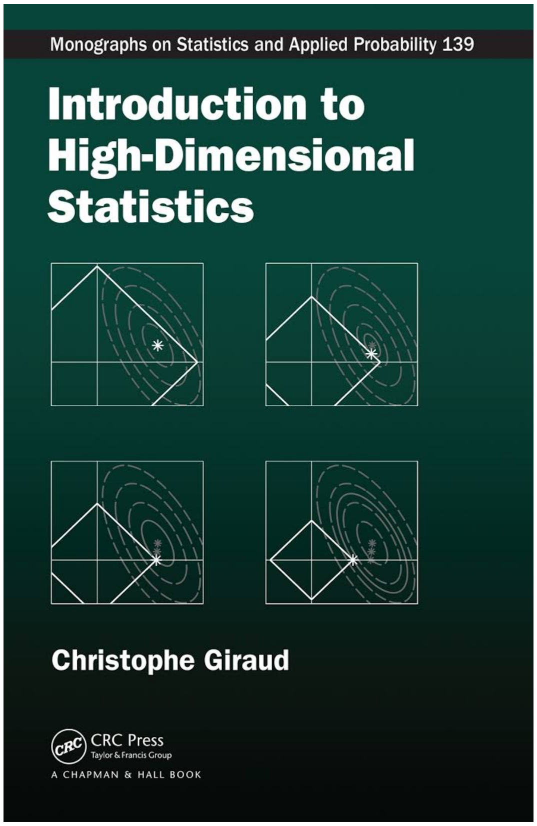
Empirical Risk Minimization, Linear Separators, Risk Convexification

Yohann De Castro & Aurélien Garivier

UNIVERSITÉ DE LYON

ENS DE LYON



Chap 9 :

Supervised Classification

→ Bounds on Risk Classification

→ proofs using VC-dimension

Experience, Task and Performance measure

- Training data : $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ (i.i.d. $\sim \mathbf{P}$)
- Predictor: $f : \mathcal{X} \to \mathcal{Y}$ measurable
- Cost/Loss function : $\ell(Y, f(\mathbf{X}))$ measure how well $f(\mathbf{X})$ "predicts" Y
- Risk:

$$\mathcal{R}(f) = \mathbb{E}\left[\ell(Y, f(\mathbf{X}))\right] = \mathbb{E}_X\left[\mathbb{E}_{Y|\mathbf{X}}\left[\ell(Y, f(\mathbf{X}))\right]\right]$$

- Often $\ell(Y, f(\mathbf{X})) = |f(\mathbf{X}) - Y|^2$ or $\ell(Y, f(\mathbf{X})) = \mathbf{1}_{Y \neq f(\mathbf{X})}$

Goal

- Learn a rule to construct a classifier $\widehat{f} \in \mathcal{F}$ from the training data $\mathcal{D}_n$ s.t. the risk $\mathcal{R}(\widehat{f})$ is small on average or with high probability with respect to $\mathcal{D}_n$.

$$R(f) = \mathbb{E}\left[\ell(Y, f(X)) \right]$$

$$\int_{\mathcal{X} \times \mathcal{Y}} \ell(y, f(x)) \, P_{(X,Y)}(x,y) \, dx \, dy$$

$$= \int_{\mathcal{X}} \left[\int_{\mathcal{Y}} \ell(y, f(x)) \, P_{(X,Y)}(x,y) \, dy \right] dx$$

$$P_{(X,Y)}(x,y) = \boxed{\frac{P_{(X,Y)}(x,y)}{\int_{\mathcal{Y}} P_{(X,Y)}(x,y)\, dy}} \times \int_{\mathcal{Y}} P_{(X,Y)}(x,y)\, dy$$

$P_{Y|X=x}(y)$

$$\int_{\mathcal{Y}} \frac{P_{(X,Y)}(x,y)}{\int_{\mathcal{Y}} P_{(X,Y)}(x,y)\, dy} \, dy = 1$$

$$R(f) = \mathbb{E}_X \left[\underbrace{\mathbb{E}_{Y|X} \left[\ell(Y, f(X)) \right]}_{\text{conditional expectation}} \right]$$

conditional expectation

- The best solution f^* (which is independent of $\mathcal{D}_n$) is

$$f^* = \arg\min_{f \in \mathcal{F}} R(f) = \arg\min_{f \in \mathcal{F}} \mathbb{E}\left[\ell(Y, f(\mathbf{X}))\right] = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[\ell(Y, f(\mathbf{x}))\right]\right]$$

Bayes Classifier (explicit solution)

- In binary classification with $0-1$ loss:

$$f^*(\mathbf{X}) = \begin{cases} +1 & \text{if } \mathbb{P}\{Y = +1|\mathbf{X}\} \geq \mathbb{P}\{Y = -1|\mathbf{X}\} \\ & \Leftrightarrow \mathbb{P}\{Y = +1|\mathbf{X}\} \geq 1/2 \\ -1 & \text{otherwise} \end{cases}$$

$= \mathrm{sgn}(\mathbb{E}[Y|X])$

- In regression with the quadratic loss

$$f^*(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$$

Issue: Explicit solution requires to know $\mathbb{E}[Y|\mathbf{X}]$ for all values of $\mathbf{X}$!

Example: Binary Classification $\mathcal{Y} = \{-1, +1\}$

Best Predictor for the misclassification loss:

$$\ell(y, y') = \mathbb{1}_{y \neq y'}$$

$$R(f) = \mathbb{E}\left[\ell(Y, f(X))\right] = \underbrace{\mathbb{P}\left[Y \neq f(X)\right]}_{\text{proportion of misclassification}}$$

$$f^*(x) = \operatorname{sgn}\left(\mathbb{E}\left[Y \mid X = x\right]\right)$$

$$\mathbb{E}\left[Y \mid X = x\right] = +1 \times \mathbb{P}\left[Y = 1 \mid X = x\right] + (-1) \times \mathbb{P}\left[Y = -1 \mid X = x\right]$$

$$= \mathbb{P}\left[Y = 1 \mid X = x\right] - \mathbb{P}\left[Y = -1 \mid X = x\right]$$

$$\to \; = \begin{cases} +1 & \text{if } \mathbb{P}[Y=1 \mid X=x] \geq \mathbb{P}[Y=-1 \mid X=x] \\ -1 & \text{if } \mathbb{P}[Y=-1 \mid X=x] \geq \mathbb{P}[Y=1 \mid X=x] \end{cases}$$

Proof :

$$R(f) = \mathbb{E}\left[\mathbb{1}_{Y \neq f(X)} \right]$$

$$= \mathbb{E}_X \left[\mathbb{E}_{Y|X} \left(\mathbb{1}_{Y \neq f(X)} \right) \right]$$

$$= \mathbb{E}_X \left[\mathbb{P}\left(Y \neq f(X) \mid X \right) \right]$$

Y: random variable

$f(X) \in \mathcal{Y} = \{-1, +1\}$ ← fixed $X = x$

$$\mathbb{P}\left(Y \neq f(x) \mid X = x \right)$$

Y: random with law $Y | X = x$

$f(x)$: fixed

If $f(x) = -1$ then :

$$\mathbb{P}\left(Y \neq -1 \mid X = x \right) = \mathbb{P}\left(Y = 1 \mid X = x \right)$$

If $f(x) = +1$ then

$$\mathbb{P}(Y \neq 1 \mid X = x) = \frac{\mathbb{P}(Y = -1 \mid X = x)}{}$$

Machine Learning

- Learn a rule to construct a classifier $\widehat{f} \in \mathcal{F}$ from the training data $\mathcal{D}_n$ s.t. the risk $\mathcal{R}(\widehat{f})$ is small on average or with high probability with respect to $\mathcal{D}_n$.

Canonical example: Empirical Risk Minimizer

- One restricts f to a subset of functions $\mathcal{S} = \{f_\theta, \theta \in \Theta\}$
- One replaces the minimization of the average loss by the minimization of the empirical loss

$$\widehat{f} = f_{\widehat{\theta}} = \underset{f_\theta, \theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\theta(\mathbf{X}_i))$$

- Examples:
  - Linear regression
  - Linear discrimination with
    $$\mathcal{S} = \{\mathbf{x} \mapsto \mathtt{sign}\{\beta^T \mathbf{x} + \beta_0\} \,/\, \beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}\}$$

$$\min_{f \text{ measurable}} \mathbb{E}[\ell(Y, f(X))]$$

Impossible → $\theta \in \mathbb{R}^k$ ← parameters

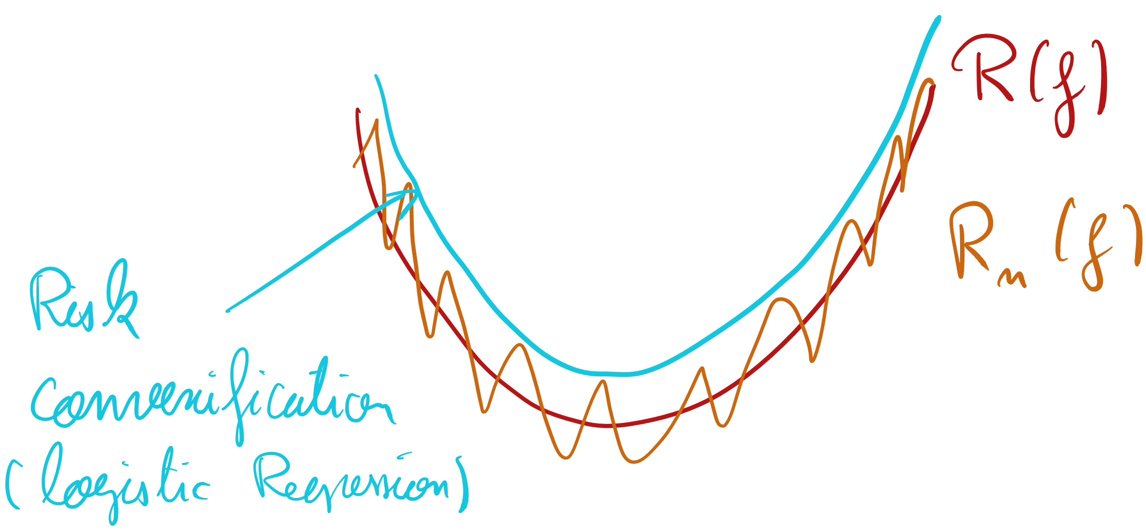
f_θ

Not access to → Empirical Risk

$$R_n(f) = \mathbb{E}_n(\ell(Y, f(X))) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$$

(random)

Training set : (X_i, Y_i)
(random)

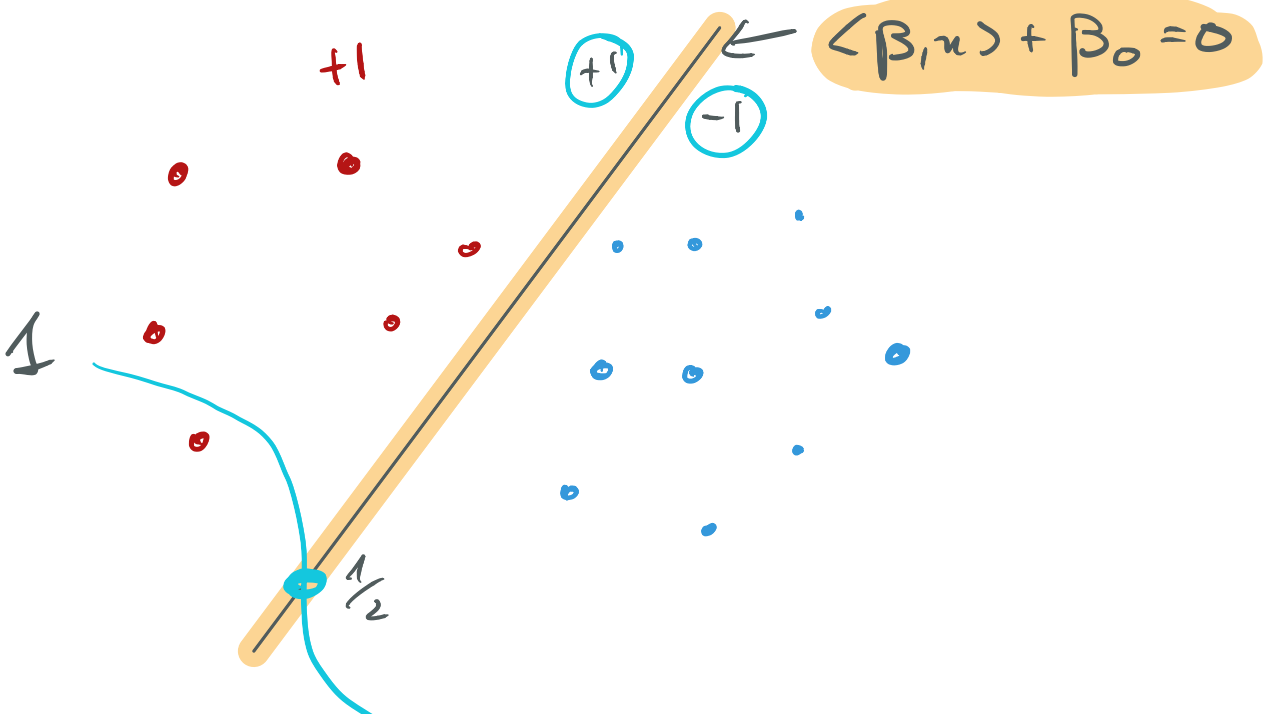


Linear Classifier

- Classifier family:

$$\mathcal{S} = \{f_\theta : \mathbf{x} \mapsto \mathtt{sign}\{\beta^T\mathbf{x} + \beta_0\} \,/\, \beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}\}$$

- Natural loss: $\ell^{0/1}(Y, f(x)) = \mathbf{1}_{y \neq f(x)}$



Linear Classifier

- Classifier family:

$$\mathcal{S} = \{f_\theta : \mathbf{x} \mapsto \mathtt{sign}\{\beta^T \mathbf{x} + \beta_0\} \,/\beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}\}$$

- Natural loss: $\ell^{0/1}(Y, f(x)) = \mathbf{1}_{y \neq f(x)}$

Empirical Risk Minimization

- ERM Classifier:

$$\widehat{f} = f_{\widehat{\theta}} = \operatorname*{argmin}_{f_\theta, \theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{Y_i \neq f_\theta(\mathbf{x}_i))}$$

- Not smooth or convex $\implies$ no easy minimization scheme!
- $\neq$ regression with quadratic loss case!

- How to go beyond?

NP-hard

Bayes Classifier and Plugin

- Best classifier given by

$$f^*(\mathbf{X}) = \begin{cases} +1 & \text{if } \mathbb{P}\{Y = +1|\mathbf{X}\} \geq \mathbb{P}\{Y = -1|\mathbf{X}\} \\ & \Leftrightarrow \mathbb{P}\{Y = +1|\mathbf{X}\} \geq 1/2 \\ -1 & \text{otherwise} \end{cases}$$

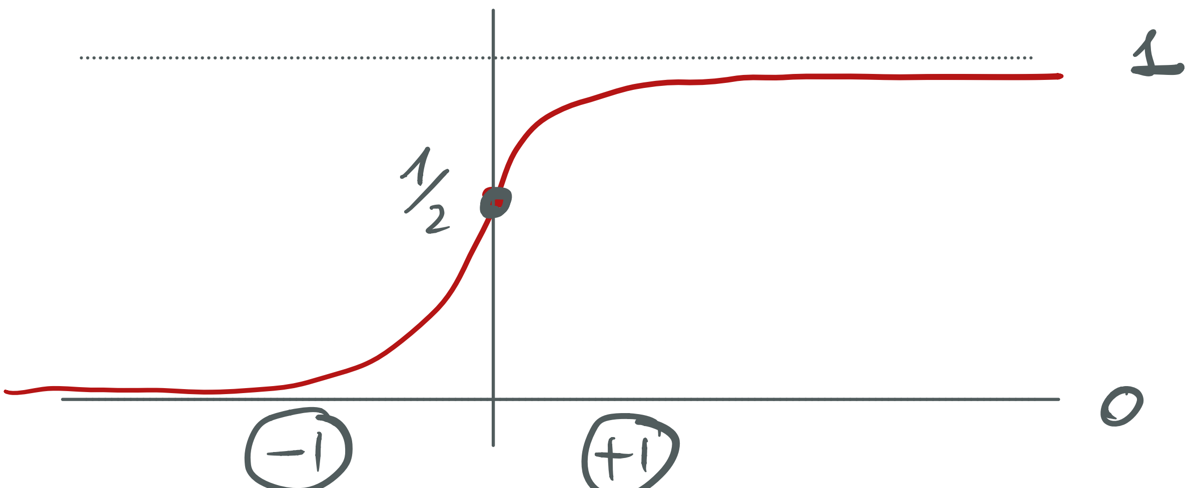
- Plugin classifier: replace $\mathbb{P}\{Y = +1|\mathbf{X}\}$ by a data driven estimate $\mathbb{P}\{\widehat{Y = +1}|\mathbf{X}\}$!

- Other strategies are possible (Risk convexification...)

Plugin Linear Discrimination

- Model $\mathbb{P}\{Y = +1|\mathbf{X}\}$ by $h(\beta^T\mathbf{X} + \beta_0)$ with h non decreasing.
- $h(\beta^T\mathbf{X} + \beta_0) > 1/2 \Leftrightarrow \beta^T\mathbf{X} + \beta_0 - h^{-1}(1/2) > 0$
- Linear Classifier: $\text{sign}(\beta^T\mathbf{X} + \beta_0 - h^{-1}(1/2))$

$$\mathbb{P}[Y=1|X] \longleftarrow h(\langle \beta, x\rangle + \beta_0)$$



Plugin Linear Discrimination

- Model $\mathbb{P}\{Y = +1|\mathbf{X}\}$ by $h(\beta^T\mathbf{X} + \beta_0)$ with h non decreasing.
- $h(\beta^T\mathbf{X} + \beta_0) > 1/2 \Leftrightarrow \beta^T\mathbf{X} + \beta_0 - h^{-1}(1/2) > 0$
- Linear Classifier: $\operatorname{sign}(\beta^T\mathbf{X} + \beta_0 - h^{-1}(1/2))$

Plugin Linear Classifier Estimation

- Classical choice for h:

$$h(t) = \frac{e^t}{1+e^t} \qquad \text{logit or logistic}$$

$$h(t) = F_{\mathcal{N}}(t) \qquad \text{probit}$$

$$h(t) = 1 - e^{-e^t} \qquad \text{log-log}$$

- Choice of the *best* β from the data.

- Need to specify the quality criterion...

Probabilistic Model

- By construction, $Y|\mathbf{X}$ follows $\mathcal{B}(\mathbb{P}\{Y=+1|\mathbf{X}\})$
- Approximation of $Y|\mathbf{X}$ by $\mathcal{B}(h(\beta^T\mathbf{X}+\beta_0))$
- *Natural* probabilistic choice for β: β minimizing the distance between $\mathcal{B}(h(X^t\beta))$ and $\mathcal{B}(\mathbb{P}\{Y=1|X\})$.

Probabilistic Model

- By construction, $Y|\mathbf{X}$ follows $\mathcal{B}(\mathbb{P}\{Y=+1|\mathbf{X}\})$
- Approximation of $Y|\mathbf{X}$ by $\mathcal{B}(h(\beta^T\mathbf{X}+\beta_0))$
- *Natural* probabilistic choice for β: β minimizing the distance between $\mathcal{B}(h(X^t\beta))$ and $\mathcal{B}(\mathbb{P}\{Y=1|X\})$.

KL Distance

- *Natural* distance: Kullback-Leibler divergence

$$\begin{aligned}
&\mathrm{KL}(\mathcal{B}(\mathbb{P}\{Y=1|X\}),\mathcal{B}(h(X^t\beta)) \\
&\quad= \mathbb{E}_X\left[\mathrm{KL}(\mathcal{B}(\mathbb{P}\{Y=1|X\}),\mathcal{B}(h(X^t\beta))\right] \\
&\quad= \mathbb{E}_X\left[\mathbb{P}\{Y=1|X\}\log\frac{\mathbb{P}\{Y=1|X\}}{h(X^t\beta)}\right. \\
&\qquad\qquad \left. +(1-\mathbb{P}\{Y=1|X\})\log\frac{1-\mathbb{P}\{Y=1|X\}}{1-h(X^t\beta)}\right]
\end{aligned}$$

log-likelihood

- KL:

$$\begin{aligned}
&\mathrm{KL}(\mathcal{B}(\mathbb{P}\{Y=1|X\}),\mathcal{B}(h(X^t\beta))\\
&\quad=\mathbb{E}_X\left[\mathbb{P}\{Y=1|X\}\log\frac{\mathbb{P}\{Y=1|X\}}{h(X^t\beta)}\right.\\
&\qquad\qquad\left.+(1-\mathbb{P}\{Y=1|X\})\log\frac{1-\mathbb{P}\{Y=1|X\}}{1-h(X^t\beta)}\right]\\
&\quad=\mathbb{E}_X\left[-\mathbb{P}\{Y=1|X\}\log(h(X^t\beta))\right.\\
&\qquad\qquad\left.-(1-\mathbb{P}\{Y=1|X\})\log(1-h(X^t\beta))\right]+C_{X,Y}
\end{aligned}$$

log-likelihood

- KL:

$$\begin{aligned}
&\mathrm{KL}(\mathcal{B}(\mathbb{P}\{Y=1|X\}),\mathcal{B}(h(X^t\beta))\\
&\quad=\mathbb{E}_X\left[\mathbb{P}\{Y=1|X\}\log\frac{\mathbb{P}\{Y=1|X\}}{h(X^t\beta)}\right.\\
&\qquad\qquad\left.+(1-\mathbb{P}\{Y=1|X\})\log\frac{1-\mathbb{P}\{Y=1|X\}}{1-h(X^t\beta)}\right]\\
&\quad=\mathbb{E}_X\left[-\mathbb{P}\{Y=1|X\}\log(h(X^t\beta))\right.\\
&\qquad\qquad\left.-(1-\mathbb{P}\{Y=1|X\})\log(1-h(X^t\beta))\right]+C_{X,Y}
\end{aligned}$$

- Empirical counterpart = opposite of the log-likelihood:

$$-\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{y_i=1}\log(h(x_i^t\beta))+\mathbf{1}_{y_i=-1}\log(1-h(x_i^t\beta))\right)$$

- Minimization of possible if h is regular...

Logistic Regression and Odd

- Logistic model: $h(t) = \frac{e^t}{1+e^t}$ (most *natural* choice...)
- The Bernoulli law $\mathcal{B}(h(t))$ satisfies then

$$\frac{\mathbb{P}\left\{Y=1\right\}}{\mathbb{P}\left\{Y=-1\right\}} = e^t \Leftrightarrow \log \frac{\mathbb{P}\left\{Y=1\right\}}{\mathbb{P}\left\{Y=-1\right\}} = t$$

- Interpretation in term of odd.
- Logistic model: linear model on the logarithm of the odd.

Associated Classifier

- Plugin strategy:

$$f_\beta(x) = \begin{cases} 1 & \text{if } \frac{e^{x^t\beta}}{1+e^{x^t\beta}} > 1/2 \Leftrightarrow x^t\beta > 0 \\ -1 & \text{otherwise} \end{cases}$$

Likelikood Rewriting

- Opposite of the log-likelihood:

$$
\begin{aligned}
& -\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{y_i=1}\log(h(x_i^t\beta)) + \mathbf{1}_{y_i=-1}\log(1-h(x_i^t\beta))\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{y_i=1}\log\frac{e^{x_i^t\beta}}{1+e^{x_i^t\beta}} + \mathbf{1}_{y_i=-1}\log\frac{1}{1+e^{x_i^t\beta}}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\log\left(1+e^{-y_i(x_i^t\beta)}\right)
\end{aligned}
$$

- Convex and smooth function of β
- Easy optimization.

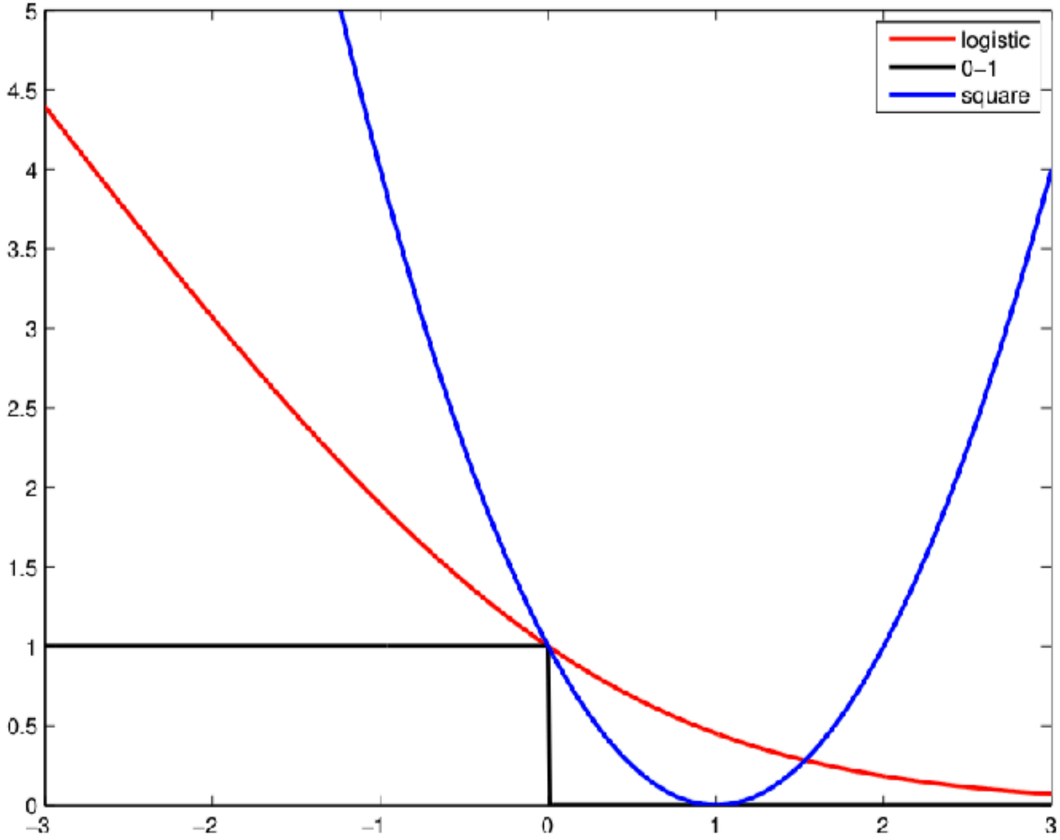
Risk Convexification Heuristic

- **Prop:** $\ell^{0/1}(y_i, f_\beta(x_i)) = \mathbf{1}_{y_i(x_i^t\beta)<0} \leq \frac{\log\left(1 + e^{-y_i(x_i^t\beta)}\right)}{\log 2}$
- Link between the empirical prediction loss and the likelihood:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{y_i \neq f_\beta(x_i)} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{y_i(x_i^t\beta)<0} \leq \frac{1}{n\log 2}\sum_{i=1}^{n}\log\left(1 + e^{-y_i(x_i^t\beta)}\right)$$

- Logistic: easy minimization of the right hand instead of the untractable left hand side...

Risk Convexification



$\ell(a, 1)$ for several classification losses

Logistic Coefficients

- Logistic regression entirely specified by β.
- Coefficientwise:
  - $\beta_i = 0$ means that the ith covariate is not used.
  - $\beta_i \sim 0$ means that the ith covariate as a low influence...

Simplified Logistic Models

- Enforce simplicity through a constraint on β!
- Support constraint: $\|\beta\|_0 = \sum_{i=1}^{d} \mathbf{1}_{\beta_i \neq 0} < C$
- Size constraint: $\|\beta\|_p < C$ with $1 \leq p$ (Often $p = 2$ or $p = 1$)

- **Rk:** $\|\beta\|_p$ is not scaling invariant if $p \neq 0$...
- Initial rescaling issue.

Constrained Optimization

- Choose a constant C.
- Compute β as

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^d, \|\beta\|_p \leq C} \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i(\beta^t x_i)})$$

Lagrangian Reformulation

- Choose λ and compute β as

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i(\beta^t x_i)}) + \lambda\|\beta\|_p^{p'}$$

  with $p' = p$ except if $p = 0$ where $p' = 1$.
- Easier calibration...

Penalized Likelihood

- Minimization of

$$\operatorname*{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i(\beta^t x_i)}) + \text{pen}(\beta)$$

where $\text{pen}(\beta)$ is a (sparsity promoting) penalty

- Variable selection if β is sparse.

Classical Penalties

- AIC: $\text{pen}(\beta) = \lambda \|\beta\|_0$ (non convex / sparsity)
- Ridge: $\text{pen}(\beta) = \lambda \|\beta\|_2^2$ (convex / no sparsity)
- Lasso: $\text{pen}(\beta) = \lambda \|\beta\|_1$ (convex / sparsity)
- Elastic net: $\text{pen}(\beta) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$ (convex / sparsity)

- Easy optimization if pen (and the loss) is convex...
- **Need to specify λ!**

Penalized Likelihood

- Minimization of

$$\frac{1}{n}\sum_{i=1}^{n}\log(1+e^{-y_i(\beta^t x_i)})+\text{pen}(\beta)$$

- Convex function in $\beta \in \mathbb{R}^d$!

Penalized Likelihood

- Minimization of

$$\frac{1}{n}\sum_{i=1}^{n}\log(1+e^{-y_i(\beta^t x_i)})+\operatorname{pen}(\beta)$$

- Convex function in $\beta \in \mathbb{R}^d$!

Practical Selection Methodology

- Choose a penalty shape $\widetilde{\operatorname{pen}}(\beta)$.
- Compute a CV error for a penalty $\lambda\widetilde{\operatorname{pen}}(\beta)$ for all $\lambda \in \Lambda$.
- Determine $\widehat{\lambda}$ the λ minimizing the CV error.
- Compute the final logistic regression with a penalty $\widehat{\lambda}\widetilde{\operatorname{pen}}(\beta)$.

Penalized Likelihood

- Minimization of

$$\frac{1}{n}\sum_{i=1}^{n}\log(1+e^{-y_i(\beta^t x_i)})+\text{pen}(\beta)$$

- Convex function in $\beta \in \mathbb{R}^d$!

Penalized Likelihood

- Minimization of

$$\frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i(\beta^t x_i)}) + \text{pen}(\beta)$$

- Convex function in $\beta \in \mathbb{R}^d$!

Convex Optimization

- A local minimum is a global minimum!
- No possibility to be trapped in a local minimum!
- Several very efficient minimization algorithm exists.
- Huge progress recently (motivated by big data...).
- Canonical algorithm: **(sub)gradient descent**.

Subgradient Descent Algorithm

- Start with a point θ_0
- for $k = 1, \ldots$ until *convergence* repeat:
  - $\theta^{k+1} \leftarrow \theta^k - \alpha_k \nabla f(\theta^k)$ where $\nabla f(\theta^k)$ is any subgradient of f at θ^k

Step/Learning Rate Choice

- Choice of α_k crucial!
- Provable convergence toward a minimum for suitable choice!

- Subject of a full course in the master!

Supervised Learning

Various approaches for Classification, a short review

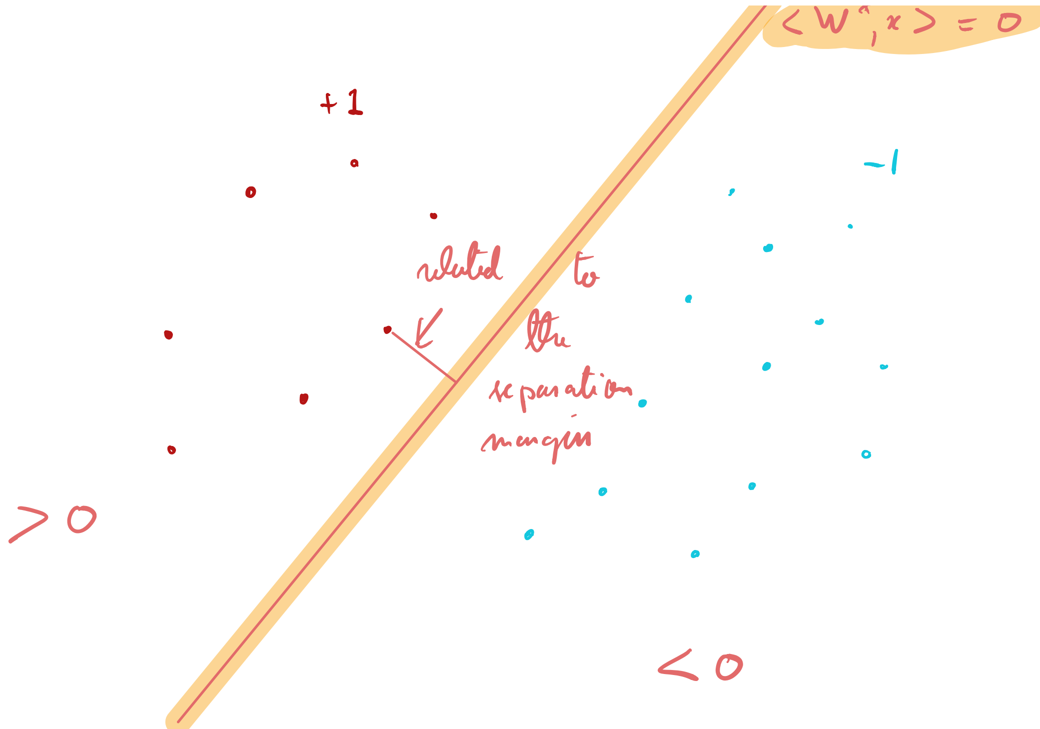
In the realizable case, there exists w^* such that $\forall i \in \{1, \ldots, m\}$, $y_i \langle w^*, x_i \rangle \geq 0$, and even such that $\forall i \in \{1, \ldots, m\}$, $y_i \langle w^*, x_i \rangle > 0$.

Then there exists $\bar{w} \in \mathbb{R}^d$ such that $\forall i \in \{1, \ldots, m\}$, $y_i \langle \bar{w}, x_i \rangle \geq 1$: if we can find one, we have an ERM.

Let $A \in \mathcal{M}_{m,d}(\mathbb{R})$ be defined by $A_{i,j} = y_i \, x_{i,j}$, and let $v = (1, \ldots, 1) \in \mathbb{R}^m$. Then any solution of the linear program

$$\max_{w \in \mathbb{R}^d} \langle 0, w \rangle \quad \text{subject to} \quad Aw \geq v$$

is an ERM. It can thus be computed in polynomial time.



<w*, x> = 0
+1
-1
related to the separation margin
> 0
< 0

Algorithm: Batch Perceptron

Data: training set $(x_1, y_1), \dots, (x_m, y_m)$

```
w_0 ← (0, ..., 0)
t ≥ 0
while ∃i_t : y_{i_t}⟨w_t, x_{i_t}⟩ ≤ 0 do
    w_{t+1} ← w_t + y_{i_t} x_{i_t} / ||x_{i_t}||
    t ← t + 1
return w_t
```

Each updates helps reaching the solution, since

$$y_{i_t}\langle w_{t+1}, x_{i_t}\rangle = y_{i_t}\left\langle w_t + y_{i_t}\frac{x_{i_t}}{\|x_{i_t}\|}, x_{i_t}\right\rangle = y_{i_t}\langle w_t, x_{i_t}\rangle + \|x_{i_t}\| \ .$$

Relates to a coordinate descent (stepsize does not matter).

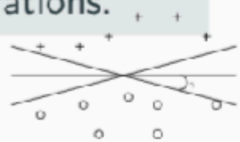
Convergence of Rosenblatt's Perceptron

Theorem

Assume that the dataset $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ is linearly separable and let the *separation margin* γ be defined as:

$$\gamma = \max_{w \in \mathbb{R}^d : \|w\|=1} \min_{1 \le i \le n} \frac{y_i \langle w, x_i \rangle}{\|x_i\|} .$$

Then the perceptron algorithm stops after at most $1/\gamma^2$ iterations.



Proof: Let w^* be such that $\forall 1 \le i \le m$, $\frac{y_i \langle w^*, x_i \rangle}{\|x_i\|} \ge \gamma$.

- If iteration t is necessary, then

$$\langle w^*, w_{t+1} - w_t \rangle = y_{i_t} \left\langle w^*, \frac{x_{i_t}}{\|x_{i_t}\|} \right\rangle \ge \gamma \quad \text{and hence } \langle w^*, w_t \rangle \ge \gamma t .$$

- If iteration t is necessary, then

$$\|w_{t+1}\|^2 = \left\| w_t + y_{i_t} \frac{x_{i_t}}{\|x_{i_t}\|} \right\|^2 = \|w_t\|^2 + \underbrace{\frac{2 y_{i_t} \langle w_t, x_{i_t} \rangle}{\|x_{i_t}\|}}_{\le 0} + y_{i_t}^2 \le \|w_t\|^2 + 1$$

  and hence $\|w_t\|^2 \le t$, or $\|w_t\| \le \sqrt{t}$.

- As a consequence, the algorithm iterates at least t times if

$$\gamma t \le \langle w^*, w_t \rangle \le \|w_t\| \le \sqrt{t} \quad \Longrightarrow \quad t \le \frac{1}{\gamma^2} .$$

In the worst case, the number of iterations can be exponentially large in the dimension d. Usually, it converges quite fast. If $\forall i, \|x_i\| = 1$, $\gamma = d(S, D)$ where $D = \{x : \langle w^*, x \rangle = 0\}$.

NP-hardness of computing the ERM for halfspaces

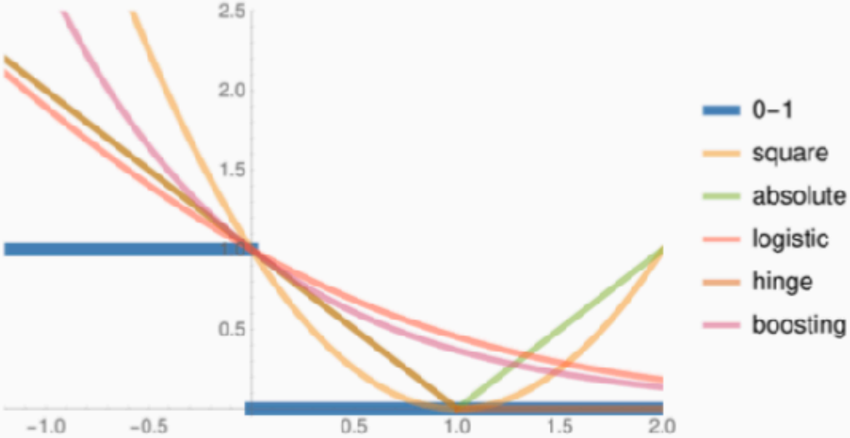
Computing an ERM in the agnostic case is NP-hard.

See *On the difficulty of approximately maximizing agreements*, by Ben-David, Eiron and Long.

Since the 0-1 loss

$$L_S(h_w) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{y_i \langle w, x_i \rangle < 0\}$$

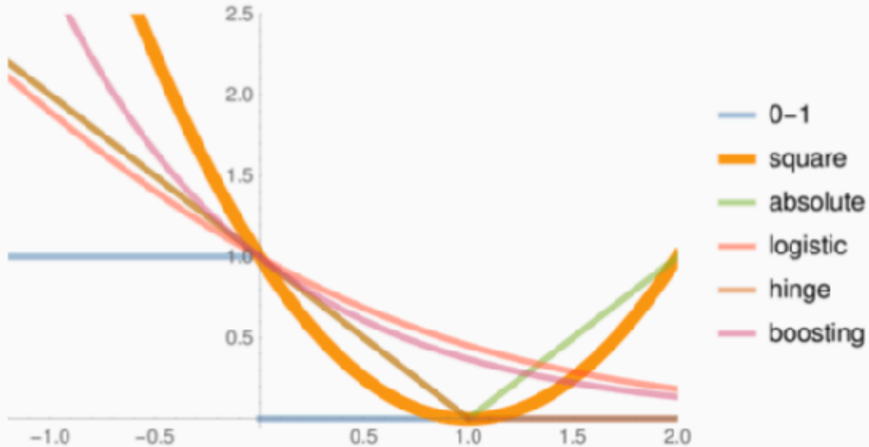
is intractable to minimize in the agnostic case, one may consider *surrogate* loss functions



$$L_S(h_w) = \frac{1}{m}\sum_{i=1}^{m} \ell(y_i \langle w, x_i \rangle) ,$$

where the loss function $\ell : \mathbb{R} \to \mathbb{R}^+$

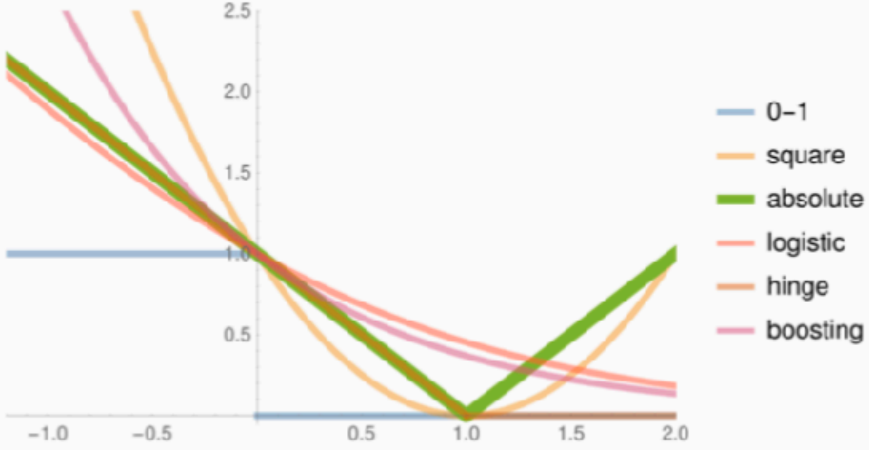
Linear regression with least squares:



$$L_S(h_w) = \frac{1}{m}\sum_{i=1}^{m}\left(h_w(x_i) - y_i\right)^2 = \frac{1}{m}\sum_{i=1}^{m}\left(1 - y_i\langle w, x_i\rangle\right)^2 .$$

If $X = (x_1, \ldots, x_m) \in \mathcal{M}_{m,d}(\mathbb{R})$ and $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$, one obtains $\hat{w} = (X^T X)^- X^T y$, where $A^- =$ generalized inverse of A.

Linear regression with absolute loss:



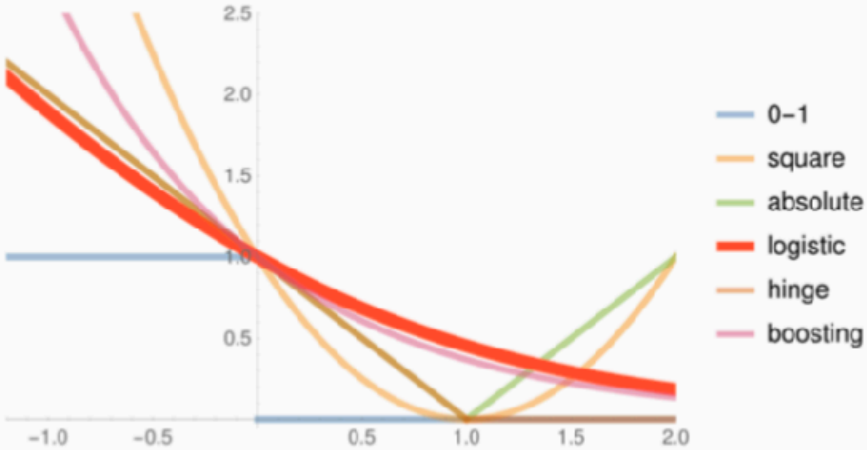
$$L_S(h_w) = \frac{1}{m}\sum_{i=1}^{m}\left|h_w(x_i) - y_i\right| = \frac{1}{m}\sum_{i=1}^{m}\left|1 - y_i h_w(x_i)\right| .$$

Can be solved by linear programming.

Interest: (statistical) robustness.

Statistics: "logistic regression":

$$P_w(Y = y|X = x) = \frac{1}{1 + \exp\left(-y\langle w, x\rangle\right)}$$

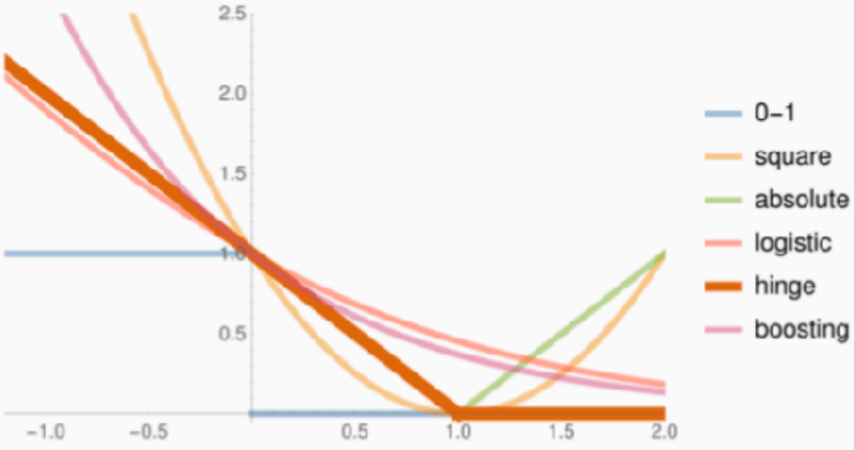


log with base 2 here so that $\ell(0) = 1$

$$L_S(h_w) = \frac{1}{m}\sum_{i=1}^{m} \log\left(1 + \exp(-y_i\langle w, x_i\rangle)\right),$$

convex minimization problem, can be solved by Newton's algorithm (in small dimension).

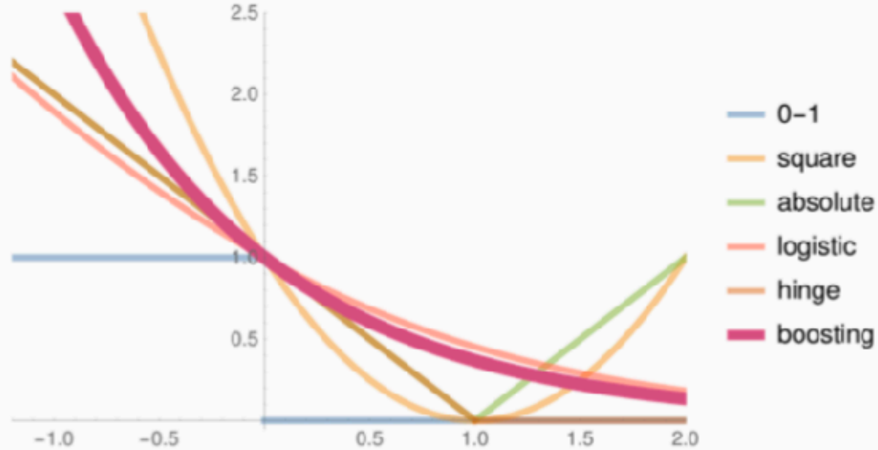
Margin maximization leads to



$$L_S(h_w) = \frac{1}{m} \sum_{i=1}^{m} \max\left\{0, 1 - y_i \langle w, x_i \rangle\right\} ,$$

convex but non-smooth minimization problem, used with a penalization term $\lambda \|w\|^2$: cf later.

Margin maximization leads to



$$L_S(h_w) = \frac{1}{m}\sum_{i=1}^{m} \exp\left(-y_i\langle w, x_i\rangle\right) ,$$

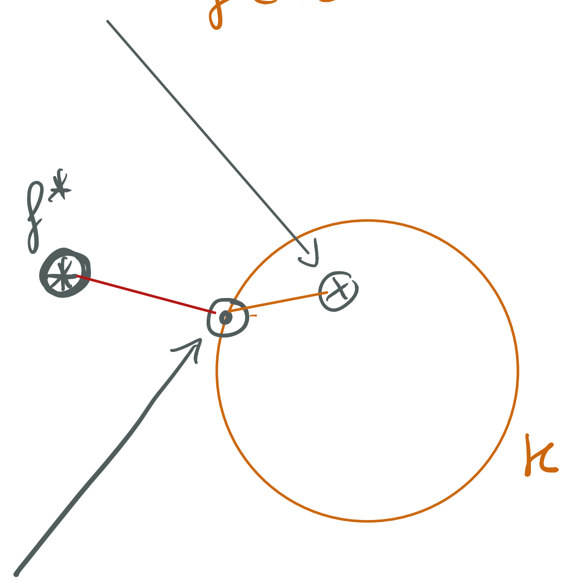
with ad-hoc optimization procedure – cf later.

$$R_n(f) = \mathbb{E}_n\left(\mathbb{1}_{Y_i = f(X_i)}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \neq f(X_i)}$$

$$\hat{f}_{\mathcal{K}} \in \underset{f \in \mathcal{K}}{\operatorname{argmin}}\ R_n(f)$$

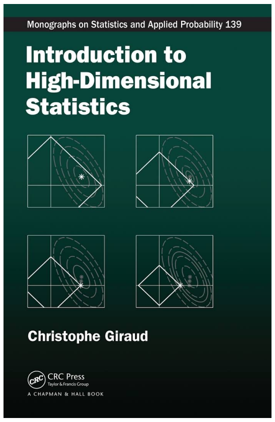
(Bayes Classifier) f^*

$$f^* \in \underset{f \text{ meas.}}{\operatorname{argmin}}\ R(f)$$



$$f^*_{\mathcal{K}} \in \underset{f \in \mathcal{K}}{\operatorname{argmin}}\ R(f)$$

$$0 \leq R(\hat{f}_{\mathcal{K}}) - R(f^*)$$
$$= \underbrace{R(f^*_{\mathcal{K}}) - R(f^*)}_{\text{approx. error}} + \underbrace{R(\hat{f}_{\mathcal{K}}) - R(f^*_{\mathcal{K}})}_{\text{stochastic error}}$$



Chap 9 p 186

Bound on stochastic error.

VC-dim of $\mathcal{H}$.