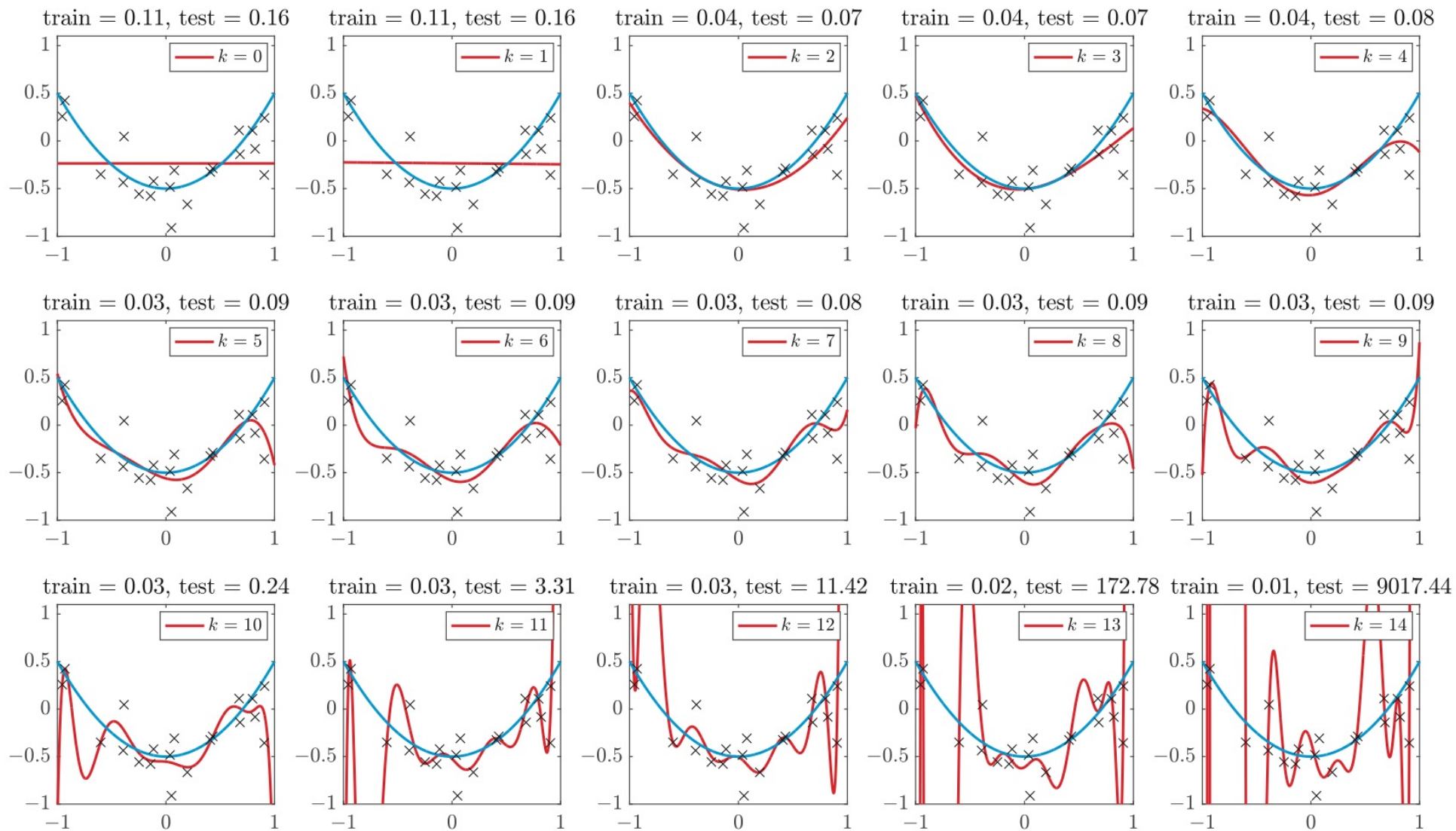


Date	Topics	Book chapters Exercises / Boards
September 17	Learning with infinite data (population setting) -Decision theory (loss, risk, optimal predictors) -Decomposition of excess risk into approximation and estimation errors -No free lunch theorems -Basic notions of concentration inequalities (MacDiarmid, Hoeffding, Bernstein)	Chapter 2
October 1	Linear Least-squares regression -Guarantees in the fixed design settings (simple in closed-form) -Ridge regression: dimension independent bounds -Guarantees in the random design settings -Lower bound of performance	Chapter 3
October 8	Empirical risk minimization -Convexification of the risk -Risk decomposition -Estimation error: finite number of hypotheses and covering numbers -Rademacher complexity -Penalized problems	Chapter 4
October 15	Optimization for machine learning -Gradient descent -Stochastic gradient descent -Generalization bounds through stochastic gradient descent	Chapter 5
October 22	Local averaging techniques -Partition estimators -Nadaraya-Watson estimators -K-nearest-neighbors	Chapter 6
October 29	Kernel methods -Kernels and representer theorems -Algorithms -Universal consistency	Chapter 7
November 5	Model selection -L0 penalty -L1 penalty -High-dimensional estimation	Chapter 8
November 19	Neural networks -Single hidden layer neural networks -Estimation error -Approximation properties and universality	Chapter 9
December 17	Exam	

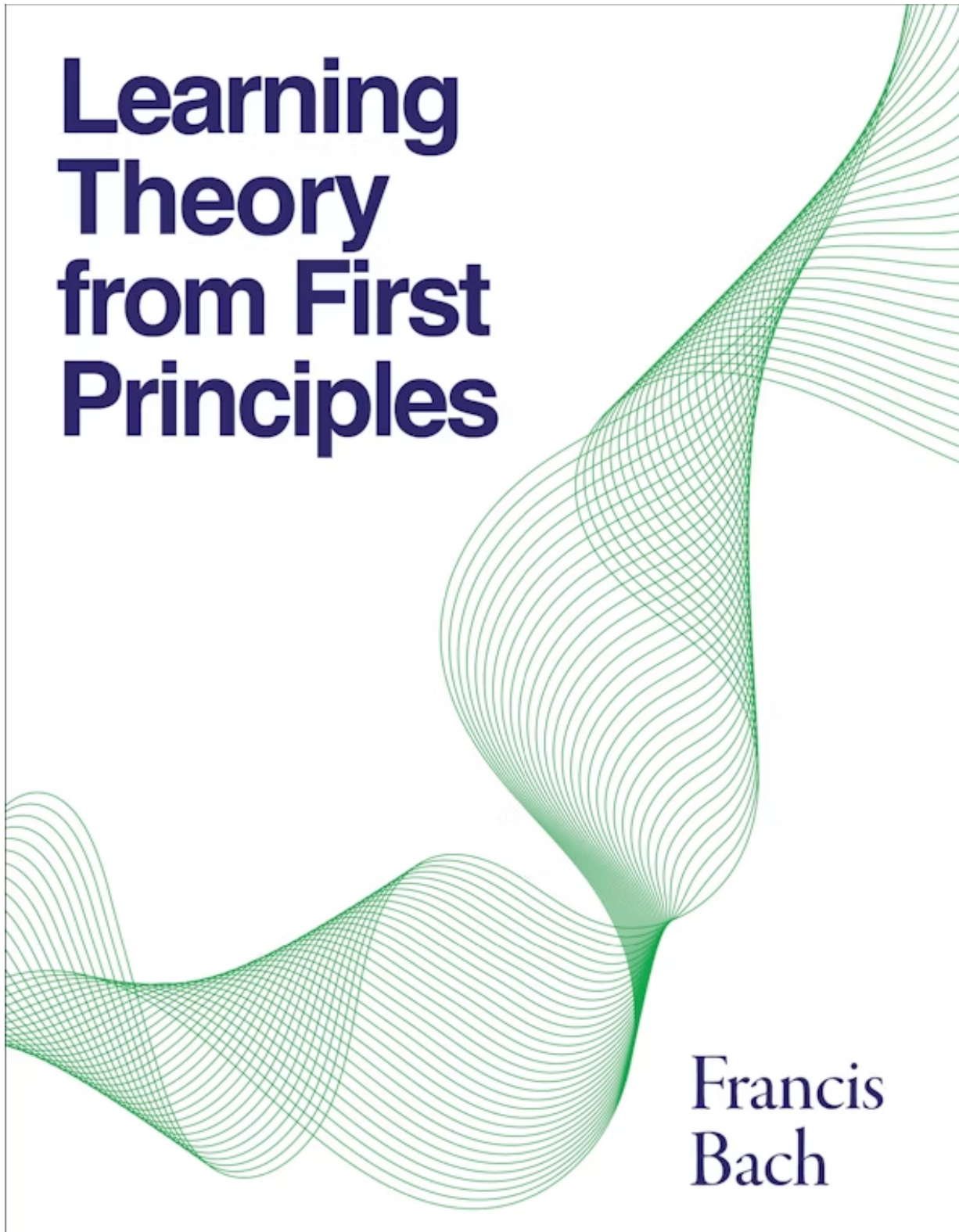
$\text{fre} =$ all poly models of degree less than h



$\mathcal{X}$	Input space
$\mathcal{Y}$	Output space
p	Joint distribution on $\mathcal{X} \times \mathcal{Y}$
$(x_1, y_1, \dots, x_n, y_n)$	Training data
$f : \mathcal{X} \rightarrow \mathcal{Y}$	Prediction function
$\ell(y, z)$	Loss function between output y and prediction z
$\mathcal{R}(f) = \mathbb{E}[\ell(y, f(x))]$	Expected risk of prediction function f
$\hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$	Empirical risk of prediction function f
$f_*(x') = \arg \min_{z \in \mathcal{Y}} \mathbb{E}[\ell(y, z) x = x']$	Bayes prediction at x'
$\mathcal{R}^* = \mathbb{E}_{x' \sim p} \inf_{z \in \mathcal{Y}} \mathbb{E}[\ell(y, z) x = x']$	Bayes risk

Learning Theory from First Principles

Francis
Bach



Supervised Learning

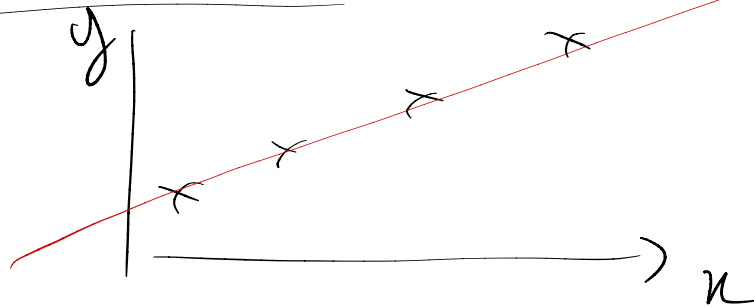
Training data $(x_i, y_i) \in X \times Y$ $i = 1, \dots, n$

input
features
covariates

output
labels
response

Regression $Y = \mathbb{R}$
Classification
 $Y = \text{finite set}$

Why is it difficult?



non linearly
 $X \rightarrow Y$
noise

model $y = f(x)$
+ noise

Formalization

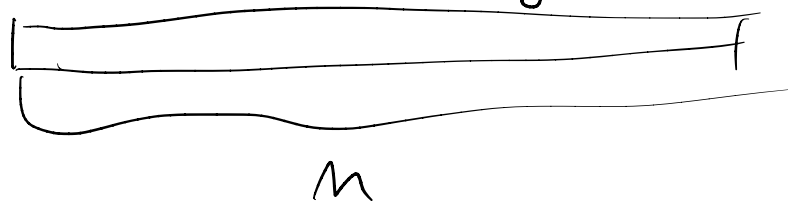
(x_i, y_i)

IID from a distribution $p(x, y)$

independent
and identically distributed

and $p(x, y)$
is the
test dist.

Measurement of test error

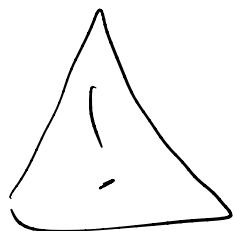


testing data

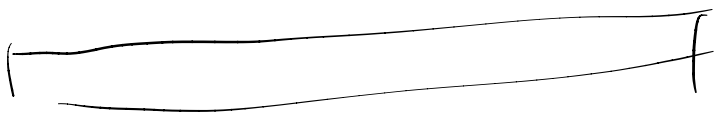
n_{test}

find $\hat{\theta}$ s.t.
as a fcn of $(x_i, y_i)_{i=1, \dots, n}$

in this case: Expectations
in practice: empirical
mean



No training on the testing data



training

validation

Decision theory : $x \in X$ input
 $y \in Y$ output
joint distribution $p(x, y)$

Loss : $\ell : Y \times Y \rightarrow \mathbb{R}$
 $\ell(y, z)$ loss of predicting z when the output is y

Examples : Y finite Classification

0-1 loss : $\ell(y, z) = 1_{y \neq z} = 1$ if $y \neq z$
 0 if $y = z$

binary : $Y = \{0, 1\}$ $\{1, 2\}$ $\{-1, 1\}$

multiclass $Y = \{1, \dots, k\}$

$Y = \mathbb{R}$ Regression : $\ell(y, z) = \frac{(y - z)^2}{|y - z|}$ square absolute value

• Surrogates (logistic)

• Risks : $X \times Y$
 $f: X \rightarrow Y$ prediction function

Def: Expected risk $R(f) = \mathbb{E}[e(y, f(x))]$
testing error
generalization performance
 $= \int e(y, f(x)) d\rho(x, y)$

R is a "functional" deterministic

$R(\hat{f})$ is random

Ex: classification
0-1 loss
error rate
 $R(f) = \mathbb{E}[1_{y \neq f(x)}] = P(y \neq f(x))$
 $= 1 - P(y = f(x))$
 $= 1 - \text{accuracy}$

Regression: $R(f) = \mathbb{E}[(y - f(x))^2]$

Def: Empirical risk: $\hat{R}(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$
training error

Impractical: compute $\hat{R}(f)$

goal of ML: find f that minimizes $R(f)$

Bayes risk, Bayes predictor

$$R(f) = \mathbb{E}[\ell(y, f(x))] = \mathbb{E}[\mathbb{E}[\ell(y, f(x)) | x]]$$

Fix $x \in X$

$$= \sum_{x' \in X} p(x=x') \mathbb{E}[\ell(y, f(x')) | x=x']$$

goal: minimize wrt $f: x \rightarrow y$ are value for each $x \in X$

$$f^*(x) \in \arg \min_{y \in Y} \mathbb{E}[\ell(y, z) | x=x']$$
$$= \int \ell(y, z) d p(y | x=x')$$

$$\min_z h(z) = \sum_i h_i(z_i) = \sum_i \min_{z_i} h_i(z_i)$$

$z_i^* \in \arg\min h_i$

$$h(z) - \sum_i h_i(z_i^*) = \sum_i h(z_i) - h_i(z_i^*)$$

Prop : Bayes predictor

Def

$$f^*(x') \in \arg \min_{z \in \mathcal{Y}}$$

$$\mathbb{E}[e(y, z) | x = x']$$

cardinal risk

Bayes
risk

$$R(f^*) = \mathbb{E} \left[\min_z \mathbb{E}[e(y, z) | x] \right]$$

Classif

0-1 loss

$$f^*(x') \in \arg \min_z \mathbb{E} [1_{y \neq z} | x]$$
$$P(y \neq z | x)$$

$$\arg \max_z P(y = z | x)$$

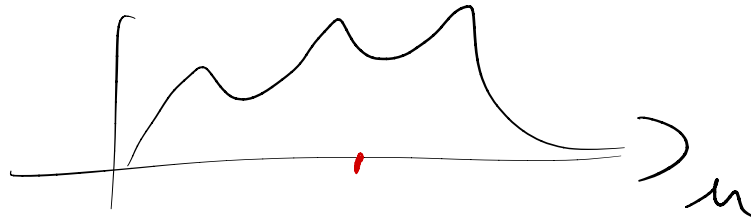
$$X = \{-1, 1\} \quad f^*(x) = \begin{cases} 1 & \text{if } P(y=1|x) > P(y=-1|x) \\ -1 & \text{if } P(y=1|x) < P(y=-1|x) \end{cases}$$

is the Bayes risk always 0? =

Regression : $\mathbb{E}[E(y, z) | n = n']$

square loss $\mathbb{E}[(y - z)^2 | n = n'] = \int (y - z)^2 dP(y | n = n')$

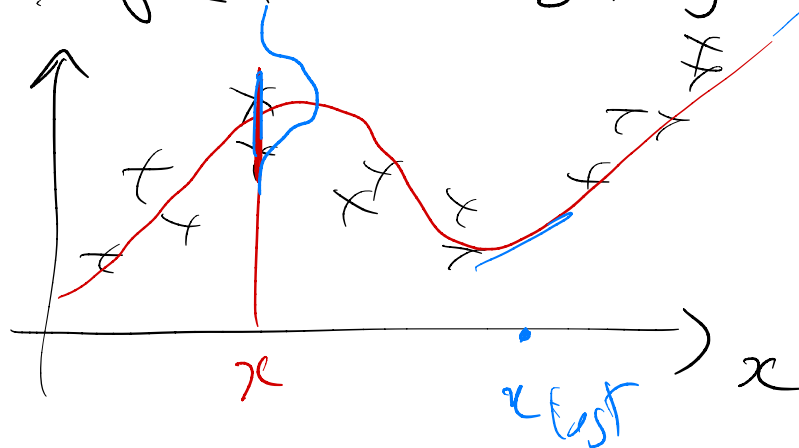
lemma $\min_z \mathbb{E}[(u - z)^2]$ $u \in \mathbb{R}$ random



$$= \mathbb{E}(u)$$

proof $\mathbb{E}[(u - z)^2] = (\mathbb{E}(u) - z)^2 + \text{var}(u) + \mathbb{E}[(u - \mathbb{E}(u))^2]$

Bayes predictor : $f^*(x) = \mathbb{E}[y | x]$



$f^*(x)$

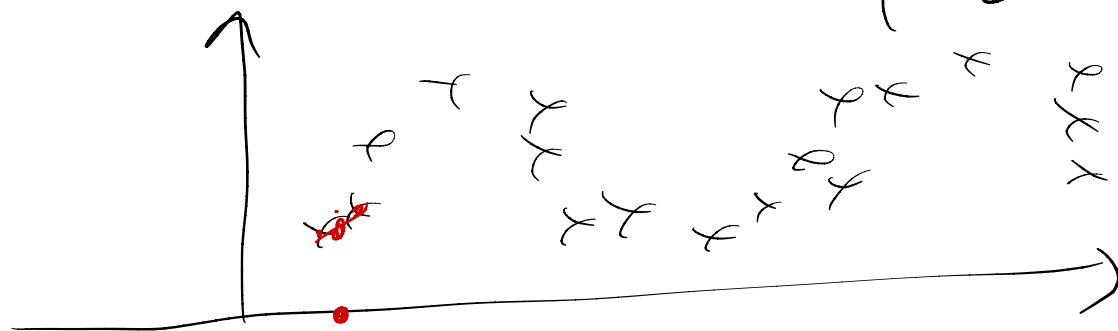
$$f^*(x') = \mathbb{E}[y | n = n']$$

interpolation
extrapolation

$\begin{bmatrix} 10^2 8 \\ 10^4 8 \end{bmatrix}$

Local averaging: k -nearest neighbor / Nadaraya-Watson

$f^*(x') \in \arg \min E[\ell(y) | n = n']$
need to know $p(y | n = n')$



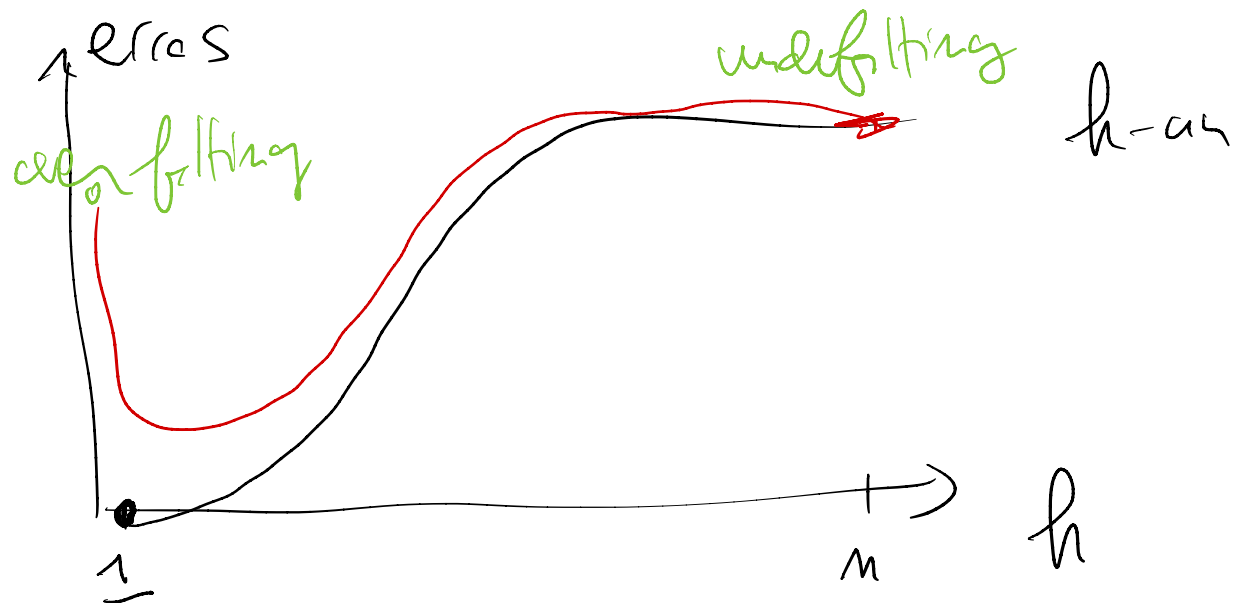
2 nearest neighbors

principle: build $\hat{p}(y | n = n') = \sum_{i=1}^n w_i(n') \delta_{y_i}$
with formula

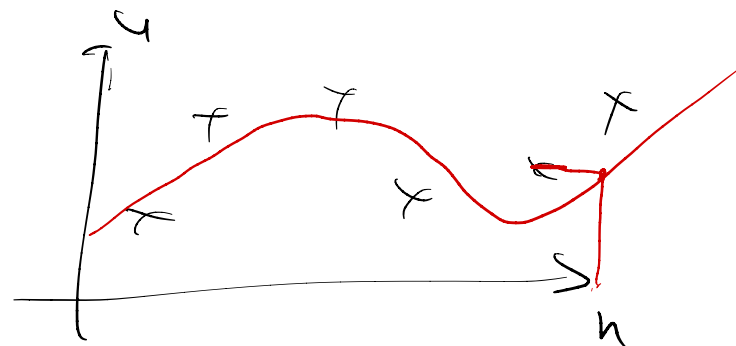
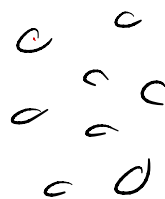
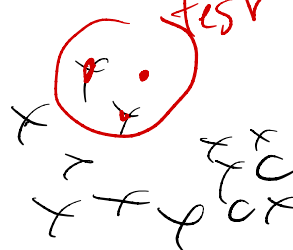
+ non-parametric
linear
no training
interpretability
guarantees

- Curse of dimensionality
costly in high dim
excess risk $\sim \frac{1}{n^{1/d}}$

train
test



2-NN
test



Empirical risk minimization (ERM)

family of prediction functions $f_{\theta}(x)$, $\theta \in \Theta$

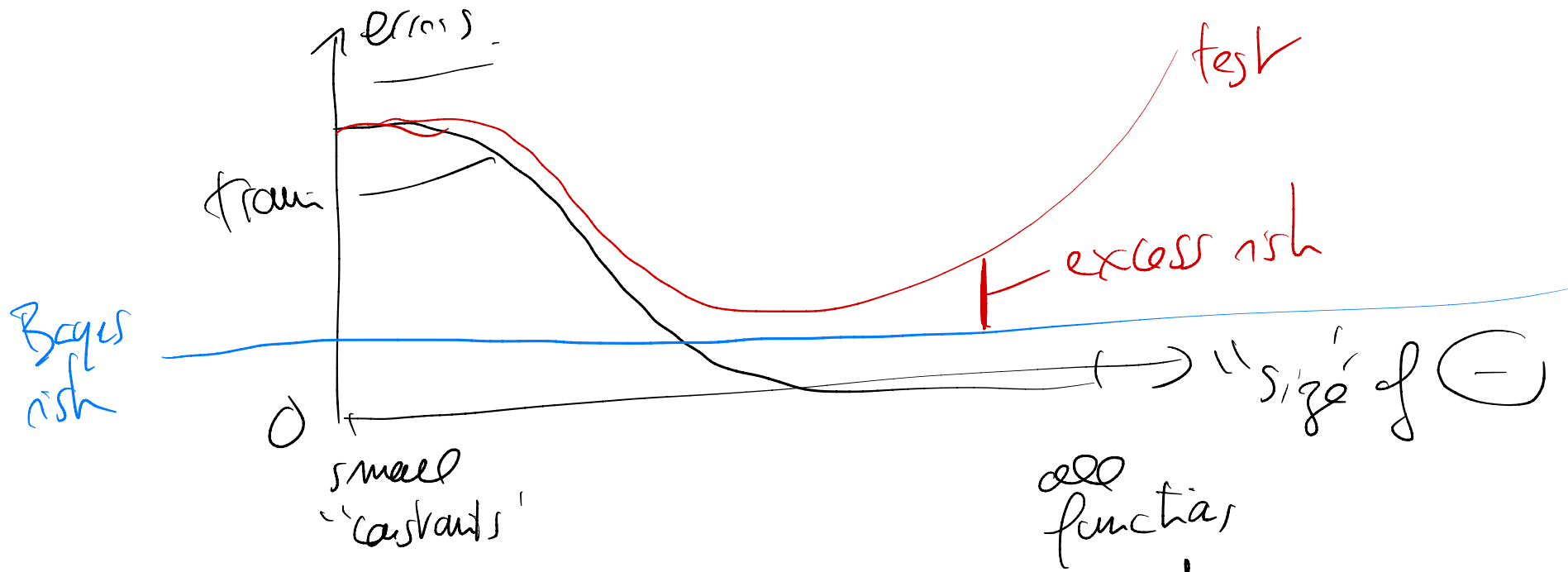
goal: find $\hat{\theta} \in \arg \min \hat{R}(f_{\theta})$

$$\frac{1}{n} \sum_{i=1}^n \ell(y_i, f_{\theta}(x_i))$$

test time: $\hat{f} = f_{\hat{\theta}}$

$\perp$ $\hat{R} \rightarrow R$ when $n \rightarrow \infty$
rely on minimization algo.

$\rightarrow$ Θ depends
on data
hard to
optimize
or helping



Regularized constrained ERM:

$$\min_{\Theta \text{ subset of } \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^{(n-1)}) + \frac{\lambda}{2} \Omega(\theta)$$

s.t. $\Omega(\theta) \leq c_n$

Risk de composition : model $\theta_0 \in \Theta$
 estimator $\hat{\theta} \in \Theta$

$$R(\hat{\theta}) - R_{\star} = \underbrace{R(\hat{\theta}) - \inf_{\theta' \in \Theta} R(\theta')}_{\substack{\text{estimation error} \\ \geq 0 \\ \text{random} \\ \text{goes down with } n}} + \underbrace{\inf_{\theta' \in \Theta} R(\theta') - R_{\star}}_{\substack{\text{approximation error} \\ \geq 0 \\ \text{deterministic} \\ \text{no change with } n \\ \text{if } \Theta \text{ fixed}}}$$

$\uparrow$
 expected risk
 excess risk

Bayes Risk

$\theta_{\star}$
 = argmin $R(\theta)$

$R(\hat{\theta}) - R(\theta_{\star})$ when $\hat{\theta}$ is an approximate minimizer of $R(\theta)$ wrt $\theta \in \Theta$

$$= \underbrace{R(\hat{\theta}) - \hat{R}(\hat{\theta})}_{\substack{\hat{\theta} \text{ random}}} + \underbrace{\hat{R}(\hat{\theta}) - \hat{R}(\theta_{\star})}_{\substack{\leq 0 \text{ if } \hat{\theta} \text{ is a minimizer} \\ \leq \hat{R}(\hat{\theta}) - \inf_{\theta' \in \Theta} \hat{R}(\theta')}} + \underbrace{\hat{R}(\theta_{\star}) - R(\theta_{\star})}_{\substack{\rightarrow 0 \\ n \rightarrow \infty}}$$

$$\leq \sup_{\theta \in \Theta} (R(\theta) - \hat{R}(\theta))$$

UNIFORM DEVIATIONS

optimization error ≥ 0

goals of learning theory

test distribution $p \rightarrow$ iid data $(x_1, y_1), \dots, (x_n, y_n)$
from p
 $D_n(p)$ dataset

ML algorithm: $D_n(p) \rightarrow$ prediction function
 A
 $A(D_n(p)) : X \rightarrow Y$

Performance measure $R_p[A(D_n(p))] - R_p^*$ ≥ 0
classical LT statement $\hookrightarrow$ Bayes risk

$\forall p \in \mathcal{P}, \quad \mathbb{E}[R_p[A(D_n(p))] - R_p^*] \leq \epsilon_n(\mathcal{P}, A)$
 $\hookrightarrow$ family of problems $\rightarrow 0$

with probability $\geq 1 - \delta$
 $R_p[A(D_n(p))] - R_p^* \leq \epsilon_n(\mathcal{P}, A, \delta)$