2.2. Empirical risk minimization

Empirical risk

- **Reminder:** we do not have access to data distribution

Definition: for fixed training data $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, we define the *empirical risk* of a predictor $f : \mathcal{X} \to \mathcal{Y}$ as

$$\hat{\mathcal{R}}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) \,.$$

- **Intuition:** good proxy for $\mathcal{R}$ if n is large enough:

$$\hat{\mathcal{R}}(f) \approx \mathcal{R}(f) \,.$$

Empirical risk minimization

- let $\mathcal{H}$ be a class of models
- ideally, we would like to find

$$f^\star \in \operatorname*{arg\,min}_{h \in \mathcal{H}} \mathcal{R}(h)\,.$$

- **Problem:** we do not know p... and even if we did it would still be a very difficult problem
- **Idea:** replace $\mathcal{R}$ by the empirical risk
- this leads to empirical risk minimization (ERM):[2]

$$\hat{f} \in \operatorname*{arg\,min}_{f \in \mathcal{H}} \hat{\mathcal{R}}(f) = \operatorname*{arg\,min}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))\,.$$

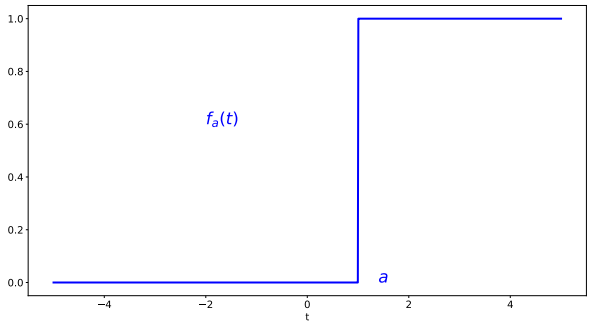
[2] Vapnik, *Principles of risk minimization for learning theory*, NIPS, 1991

Empirical risk minimization: example

- let us give a simple example
- take $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$, 0-1 loss, and "bump functions:"

$$\mathcal{H} = \{f_a : \mathbb{R} \to \mathbb{R}, \forall t \in \mathbb{R}, f_a(t) = \mathbb{1}_{t \geq a}\}.$$

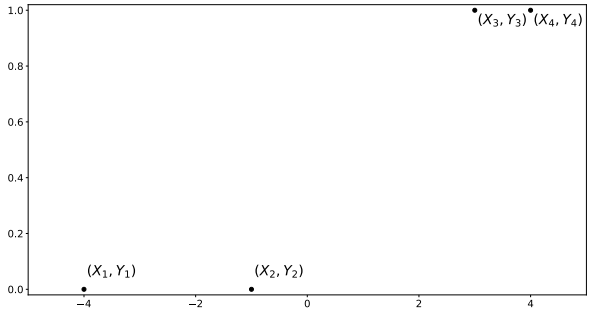
- **Visually,** elements of $\mathcal{H}$ look like:



Empirical risk minimization: example

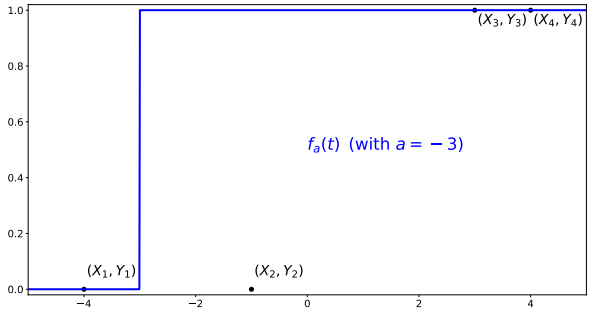
- take the following datapoints:

$$(X_1, Y_1) = (-4, 0),\ (X_2, Y_2) = (-1, 0),\ (X_3, Y_3) = (3, 1),\ (X_4, Y_4) = (4, 1)\,.$$



Empirical risk minimization: example

- for each candidate f_a, we can compute the associated empirical risk:

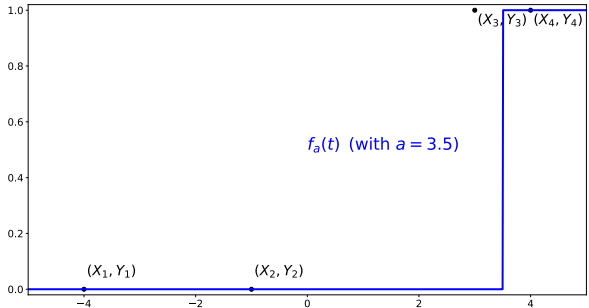


- here we have

$$\hat{\mathcal{R}}(f_a) = \frac{1}{4}(0 + 1 + 0 + 0) = \frac{1}{4}\,.$$

Empirical risk minimization: example

- for each candidate f_a, we can compute the associated empirical risk:

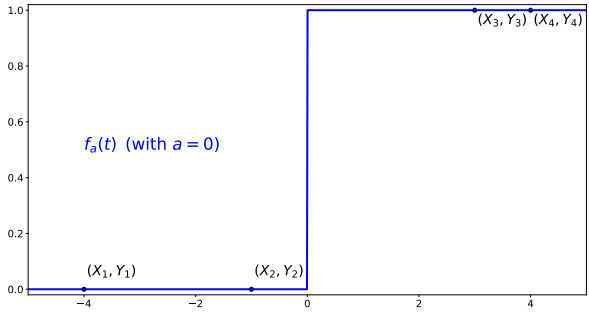


- here we have

$$\hat{\mathcal{R}}(f_a) = \frac{1}{4}(0 + 0 + 1 + 0) = \frac{1}{4}\,.$$

Empirical risk minimization: example

- we notice that several candidates achieve empirical risk $= 0$:

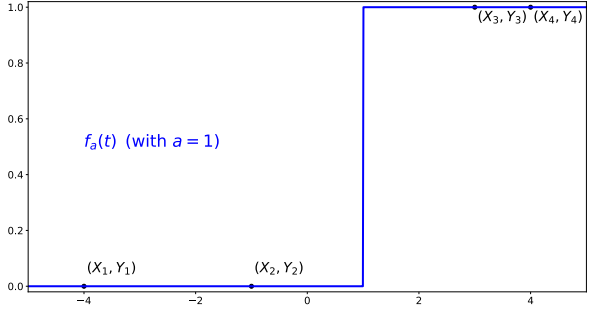


- here we have

$$\hat{\mathcal{R}}(f_a) = \frac{1}{4}(0 + 0 + 0 + 0) = 0\,.$$

Empirical risk minimization: example

- we notice that several candidates achieve empirical risk $= 0$:

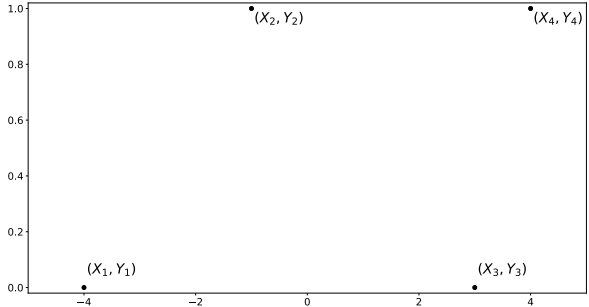


- here we have

$$\hat{\mathcal{R}}(f_a) = \frac{1}{4}(0 + 0 + 0 + 0) = 0\,.$$

Empirical risk minimization: example

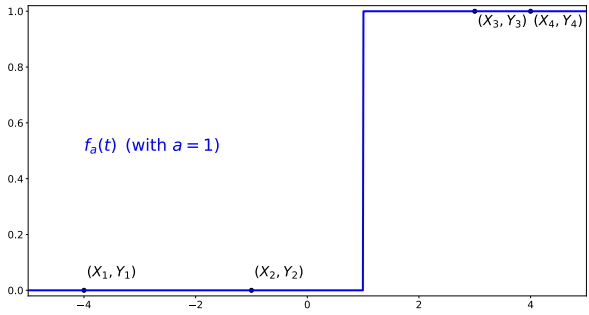
- f_a with $a \in (-1, 3)$ are all **empirical risk minimizers**
- we can pick any of them
- not always the case:



- **Question:** can you find a candidate with empirical risk $= 0$?

Generalization

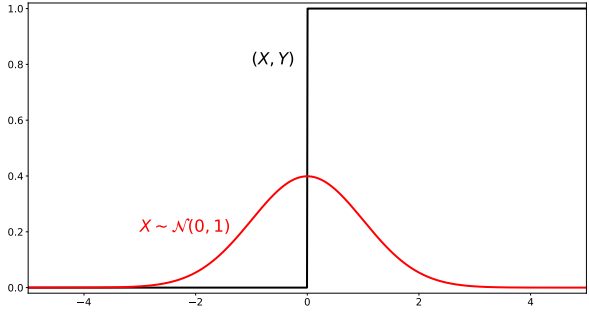
- back to the "separable" case:



- **Question:** does $\hat{\mathcal{R}}(f) = 0$ say something about $\mathcal{R}(f)$?

Generalization

- **Answer:** it depends (on the true data distribution)
- **Example:** assume $X \sim \mathcal{N}(0,1)$, and $Y = \mathbb{1}_{X \geq 0}$



- we can compute the (true) risk for different candidates

Generalization

- **Example:**

$$\begin{aligned}
\mathcal{R}(f_1) &= \mathbb{P}\left(f_1(X) \neq Y\right) && \text{(definition of the risk)} \\
&= \mathbb{P}\left(\mathbb{1}_{X \geq 1} \neq \mathbb{1}_{X \geq 0}\right) && \text{(definition of } f_a \text{ and data distribution)} \\
&= \mathbb{P}\left(X \in [0,1]\right) \\
&= \frac{1}{\sqrt{2\pi}} \int_0^1 \mathrm{e}^{\frac{-x^2}{2}} \mathrm{d}x && \text{(density of a } \mathcal{N}(0,1)\text{)} \\
\mathcal{R}(f_1) &\approx 0.34
\end{aligned}$$

- this is not zero!
- one predictor, though, has zero risk in that case: f_0
- it is the **Bayes predictor**

Overfitting

- **Problem:** in extreme cases, this can be a severe issue
- this is in particular true when the hypotheses class $\mathcal{H}$ is too large
- **Example:** assume $\mathcal{H}$ is the set of all measurable functions
- consider a fixed training set (x_i, y_i) and let

$$h(x) = \begin{cases} y_i & \text{if } \exists i \in \{1, \dots, n\} \text{ s.t. } x = x_i \\ 0 & \text{otherwise.} \end{cases}$$

- in particular, $h \in \mathcal{H}$ (since $\mathcal{H}$ contains all functions), and

$$\forall i \in [n], \quad h(x_i) = y_i \,.$$

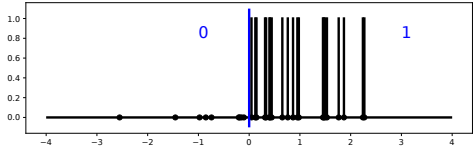
- in that case,

$$\hat{\mathcal{R}}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{h(x_i) \neq y_i} = 0 \,.$$

- empirical risk $= 0$ (interpolating)

Overfitting, ctd.

- **As in the previous example:** assume $Y = \mathbb{1}_{X \geq 0}$ and $X \sim \mathcal{N}(0, 1)$
- h looks like:



- since X has a density, $\mathbb{P}(X = x_i) = 0$
- thus we will always predict 0 on new datapoints
- let us compute the true risk:

$$\mathcal{R}(h) = \mathbb{P}(h(X) \neq Y) = \mathbb{P}(0 \neq \mathbb{1}_{X \geq 0}) = 1/2\,.$$

- this is essentially the **worst we can get**, despite having 0 training error

How to prevent overfitting?

- **Solution I:** reduce size of $\mathcal{H}$
- typical situation: parameterized space $f_\theta : \mathcal{X} \to \mathcal{Y}$, with $\theta \in \Theta$
- in this situation, ERM becomes

$$\hat{\theta} \in \operatorname*{arg\,min}_{\theta \in \Theta} \hat{\mathcal{R}}(f_\theta) = \operatorname*{arg\,min}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i))$$

- we can control the number of parameters
- **Solution II:** regularize (not exclusive), that is, minimize

$$\hat{\mathcal{R}}(f_\theta) + \lambda \Omega(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i)) + \lambda \Omega(\theta)\,.$$

- **Example:** $\Omega(\theta) = \lambda \|\theta\|^2$ with $\lambda > 0$ some hyperparameter

Empirical risk minimization: summary

- **Pros:**
  - general framework
  - can be solved approximately when $\mathcal{H}$ is parameterized
- **Cons:**
  - non-separable data
  - non-convexity $\rightarrow$ optimization problem can be hard
  - overfitting
- **Other approaches:** local averaging
- **Idea:** we know $\mathbb{E}[Y \mid X = x]$ or $\mathbb{P}(Y = 1 \mid X = x)$ are "the best we can do"
- $\rightarrow$ let us approximate them directly
- typical example = k-nearest neighbors[3]

[3] Fix, Hodges, *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*, USAF report, 1951

3. Linear least-square regression