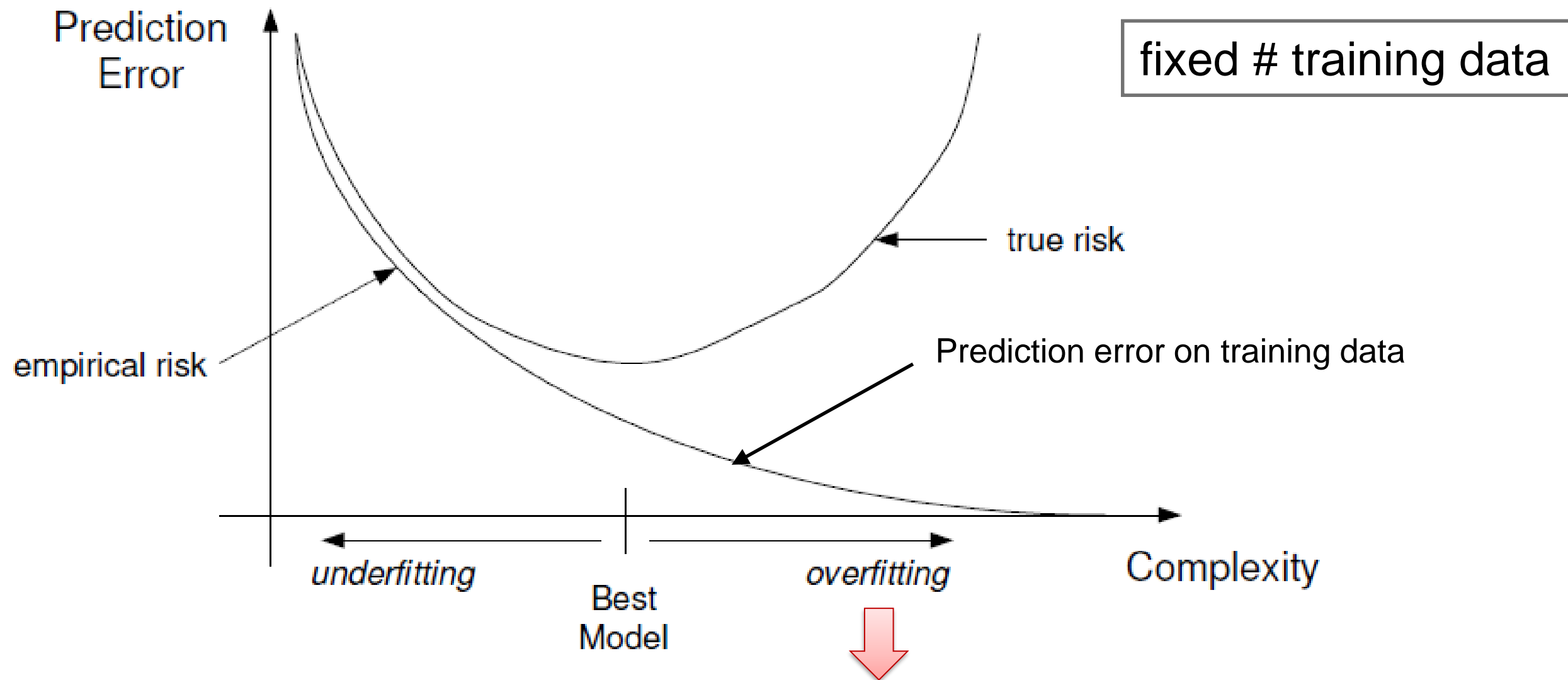
Quadratic loss  $\Rightarrow$  convex  $\hat{R}_n(f)$

# Approximation with the Hinge loss and quadratic loss



# Effect of Model Complexity

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



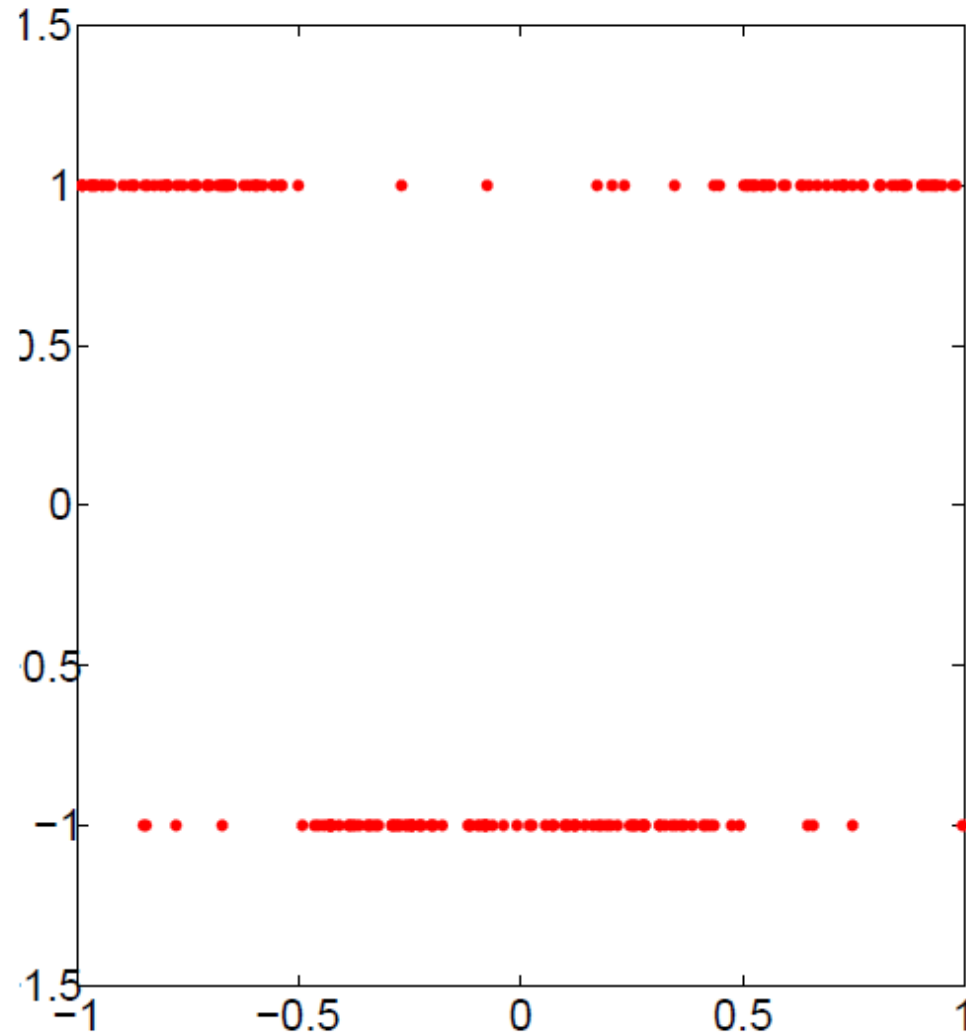
Empirical risk is no longer a good indicator of true risk



# Underfitting

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

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



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

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

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

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

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

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

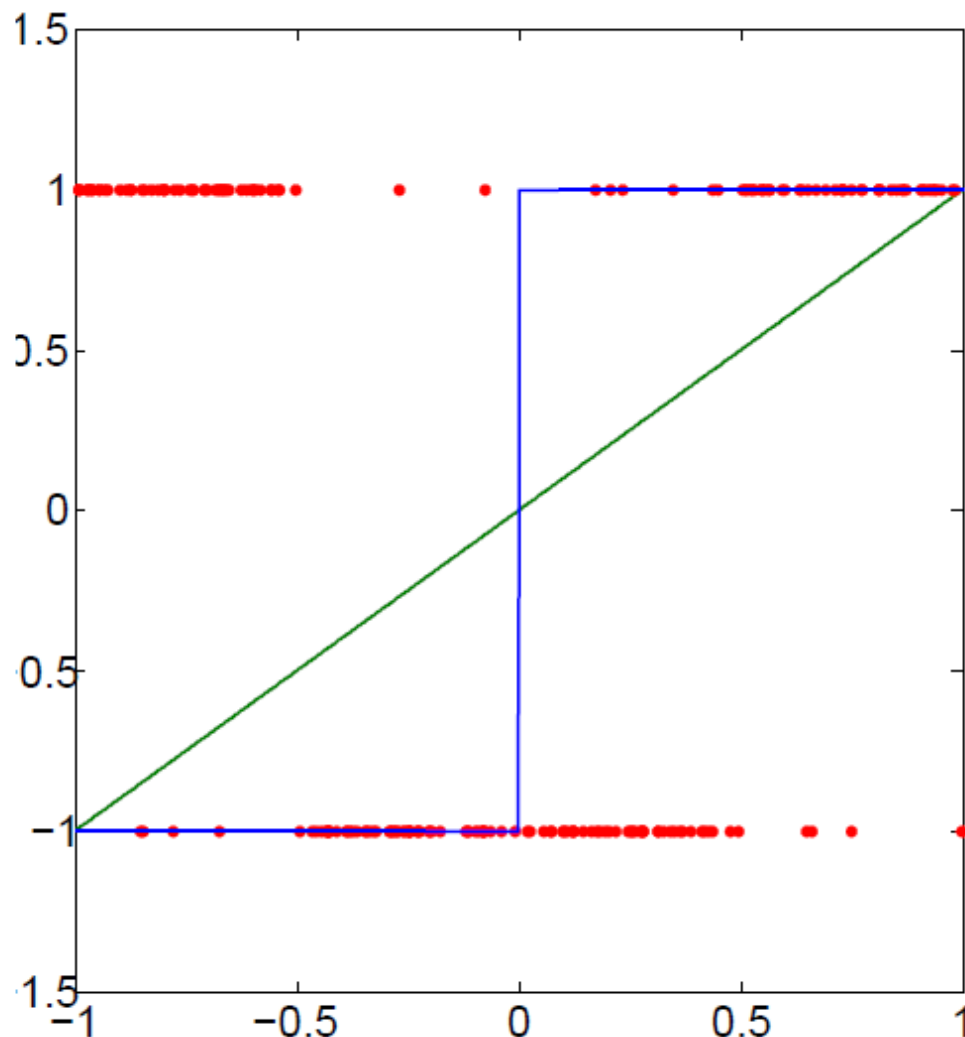
$$R_{\mathcal{F}}^* = \inf_{f \in \mathcal{F}} R(f) = \inf_{f \in \mathcal{F}} \mathbb{E}[L(Y, f(X))]$$

Bayes risk = 0.1

# Underfitting

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

Best linear classifier:



$$\begin{aligned} R_{\mathcal{F}}^* &= R(f_{\mathcal{F}}^*) = \inf_{f \in \mathcal{F}} \Pr[Y \neq f(X)] \\ &= \frac{1}{4} \times 0.9 + \frac{1}{4} \times 0.1 + \frac{1}{4} \times 0.9 + \frac{1}{4} \times 0.1 = 0.5 \end{aligned}$$

The empirical risk of the best linear classifier:

$$\hat{R}_n(f_{\mathcal{F}}^*) \approx 0.5$$

# Underfitting

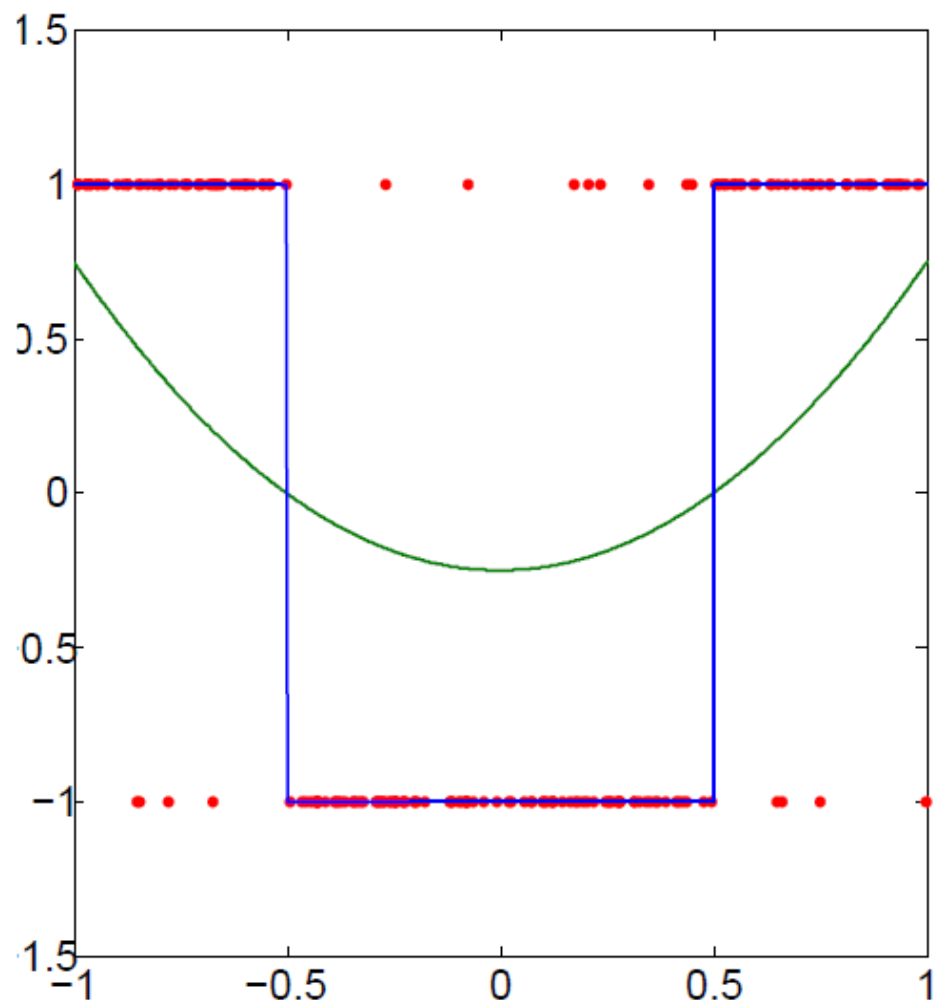
$$\mathcal{F} = \{f : f(x) = \text{sign}(ax^2 + bx + c), a, b, c \in \mathbb{R}\}$$

Best quadratic classifier:

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

$$\begin{aligned} R_{\mathcal{F}}^* &= R(f_{\mathcal{F}}^*) = \inf_{f \in \mathcal{F}} \Pr[Y \neq f(X)] \\ &= \frac{1}{4} \times 0.1 + \frac{1}{4} \times 0.1 + \frac{1}{4} \times 0.1 + \frac{1}{4} \times 0.1 = 0.1 \end{aligned}$$

Same as the Bayes risk ) good fit!



# Classification using the 0-1 loss

# The Bayes Classifier

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

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

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

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

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

# Proofs

Lemma I: Trivial from definition

Lemma II: Surprisingly long calculation

# The Bayes Classifier

$$R(f) = \Pr[Y \neq f(X)]$$

$$R^* = R(f^*) = \inf_{f: \mathcal{X} \rightarrow \mathbb{R}} R(f)$$

$$f^* = \arg \min_{f: \mathcal{X} \rightarrow \mathbb{R}} R(f)$$

$$R_{\mathcal{F}}^* = R(f_{\mathcal{F}}^*) = \inf_{f \in \mathcal{F}} R(f)$$

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

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

$$\hat{R}_{n,\mathcal{F}}^* = \inf_{f \in \mathcal{F}} \hat{R}_n(f)$$

$$f_{n,\mathcal{F}}^* = \arg \min_{f \in \mathcal{F}} \hat{R}_n(f)$$

This is what the learning algorithm produces

**We will need these definitions, please copy it!**

$R(f)$  = Risk

$R^*$  = Bayes risk

$\hat{R}_n(f)$  = Empirical risk

$f^*$  = Bayes classifier

$f_n^* = f_{n,\mathcal{F}}^*$  = the classifier that the learning algorithm produces

# The Bayes Classifier

$$R(f) = \Pr[Y \neq f(X)]$$

$$R^* = R(f^*) = \inf_{f: \mathcal{X} \rightarrow \mathbb{R}} R(f)$$

$$f^* = \arg \min_{f: \mathcal{X} \rightarrow \mathbb{R}} R(f)$$

$$R_{\mathcal{F}}^* = R(f_{\mathcal{F}}^*) = \inf_{f \in \mathcal{F}} R(f)$$

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

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

$$\hat{R}_{n,\mathcal{F}}^* = \inf_{f \in \mathcal{F}} \hat{R}_n(f)$$

$$f_{n,\mathcal{F}}^* = \arg \min_{f \in \mathcal{F}} \hat{R}_n(f)$$

This is what the learning algorithm produces

## Theorem I: Bound on the Estimation error

The true risk of what the learning algorithm produces

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

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

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

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

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

**Proof:**

# The Bayes Classifier

$$R(f) = \Pr[Y \neq f(X)]$$

$$R^* = R(f^*) = \inf_{f: \mathcal{X} \rightarrow \mathbb{R}} R(f)$$

$$f^* = \arg \inf_{f: \mathcal{X} \rightarrow \mathbb{R}} R(f)$$

$$R_{\mathcal{F}}^* = R(f_{\mathcal{F}}^*) = \inf_{f \in \mathcal{F}} R(f)$$

$$f_{\mathcal{F}}^* = \arg \inf_{f \in \mathcal{F}} R(f)$$

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

$$\hat{R}_{n,\mathcal{F}}^* = \inf_{f \in \mathcal{F}} \hat{R}_n(f)$$

$$f_{n,\mathcal{F}}^* = \arg \min_{f \in \mathcal{F}} \hat{R}_n(f)$$

This is what the learning algorithm produces

## Theorem II:

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

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

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

# Proofs

Theorem I: Not so long calculations.

Theorem II: Trivial

Corollary:

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

True risk of the best possible classifier 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

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

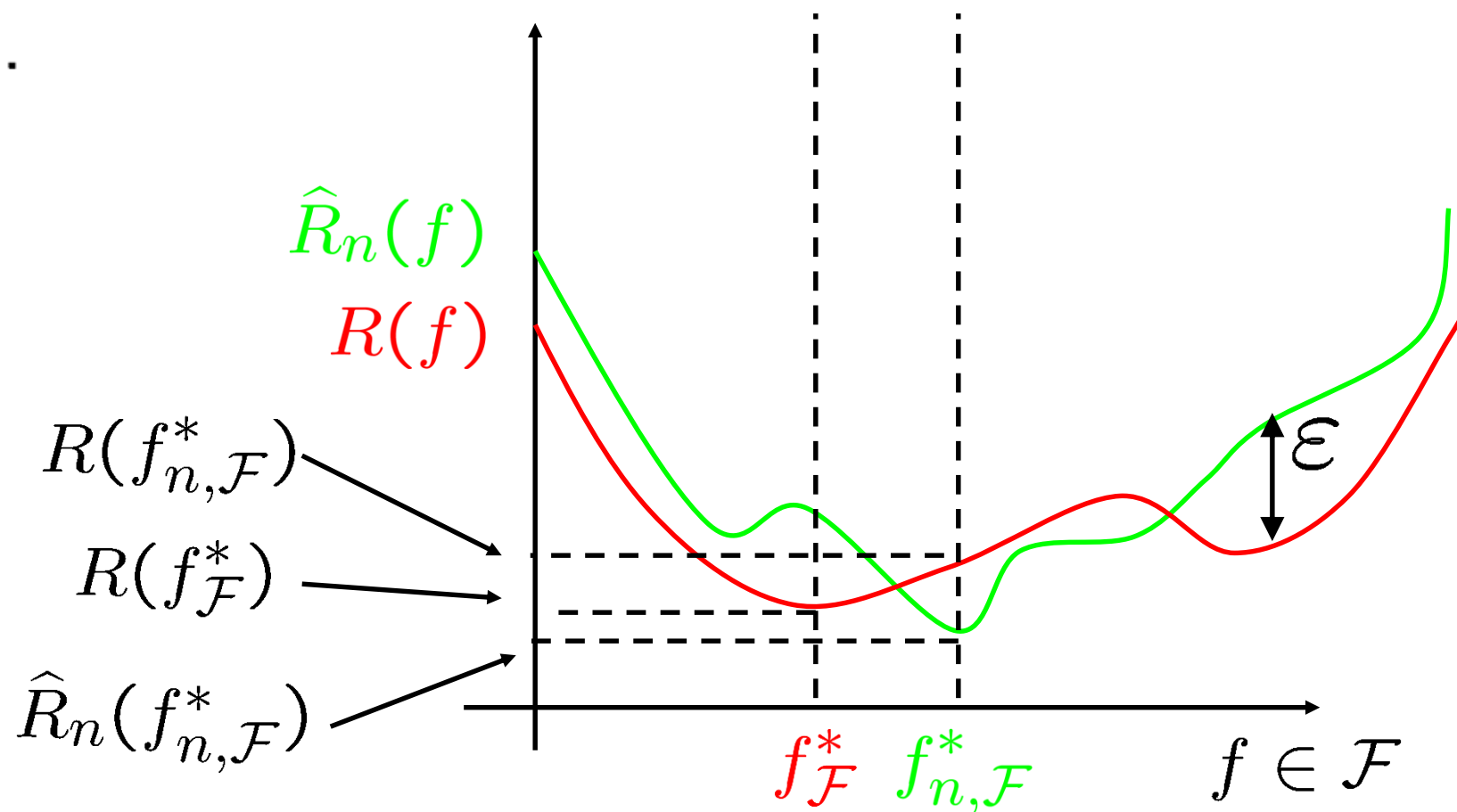
# Illustration of the Risks

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

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

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

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



It's enough to derive upper bounds for

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

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

# Hoeffding's inequality (1963)

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

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

## Special case

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

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

# Binomial distributions

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

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

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

$$\text{Let } Z_i = 1_{\{Y_i \neq f(X_i)\}} \sim \text{Bernoulli}(p)$$
$$\Rightarrow |\hat{R}_n(f) - R(f)| = \left| \frac{1}{n} \sum_{i=1}^n 1_{\{Y_i \neq f(X_i)\}} - p \right| = \left| \frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}[Z_i] \right|$$

Therefore, from Hoeffding we have:

$$\Pr(|\hat{R}_n(f) - R(f)| > \varepsilon) \leq 2 \exp(-2n\varepsilon^2)$$

Yuppie!!!

# Inversion

From Hoeffding we have:

$$\Pr(|\hat{R}_n(f) - R(f)| \geq \varepsilon) \leq 2 \exp(-2n\varepsilon^2)$$

$$\begin{aligned} \text{Let } 2 \exp(-2n\varepsilon^2) &\leq \delta \\ -2n\varepsilon^2 &\leq \log(\delta/2) \\ \varepsilon^2 &\geq \frac{\log(2/\delta)}{2n} \end{aligned}$$

Therefore,

$$\begin{aligned} \Pr\left(|\hat{R}_n(f) - R(f)| \geq \sqrt{\frac{\log(2/\delta)}{2n}}\right) &\leq \delta \\ \Pr\left(|\hat{R}_n(f) - R(f)| < \sqrt{\frac{\log(2/\delta)}{2n}}\right) &\geq 1 - \delta \end{aligned}$$

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



# Union Bound

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

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

**Theorem:** [tail bound on the ‘deviation’ in the worst case]

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

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

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

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

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

# Inversion of Union Bound

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

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

Let  $2N \exp(-2n\varepsilon^2) \leq \delta \Rightarrow -2n\varepsilon^2 \leq \log(\delta/(2N)) \Rightarrow \varepsilon^2 \geq \frac{\log(2N/\delta)}{2n}$

Therefore,

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

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

# Inversion of Union Bound

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

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

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

**We will see later how to fix that.** (Hint: McDiarmid, VC dimension...)

# The Expected Error

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

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

(Tail bound, Concentration inequality)

**Theorem:** [Expected ‘deviation’ in the worst case]

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

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

 **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  $Y \geq 0$ , then  $\mathbb{E}[Y] = \int_0^\infty \Pr(Y \geq z) dz$

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

Thanks for your attention 😊