Machine Learning

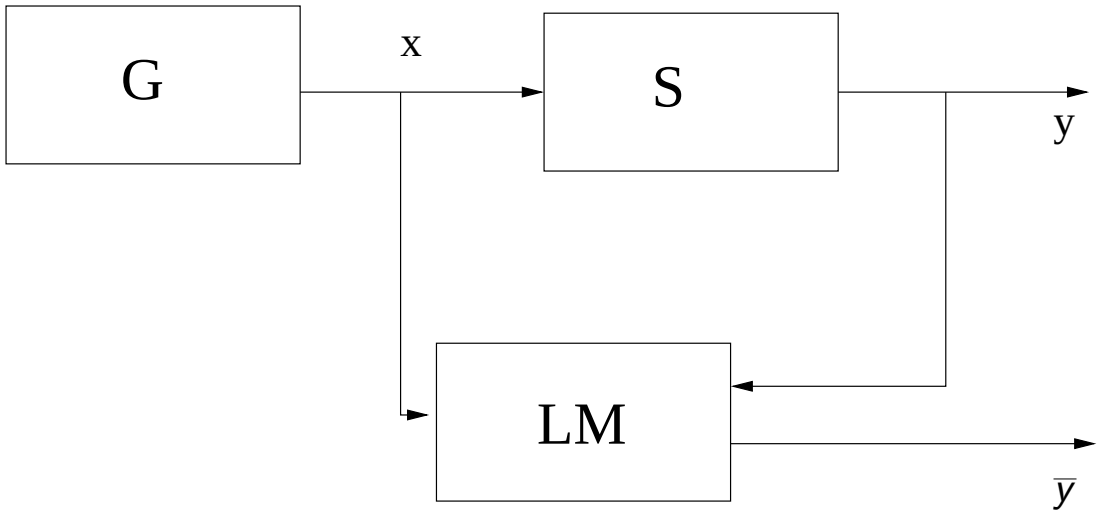
Volker Roth

Department of Mathematics & Computer Science
University of Basel

Section 6

Elements of Statistical Learning Theory

A 'black box' model of learning



G Generates i.i.d. samples $\boldsymbol{x}$ according to unknown pdf $p(\boldsymbol{x})$.

S Outputs values y according to unknown $p(y|\boldsymbol{x})$.

LM Observes pairs $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \sim p(\boldsymbol{x}, y) = p(\boldsymbol{x})p(y|\boldsymbol{x})$.
Tries to capture the relation between $\boldsymbol{x}$ and y.

Expected Risk

Learning Process: frequentist view

The learning process is the process of choosing an **appropriate** function from a **given set** of functions.

Note: from a **Bayesian** viewpoint we would rather define a **distribution over functions.**
A good function should incur only a few errors $\rightsquigarrow$ small expected risk:

Expected Risk

The quantity

$$R[f] = E_{(\boldsymbol{x},y)\sim p}\{\text{Loss}(y, f(\boldsymbol{x}))\}$$

is called the expected risk and measures the loss averaged over the unknown distribution.

Empirical Risk

- The best possible risk is $\inf_f R[f]$. The infimum is often achieved at a minimizer f_ρ that we call **target function**.
- ...but we just have a sample S. To find "good" functions we typically have to restrict ourselves to a **hypothesis space** $\mathcal{H}$ containing functions with some property (regularity, smoothness etc.)
- In a given hypothesis space $\mathcal{H}$, denote by f^* the **best possible function** that can be **implemented** by the learning machine.

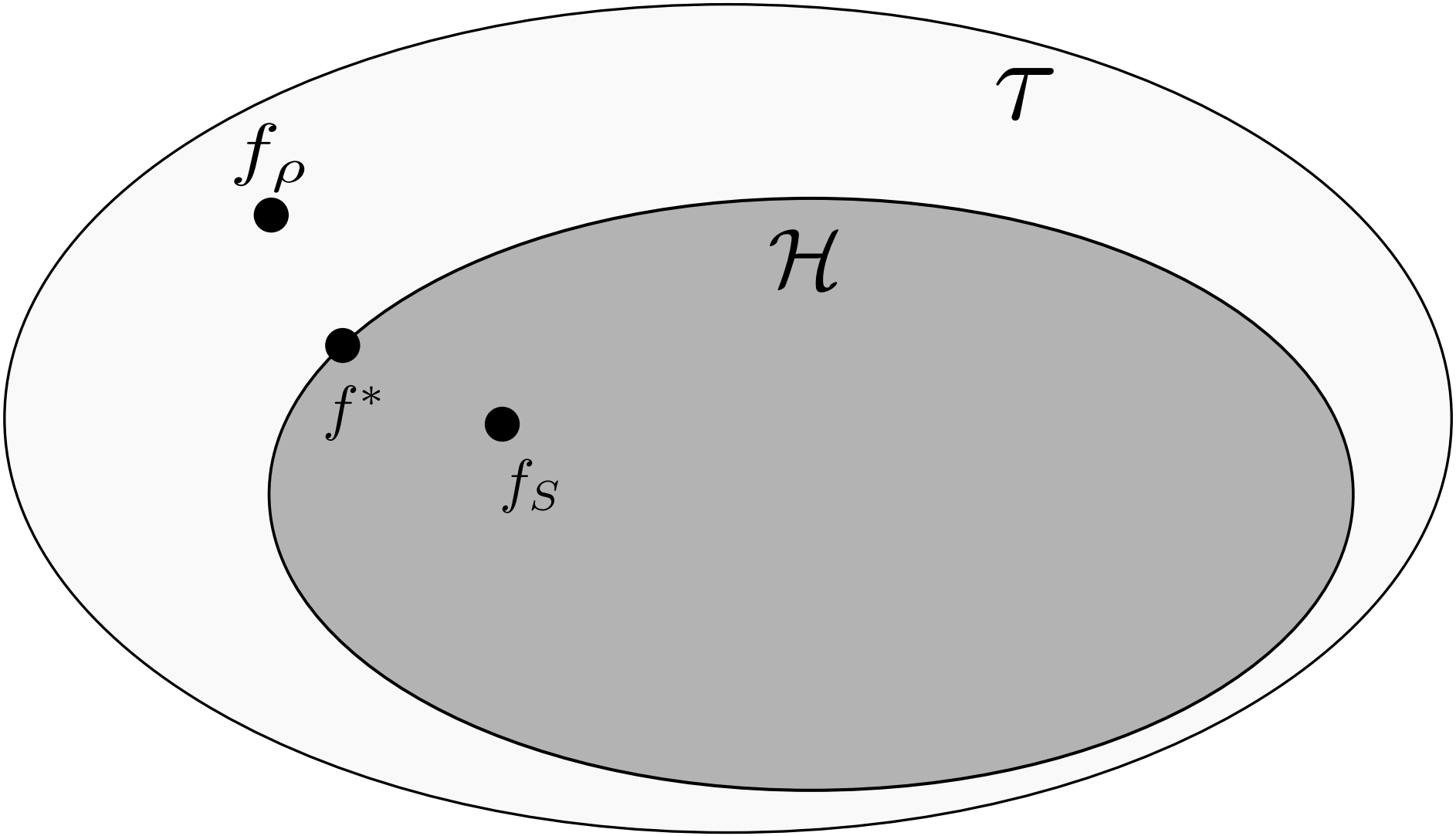
Empirical risk

The empirical risk of a function f is

$$R_{\text{emp}}[f] = \frac{1}{n} \sum_{i=1}^{n} \text{Loss}(y_i, f(\boldsymbol{x}_i)).$$

Denote by $f_S \in \mathcal{H}$ the **empirical risk minimizer** on sample S.

Hypothesis space



Generalization

- SLT gives results for bounding the error on the **unseen** data, given only the training data.
- There needs to be a relation that **couples past and future**.

Sampling Assumption

The only assumption of SLT is that all samples (past and future) are iid.

Typical structure of a bound: With probability $1 - \delta$ it holds that

$$R[f_n] \leq \overbrace{R_{\text{emp}}[f_n]}^{\text{known}} + \underbrace{\sqrt{\frac{a}{n}\left(\text{capacity}(\mathcal{H}) + \ln\frac{b}{\delta}\right)}}_{\text{confidence term}},$$

with some constants $a, b > 0$.

Convergence of random variables

Definition (Convergence in Probability)

Let $X_1, X_2, \ldots$ be random variables. We say that X_n **converges in probability** to the random variable X as $n \to \infty$, iff, for all $\varepsilon > 0$,

$$P(|X_n - X| > \varepsilon) \to 0, \text{ as } n \to \infty.$$

We write $X_n \xrightarrow{p} X$ as $n \to \infty$.

The simplest case: Binary classification

Binary classification with 0/1 loss

Our analysis considers the case where

$$f : X \to \{-1, 1\}, \quad \text{and } L(y, f(\boldsymbol{x})) = \frac{1}{2}|1 - f(\boldsymbol{x})y|.$$

- Note: We can use any hypothesis space and apply the signum function:

$$\mathcal{H}' = \{f' = \text{sign}(f) | f \in \mathcal{H}\}$$

- Similar results from SLT also available other classification loss functions and for regression (but we will not discuss this here).

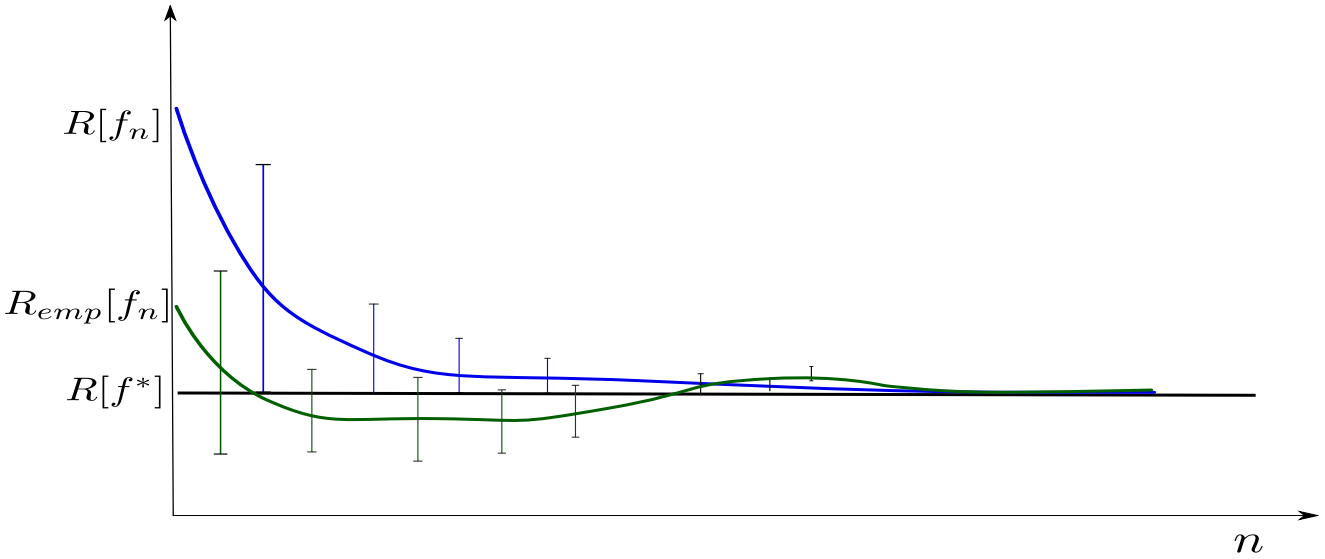
Consistency of ERM

The principle of empirical risk minimization is consistent if for any $\epsilon > 0$,

$$\lim_{n \to \infty} P(|R[f_n] - R[f^*]| > \epsilon) = 0$$

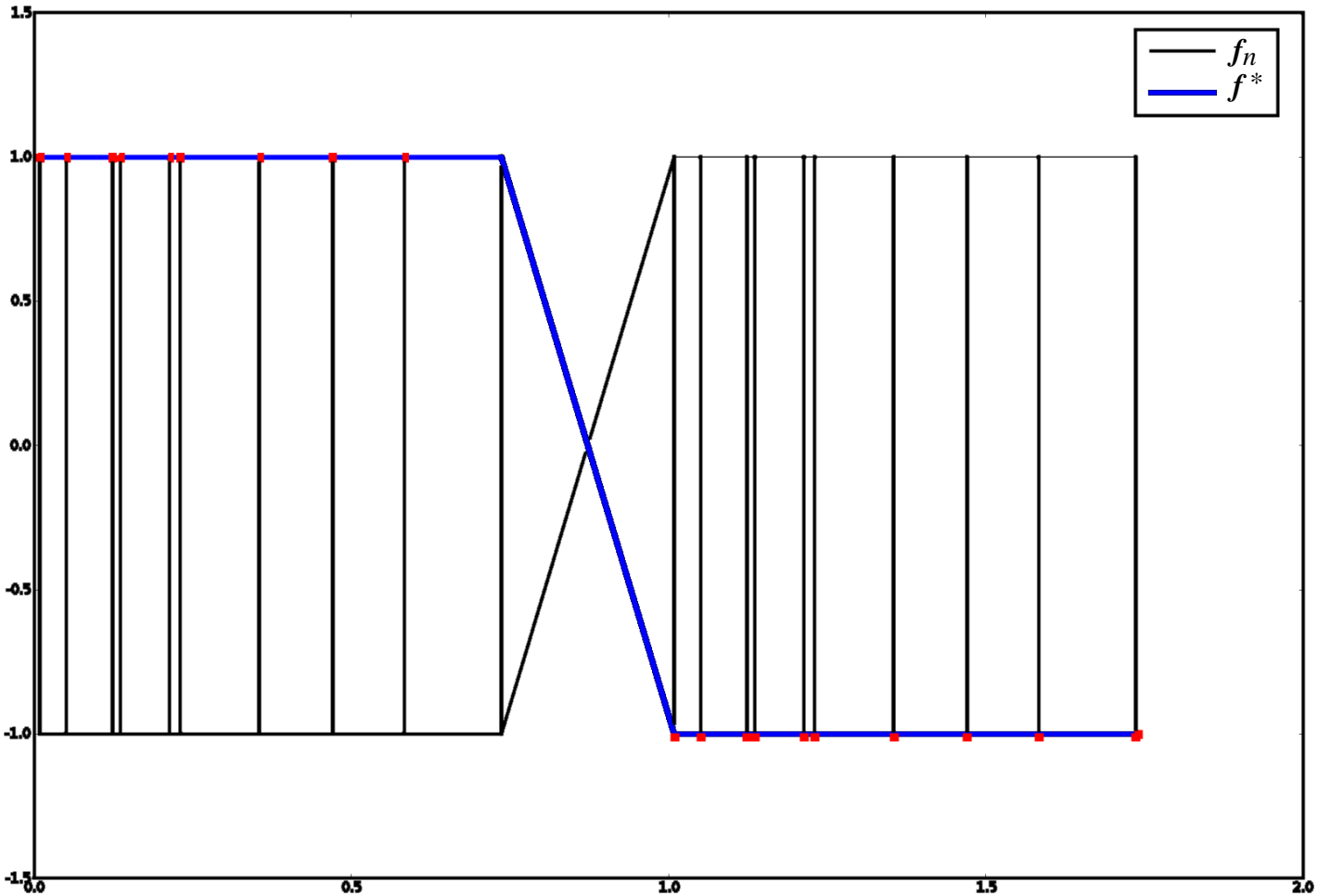
and

$$\lim_{n \to \infty} P(|R_{\text{emp}}[f_n] - R[f_n]| > \epsilon) = 0$$



A counter example

Why is bounding $P(|R_{\text{emp}}[f_n] - R[f^*]| > \epsilon)$ not sufficient?



Hoeffding's inequality

Theorem (Hoeffding)

Let $\xi_i, i \in [0, n]$ be n independent instances of a bounded random variable ξ, with values in $[a, b]$. Denote their average by $Q_n = \frac{1}{n}\sum_i \xi_i$. Then for any $\epsilon > 0$,

$$\left.\begin{array}{l} P(Q_n - E(\xi) \geq \epsilon) \\ P(E(\xi) - Q_n \geq \epsilon) \end{array}\right\} \leq \exp(-\frac{2n\epsilon^2}{(b-a)^2}) \tag{1}$$

and

$$P(|Q_n - E(\xi)| \geq \epsilon) \leq 2\exp(-\frac{2n\epsilon^2}{(b-a)^2}) \tag{2}$$

Hoeffding's inequality

Let ξ be the 0/1 loss function:

$$\xi = \frac{1}{2}|1 - f(\boldsymbol{x})y| = L(y, f(\boldsymbol{x})).$$

Then

$$Q_n[f] = \frac{1}{n}\sum_{i=1}^{n}\xi_i = \frac{1}{n}\sum_{i=1}^{n}L(y_i, f(\boldsymbol{x}_i)) = R_{\text{emp}}[f]$$

and

$$E[\xi] = E[L(y, f(\boldsymbol{x}))] = R[f].$$

I.i.d. sampling assumption: ξ_i are independent instances of bounded random variable ξ, with values in $[0, 1]$.

Hoeffding's Inequality for fixed functions

$$P(|R_{\text{emp}}[f] - R[f]| > \epsilon) \leq 2\exp(-2n\epsilon^2)$$

Hoeffding's inequality

Hoeffding's inequality gives us rates of convergence for any fixed function.

Example: Let $f \in \mathcal{H}$ be an arbitrary fixed function

- For $\epsilon = 0.1$ and $n = 100$,

$$P(|R_{\text{emp}}[f] - R[f]| > 0.1) \leq 0.28$$

- For $\epsilon = 0.1$ and $n = 200$,

$$P(|R_{\text{emp}}[f] - R[f]| > 0.1) \leq 0.04$$

Caution!

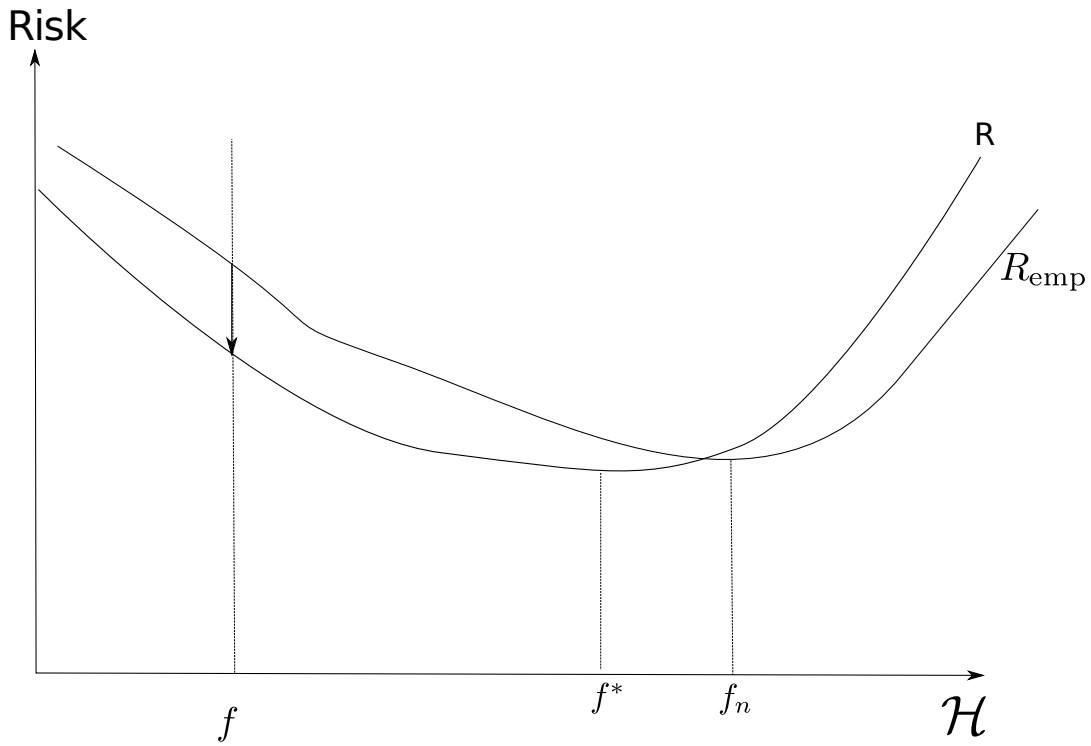
Hoeffding's inequality does **not** tell us that

$$P(|R_{\text{emp}}[f_n] - R[f_n]| > \epsilon) \leq 2\exp(-2n\epsilon^2).$$

Because:

- f_n is chosen to minimize R_{emp}.
- This is not a fixed function!!

Consistency



For each **fixed** function f, $R_{\text{emp}}[f] \xrightarrow[n\to\infty]{P} R[f]$ (downward arrow).

This does not mean that the empirical risk minimizer f_n will lead to a value of risk that is as good as possible, $R[f^*]$ (consistency).

Conditions for consistency

Let

$$f_n := \arg\min_{f \in \mathcal{H}} R_{\text{emp}}[f]$$

$$f^* := \arg\min_{f \in \mathcal{H}} R[f]$$

then

$$R[f] - R[f^*] \geq 0, \forall f \in \mathcal{H}$$

$$R_{\text{emp}}[f] - R_{\text{emp}}[f_n] \geq 0, \forall f \in \mathcal{H}$$

Conditions for consistency

Let

$$f_n := \arg\min_{f \in \mathcal{H}} R_{\text{emp}}[f]$$
$$f^* := \arg\min_{f \in \mathcal{H}} R[f]$$

then

$$R[f_n] - R[f^*] \geq 0,$$
$$R_{\text{emp}}[f] - R_{\text{emp}}[f_n] \geq 0, \; \forall f \in \mathcal{H}$$

Conditions for consistency

Let

$$f_n := \arg\min_{f \in \mathcal{H}} R_{\text{emp}}[f]$$

$$f^* := \arg\min_{f \in \mathcal{H}} R[f]$$

then

$$R[f_n] - R[f^*] \geq 0, \forall f \in \mathcal{H}$$

$$R_{\text{emp}}[f^*] - R_{\text{emp}}[f_n] \geq 0, \forall f \in \mathcal{H}$$

Conditions for consistency

$$
\begin{aligned}
0 &\leq \overbrace{R[f_n] - R[f^*]}^{\geq 0} + \overbrace{R_{\text{emp}}[f^*] - R_{\text{emp}}[f_n]}^{\geq 0} \\
&= R[f_n] - R_{\text{emp}}[f_n] + R_{\text{emp}}[f^*] - R[f^*] \\
&\leq \underbrace{\sup_{f \in \mathcal{H}} \left(R[f] - R_{\text{emp}}[f] \right)}_{\text{Assumption: } \xrightarrow[n \to \infty]{P} 0} + \underbrace{R_{\text{emp}}[f^*] - R[f^*]}_{\text{Hoeffding: } \xrightarrow[n \to \infty]{P} 0}
\end{aligned}
$$

Assume

$$\sup_{f \in \mathcal{H}} \left(R[f] - R_{\text{emp}}[f] \right) \xrightarrow[n \to \infty]{P} 0$$

One-sided uniform convergence over all functions in $\mathcal{H}$

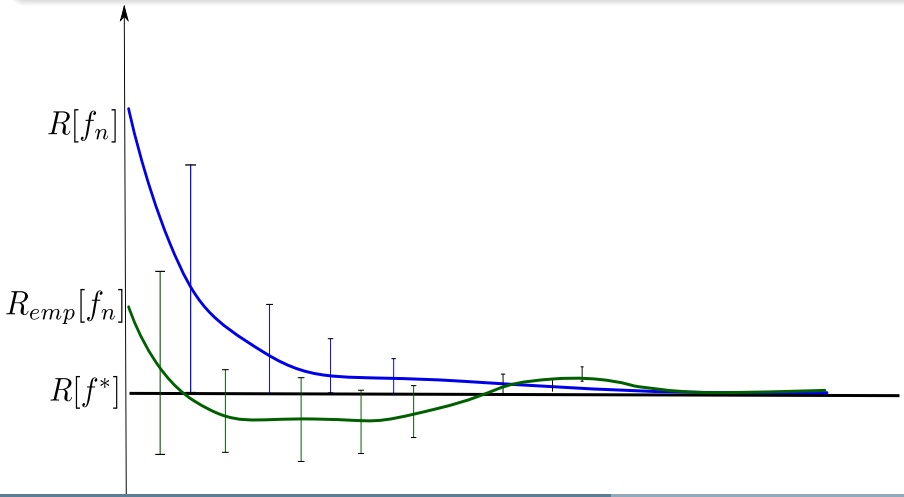
Conditions for consistency

$$0 \leq \overbrace{R[f_n] - R[f^*]}^{\geq 0} + \overbrace{R_{\text{emp}}[f^*] - R_{\text{emp}}[f_n]}^{\geq 0} \xrightarrow[n\to\infty]{P} 0$$

$$- R[f^*] + R_{\text{emp}}[f^*] \xrightarrow[n\to\infty]{P} 0$$

$$R[f_n] - \qquad\qquad R_{\text{emp}}[f_n] \xrightarrow[n\to\infty]{P} 0$$

$\sup_{f\in\mathcal{H}}(R[f] - R_{\text{emp}}[f]) \xrightarrow[n\to\infty]{P} 0 \Rightarrow$ **consistency** of ERM.
Thus, it is a sufficient condition for consistency.



The key theorem of learning theory

Theorem (Vapnik & Chervonenkis '98)

Let $\mathcal{H}$ be a set of functions with bounded loss for the distribution $F(x, y)$,

$$A \leq R[f] \leq B, \ \forall f \in \mathcal{H}.$$

For the ERM principle to be consistent, it is necessary and sufficient that

$$\lim_{n \to \infty} P(\sup_{f \in \mathcal{H}}(R[f] - R_{emp}[f]) > \epsilon) = 0, \ \forall \epsilon > 0.$$

Note: here, we looked only at the sufficient condition for consistency.
For the necessary condition see (Vapnik & Chervonenkis '98).

The key theorem of learning theory

- The key theorem asserts that any analysis of the convergence of ERM must be a **worst case analysis.**
- We will show:
  Consistency depends on the capacity of the hypothesis space.

But there are some open questions:

- How can we check the condition for the theorem (uniform one-sided convergence) in practice?
- Are there "simple" hypothesis classes with guaranteed consistency?
- Analysis is still asymptotic.
  What can we say about finite sample sizes?

Finite hypothesis spaces

Assume the set $\mathcal{H}$ contains only 2 functions:

$$\mathcal{H} = \{f_1, f_2\}.$$

Let

$$C_\epsilon^i := \{(x_1, y_1), \ldots, (x_n, y_n) | R[f_i] - R_{\text{emp}}[f_i] > \epsilon\}$$

be the set of samples for which the risks of f_i differ by more than ϵ.

Hoeffding's inequality:

$$P(C_\epsilon^i) \leq \exp(-2n\epsilon^2)$$

Union bound:

$$P(\sup_{f \in \mathcal{H}}(R[f] - R_{\text{emp}}[f] > \epsilon)) = P(C_\epsilon^1 \cup C_\epsilon^2) = P(C_\epsilon^1) + P(C_\epsilon^2) - P(C_\epsilon^1 \cap C_\epsilon^2)$$
$$\leq P(C_\epsilon^1) + P(C_\epsilon^2) \leq 2\exp(-2n\epsilon^2).$$

Finite hypothesis spaces

Assume $\mathcal{H}$ contains a finite number of functions: $\mathcal{H} = \{f_1, \ldots, f_N\}$.

$$C_\epsilon^i := \{(x_1, y_1), \ldots, (x_n, y_n) | R[f_i] - R_{\text{emp}}[f_i] > \epsilon\}$$

Hoeffding's inequality: $P(C_\epsilon^i) \leq \exp(-2n\epsilon^2)$

Union bound: $P(\cup_{i=1}^N C_\epsilon^i) \leq \sum_{i=1}^N P(C_\epsilon^i) \leq N \exp(-2n\epsilon^2)$

$$P(\sup_{f \in \mathcal{H}}(R[f] - R_{\text{emp}}[f] > \epsilon)) \leq N \exp(-2n\epsilon^2) = \exp(\ln N - 2n\epsilon^2)$$

- For any finite hypothesis space, the ERM is consistent.
- The convergence is exponentially fast.

Some consequences

$$P(\sup_{f \in \mathcal{H}} R[f] - R_{\text{emp}}[f] > \epsilon) \leq \exp(\ln N - 2n\epsilon^2)$$

- Bound holds uniformly for all functions in $\mathcal{H}$
  $\rightsquigarrow$ can use it for the functions that minimize R_{emp}.
  $\rightsquigarrow$ We can **bound the test error:**

$$P(R[f_n] - R_{\text{emp}}[f_n] > \epsilon) \leq \exp(\ln N - 2n\epsilon^2).$$

Some consequences

- Can derive a **confidence interval:** equate r.h.s. to δ and solve for ϵ:

$$P(R[f_n] - R_{\text{emp}}[f_n] > \epsilon) \leq \delta(\epsilon)$$

$$P(R[f_n] - R_{\text{emp}}[f_n] \leq \epsilon) \geq 1 - \delta(\epsilon)$$

- With probability at least $(1 - \delta)$ it holds that

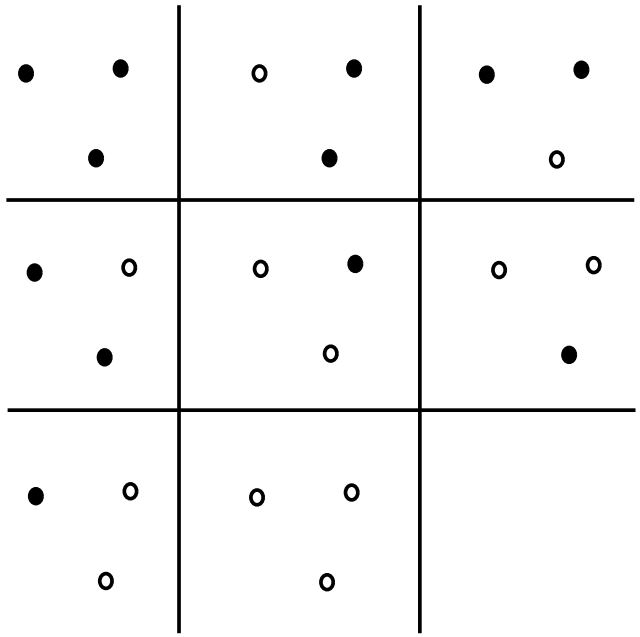
$$R[f_n] \leq R_{\text{emp}}[f_n] + \epsilon(\delta)$$

$$R[f_n] \leq R_{\text{emp}}[f_n] + \sqrt{\frac{a}{n}\Big(\underbrace{\ln N}_{\text{Capacity}(\mathcal{H})} + \ln\frac{b}{\delta}\Big)}, \quad \text{with } a = 1/2, b = 1.$$

- Bound depends only on $\mathcal{H}$ and n.
- However: "Simple" spaces (like the space of linear functions) contain **infinitely many functions.**

Infinite to finite (?)

Observation: $R_{\text{emp}}[f]$ effectively refers only to a **finite** function class: for n sample points $x_1, \ldots, x_n$, the functions in $f \in \mathcal{H}$ can take at most 2^n different values $y_1, \ldots, y_n$.



But this does not yet solve our problem: Confidence term $\ln(2^n)/n = \ln 2$ does not converge to 0 as $n \to \infty$. But let's formalize this idea first...

Infinite case: Shattering Coefficient

Let a sample: $Z_n := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be given.
Denote by $\mathcal{N}(\mathcal{H}, Z_n)$ the cardinality of $\mathcal{H}$ **when restricted to** $\{x_1, \ldots, x_n\}$, $\mathcal{H}|Z_n$, i.e. the **number of functions** from $\mathcal{H}$ that can be **distinguished on the given sample.**
Consider now the maximum (over all possible n-samples):

Definition (Shattering Coefficient)

The *Shattering Coefficient* is the maximum number of ways into which n points can be classified by the function class:

$$\mathcal{N}(\mathcal{H}, n) = \max_{Z_n} \mathcal{N}(\mathcal{H}, Z_n).$$

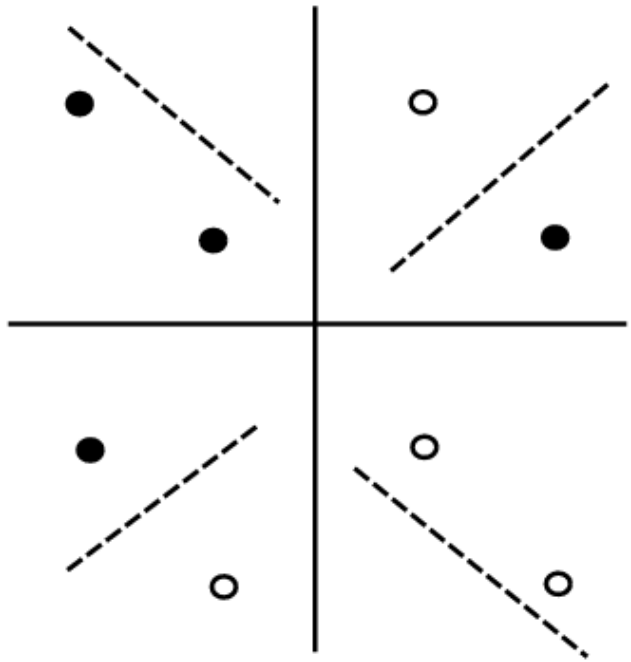
Since $f(x) \in \{-1, 1\}$, $\mathcal{N}(\mathcal{H}, n)$ is finite.

$$\mathcal{N}(\mathcal{H}, Z_n) \leq \mathcal{N}(\mathcal{H}, n) \leq 2^n$$

Example

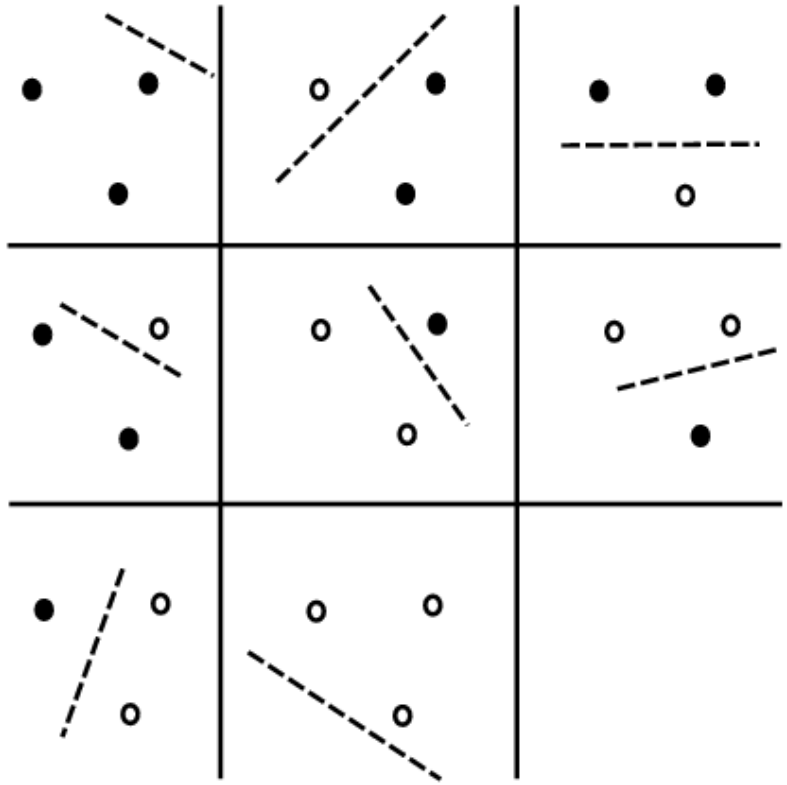
Linear functions

$$\mathcal{H} = \{\text{sign}(\langle w, x \rangle + b) | w \in \mathbb{R}^2, b \in \mathbb{R}\}$$



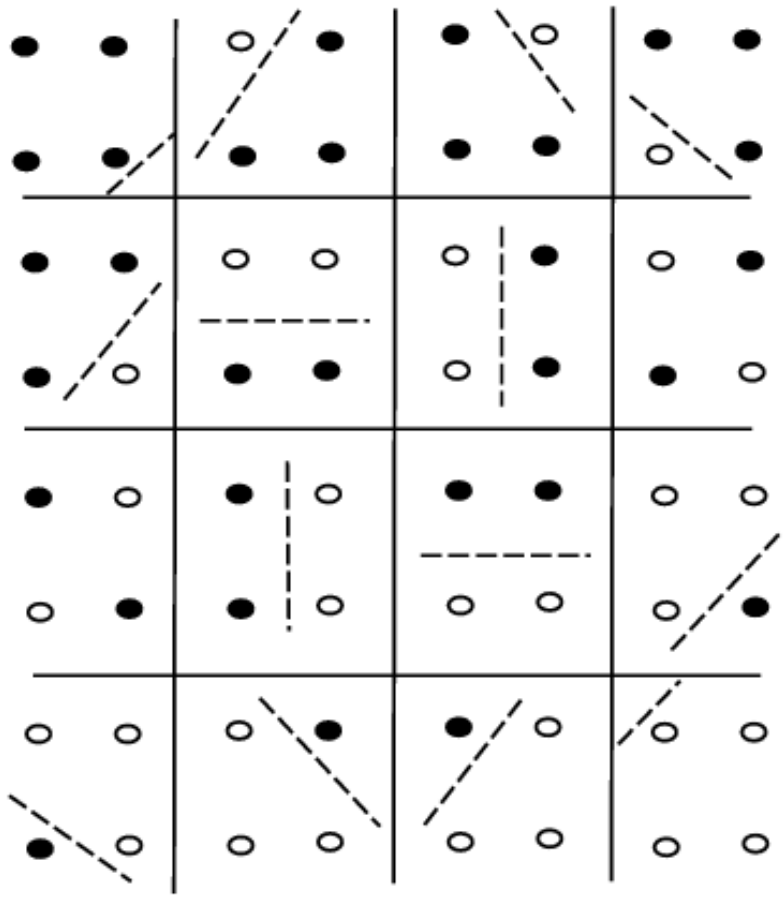
- $\mathcal{N}(\mathcal{H}, 2) = 4 = 2^2$

Example



- $\mathcal{N}(\mathcal{H}, 3) = 8 = 2^3$

Example



- $\mathcal{N}(\mathcal{H}, 4) = 14 < 2^4$

Capacity concepts

- Recall: we search for other capacity measures of $\mathcal{H}$ replacing $\ln N$.
- We know $\underbrace{\mathcal{N}(\mathcal{H}, Z_n)}_{\text{depends on sample}} \leq \mathcal{N}(\mathcal{H}, n) \leq \underbrace{2^n}_{\text{too loose}}$
- Dependency on sample can be removed by averaging over all samples: $E[\mathcal{N}(\mathcal{H}, Z_n)]$. It turns out that this is a valid capacity measure:

Theorem (Vapnik and Chervonenkis)

Let $Z_{2n} = ((x_1, y_1), \ldots, (x_{2n}, y_{2n})$ be a sample of size $2n$. For any $\epsilon > 0$ it holds that

$$P(\sup_{f \in \mathcal{H}} R[f] - R_{emp}[f] > \epsilon) \leq 4 \exp(\ln E[\mathcal{N}(\mathcal{H}, Z_{2n})] - \frac{n\epsilon^2}{8})$$

- If $\ln E[\mathcal{N}(\mathcal{H}, Z_{2n})]$ grows **sublinearly**, we get a **nontrivial bound.**

Some consequences

$$P(\sup_{f \in \mathcal{H}} R[f] - R_{\text{emp}}[f] > \epsilon) \leq 4 \exp(\ln E[\mathcal{N}(\mathcal{H}, Z_{2n})] - \frac{n\epsilon^2}{8})$$

- Bound holds uniformly for all functions in $\mathcal{H}$
  $\rightsquigarrow$ can use it for the functions that minimize R_{emp}.
  $\rightsquigarrow$ We can **bound the test error:**

$$P(R[f_n] - R_{\text{emp}}[f_n] > \epsilon) \leq 4E[\mathcal{N}(\mathcal{H}, Z_{2n})] \exp(-\frac{n\epsilon^2}{8}).$$

- Can derive a **confidence interval:** equate r.h.s. to δ and solve for ϵ:
  With probability at least $(1 - \delta)$ it holds that

$$R[f_n] \leq R_{\text{emp}}[f_n] + \epsilon(\delta)$$

$$R[f_n] \leq R_{\text{emp}}[f_n] + \sqrt{\frac{8}{n}\left(\ln E[\mathcal{N}(\mathcal{H}, Z_{2n})] + \ln \frac{4}{\delta}\right)}$$

- Bound depends on $\mathcal{H}$, n and the **unknown probability** $P(Z)$.

VC Dimension and other capacity concepts

- **Growth function**: upper bound expectation by maximum:

$$\mathcal{G}_{\mathcal{H}}(n) = \ln[\max_{Z_n} \mathcal{N}(\mathcal{H}, Z_n)] = \ln \underbrace{\mathcal{N}(\mathcal{H}, n)}_{\text{Shattering coeff.}} .$$

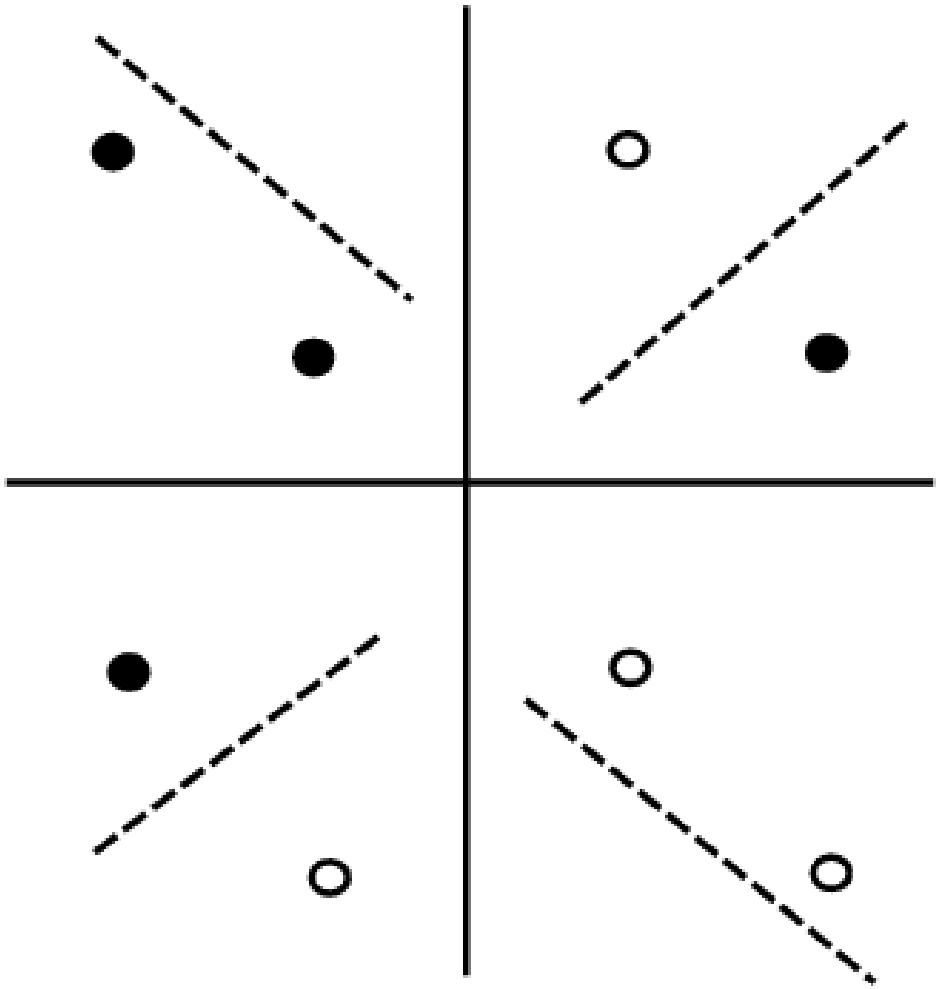
- **VC-Dimension**: recall that $\mathcal{N}(\mathcal{H}, n) \leq 2^n$. Vapnik & Chervonenkis showed that either $\mathcal{N}(\mathcal{H}, n) = 2^n$ for **all** n, or there exists some **maximal** n for which this is the case.

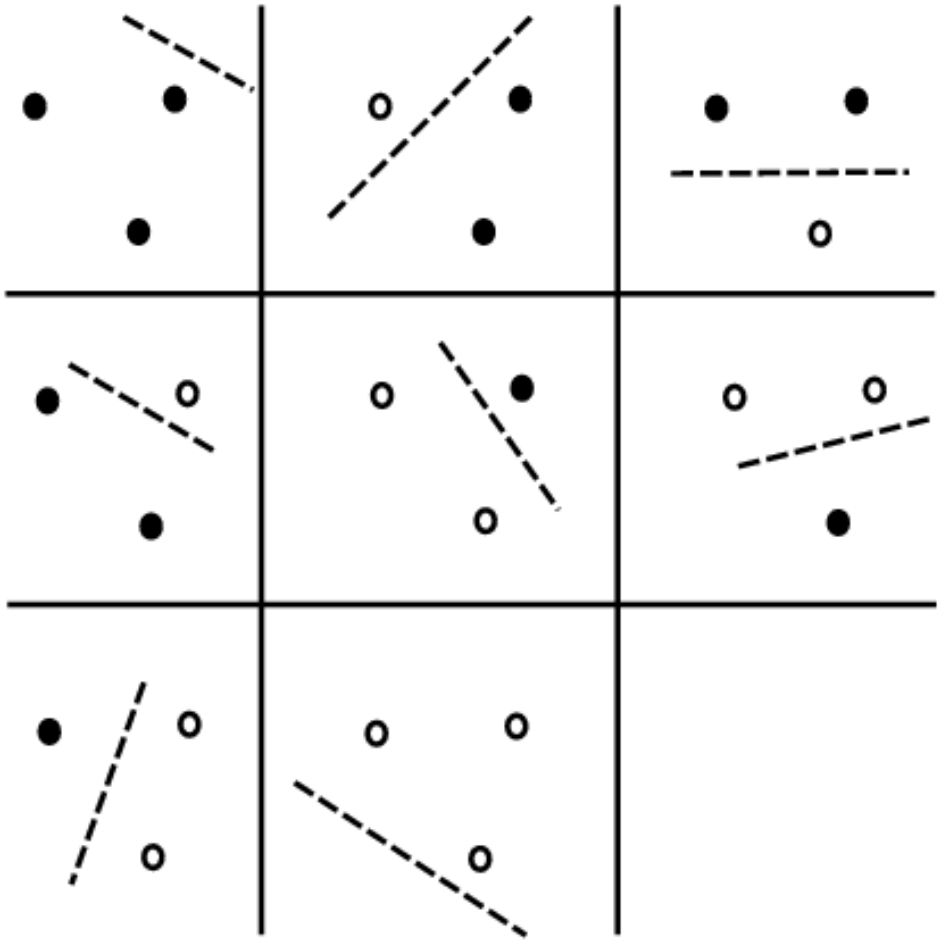
Definition

The *VC dimension* h of a class $\mathcal{H}$ is the largest n such that $\mathcal{N}(\mathcal{H}, n) = 2^n$, or, equivalently $\mathcal{G}_{\mathcal{H}}(n) = n \ln(2)$.

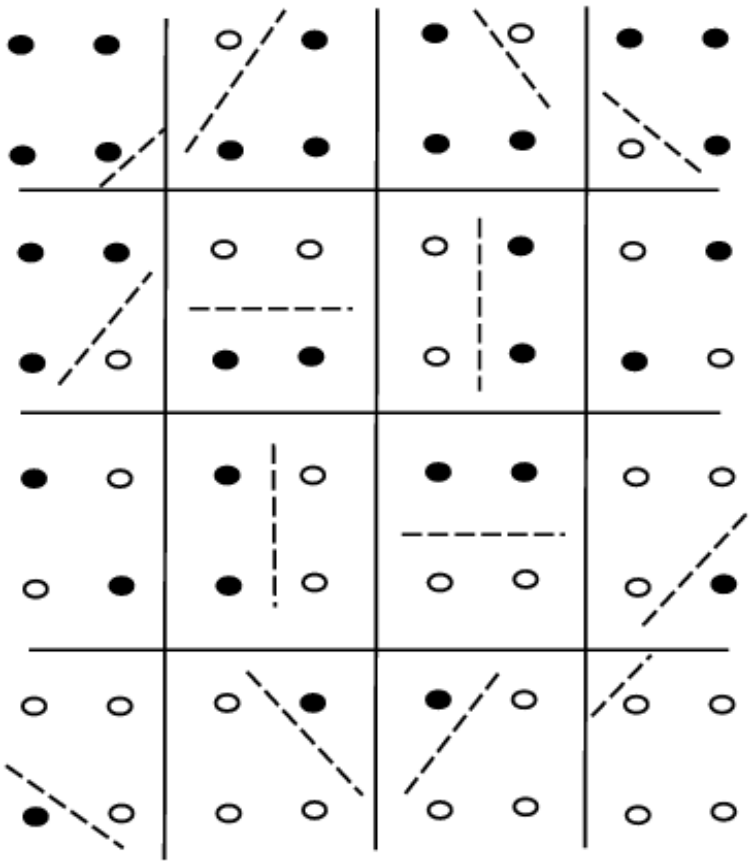
Interpretation: The VC-Dimension is the maximal number of samples that can be classified in all 2^n possible ways.

VC Dimension





VC Dimension



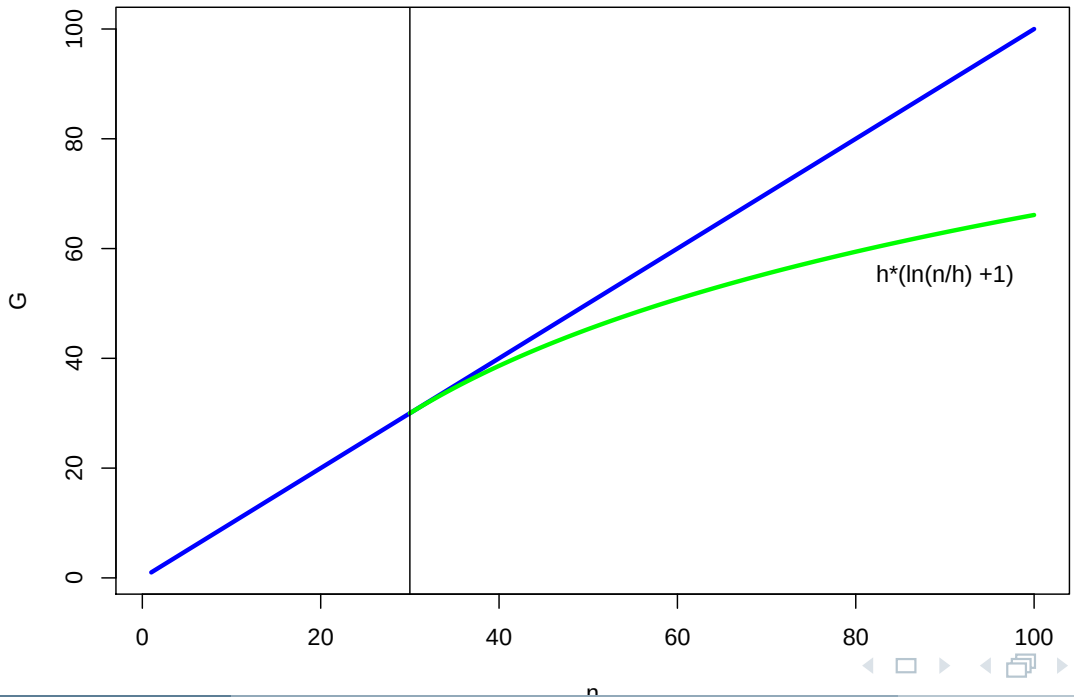
4 Points in $2D$ cannot be labeled in all possible ways by linear functions. The VC-Dimension is 3!

A remarkable property of the growth function

Theorem (Vapnik & Chervonenkis)

Let $\mathcal{H}$ be a class of functions with finite VC-dimension h. Then for $n \leq h$, $\mathcal{G}_{\mathcal{H}}(n)$ grows linearly with the sample size, and for all $n > h$

$$\mathcal{G}_{\mathcal{H}}(n) \leq h \left(\ln \frac{n}{h} + 1 \right).$$



Capacity concepts

Relation of capacity concepts:

$$\underbrace{\ln E[\mathcal{N}(\mathcal{H}, Z_{2n})]}_{\text{distribution dependent}} \leq \underbrace{\mathcal{G}_{\mathcal{H}}(n) \leq \overbrace{h\left(\ln \frac{n}{h} + 1\right)}^{\text{(sometimes) easy to compute}}}_{\text{distribution independent}}$$

Structure of bounds:

$$R[f_n] \leq R_{\text{emp}}[f_n] + \sqrt{\frac{a}{n}\left(\text{capacity}(\mathcal{H}) + \ln \frac{b}{\delta}\right)}$$

If the VC Dimension is finite, we get non-trivial bounds!

VC-Dimension for linear functions

Theorem

The VC dimension of linear functions in d-dimensional space is $d + 1$.

Question: Does the number of parameters coincide with the VC-Dimension? No!! Counter example:

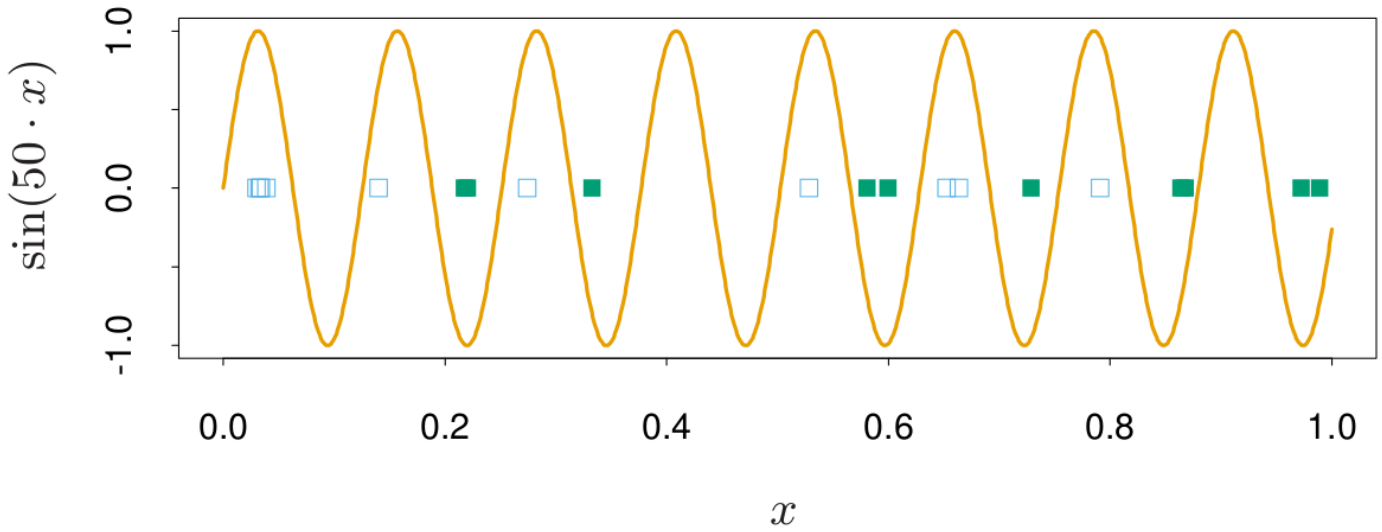


FIGURE 7.5 in (Hastie et al.: The Elements of Statistical Learning). Solid curve: sin(50x) for $x \in [0, 1]$. Blue and green points illustrate how sign(sin(αx)) can separate an arbitrarily large number of points by choosing a high frequency α.

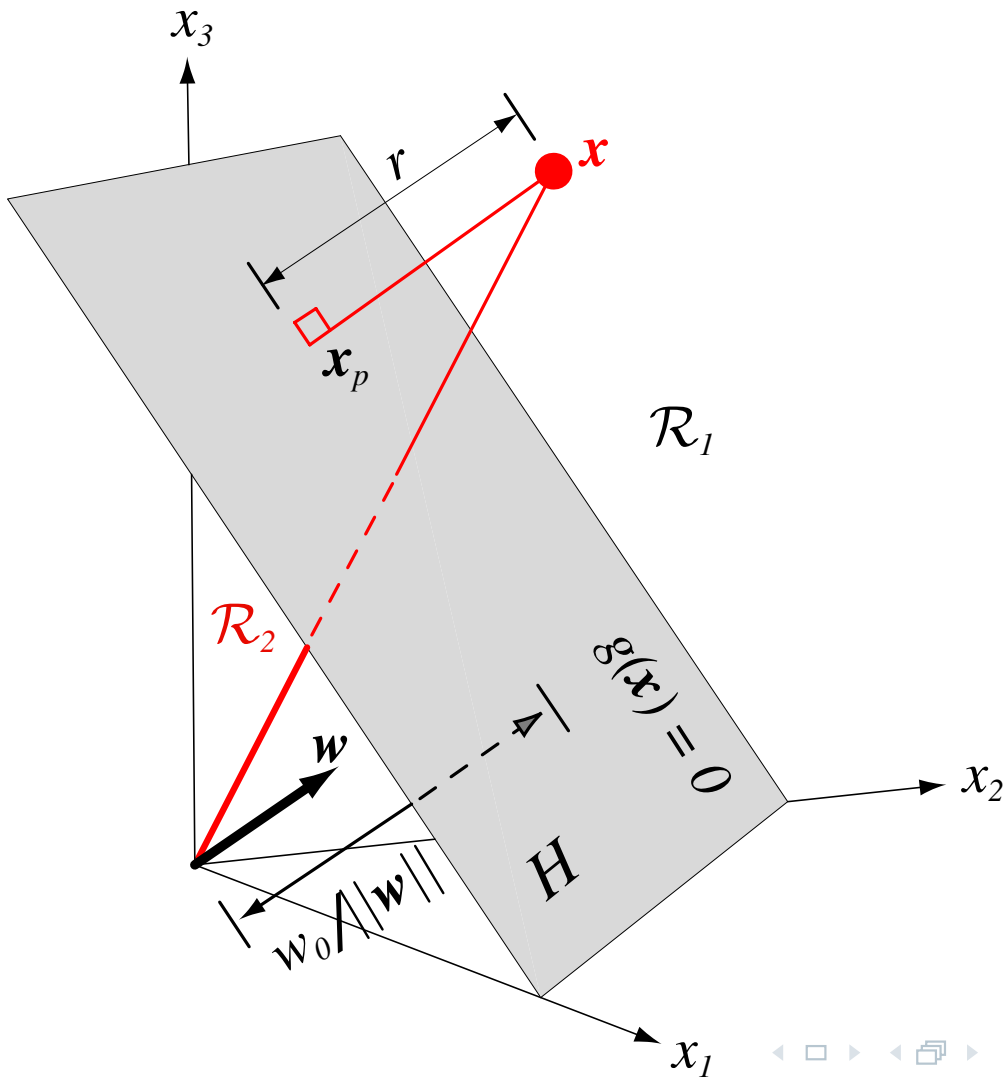
The VC-Dimension of $\{\text{sign}(sin(\alpha x))|\alpha \in \mathbb{R}\}$ is infinite.

Linear functions: Role of the margin

- Recall: the VC dimension of linear functions on $\mathbb{R}^d$ is $d + 1$.
- We need finite VC dimension for "simple" nontrivial bounds.
- Question: is learning impossible in **infinite** dimensional spaces (e.g. Gaussian RBF kernels)?
- **Not necessarily!** The capacity of the subset of hyperplanes with **large classification margin** can be much smaller than the general VC dimension of all hyperplanes.

Recall: Decision hyperplanes

- $f(\boldsymbol{x}; \boldsymbol{w})$ defines distance r from $\boldsymbol{x}$ to the hyperplane: $\boldsymbol{x} = \boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$.
- $f(\boldsymbol{x}_p) = 0 \Rightarrow f(\boldsymbol{x}) = r\|\boldsymbol{w}\| \quad \Leftrightarrow \quad r = f(\boldsymbol{x})/\|\boldsymbol{w}\|$.



Canonical hyperplanes

- Definition of hyperplane is not unique: weight vector $\boldsymbol{w}$ can be multiplied by any nonzero constant.
- The definition of a **canonical** hyperplane overcomes this ambiguity by additionally requiring

$$\min_{i=1,\ldots,n} \left| \boldsymbol{w}^t \boldsymbol{x}_i + w_0 \right| = 1.$$

- Distance between canonical hyperplane and the closest point: **margin** $r = 1/\|\boldsymbol{w}\|$.

Structure on canonical hyperplanes

Theorem (Vapnik, 1982)

Let R be the radius of the smallest ball containing the points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$: $B_R(\boldsymbol{a}) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{a}\| < R, \boldsymbol{a} \in \mathbb{R}^d\}$. The set of canonical hyperplane decision functions $f(\boldsymbol{w}, w_0) = sign\{\boldsymbol{w}^t \boldsymbol{x} + w_0\}$ satisfying $\|\boldsymbol{w}\| \leq A$ has VC dimension h bounded by

$$h \leq R^2 A^2 + 1.$$

Intuitive interpretation: $\text{margin} = 1/\|\boldsymbol{w}\|$

$\rightsquigarrow$ **minimizing capacity($\mathcal{H}$) corresponds to maximizing the margin.**

Structure of bounds:

$$R[f_n] \leq R_{\text{emp}}[f_n] + \sqrt{\frac{a}{n}\left(\text{capacity}(\mathcal{H}) + \ln\frac{b}{\delta}\right)}$$

$\rightsquigarrow$ **Large margin classifiers.**