Risks

Matthieu Geist (CentraleSupélec)
matthieu.geist@centralesupelec.fr

The paradigm

To formalize the learning problem, we assume that we have:

- a random generator of vectors $x \in \mathcal{X}$, sampled i.i.d. from $\mathrm{P}(x)$, *fixed* but *unknown*;
- an oracle that for each input x provides an output $y \in \mathcal{Y}$, sampled according to $\mathrm{P}(y \mid x)$, also *fixed* but *unknown*.
- a machine that can implement a set of functions, this set being called the hypothesis space $\mathcal{H} = \{f : \mathcal{X} \to \mathcal{Y}\} \subset \mathcal{Y}^{\mathcal{X}}$.

The problem, roughly

- pick $f \in \mathcal{H}$ that predict "the best" the responses of the oracle;
- f must be chosen based on a dataset $\mathcal{D} = \{(x_i, y_i)_{1 \leq i \leq n}\}$ of n examples sampled i.i.d. from the joint distribution $\mathrm{P}(x, y) = \mathrm{P}(x)\mathrm{P}(y \mid x)$;

The paradigm

To formalize the learning problem, we assume that we have:

- a random generator of vectors $x \in \mathcal{X}$, sampled i.i.d. from $\mathrm{P}(x)$, *fixed* but *unknown*;
- an oracle that for each input x provides an output $y \in \mathcal{Y}$, sampled according to $\mathrm{P}(y \mid x)$, also *fixed* but *unknown*.
- a machine that can implement a set of functions, this set being called the hypothesis space $\mathcal{H} = \{f : \mathcal{X} \to \mathcal{Y}\} \subset \mathcal{Y}^{\mathcal{X}}$.

The problem, roughly

- pick $f \in \mathcal{H}$ that predict "the best" the responses of the oracle;
- f must be chosen based on a dataset $\mathcal{D} = \{(x_i, y_i)_{1 \leq i \leq n}\}$ of n examples sampled i.i.d. from the joint distribution $\mathrm{P}(x, y) = \mathrm{P}(x)\mathrm{P}(y \mid x)$;

Examples of hypothesis spaces

- linear predictions
  $\mathcal{H} = \left\{ f_{\alpha,\beta} : \mathcal{X} \to \mathbb{R}, \quad f_{\alpha,\beta}(x) = \alpha^\top x + \beta, \alpha \in \mathbb{R}^p, \beta \in \mathbb{R} \right\}$
  $\mathcal{H} =$
  $\left\{ f_{\alpha,\beta} : \mathcal{X} \to \{-1, 1\}, \quad f_{\alpha,\beta}(x) = \mathrm{sgn}(\alpha^\top x + \beta), \alpha \in \mathbb{R}^p, \beta \in \mathbb{R} \right\}$
- RBFNs
  $\mathcal{H} = \left\{ f_{\alpha,\beta} : x \to \sum_{i=1}^d \alpha_i \exp\left(\frac{1}{2}(x - \mu_i)^\top \Sigma_i^{-1}(x - \mu_i)\right) + \beta \right\}$
- linear parameterizations
  $\mathcal{H} = \left\{ f_\alpha : x \to \alpha^\top \phi(x), \alpha \in \mathbb{R}^d \right\}$
- nonlinear parameterization (eg. neural networks)
- RKHS
  $\mathcal{H} = \{ f_\alpha : x \to \sum_{i=1}^n \alpha_i K(x, x_i), \alpha \in \mathbb{R}^n \}$

Measuring locally the quality of the prediction

Loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, $L(y, f(x))$ measures the error between the response y of the oracle for a given input x and the prediction $f(x)$ of the machine for the same input

- ℓ_2-loss

$$L(y, f(x)) = (y - f(x))^2$$

- ℓ_1-loss

$$L(y, f(x)) = |y - f(x)|$$

- binary loss

$$L(y, f(x)) = \mathbb{I}_{\{y \neq f(x)\}} = \begin{cases} 1 & \text{if } f(x) \neq y \\ 0 & \text{else} \end{cases}$$

Measuring locally the quality of the prediction

Loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, $L(y, f(x))$ measures the error between the response y of the oracle for a given input x and the prediction $f(x)$ of the machine for the same input

- ℓ_2-loss
$$L(y, f(x)) = (y - f(x))^2$$
- ℓ_1-loss
$$L(y, f(x)) = |y - f(x)|$$
- binary loss
$$L(y, f(x)) = \mathbb{I}_{\{y \neq f(x)\}} = \begin{cases} 1 & \text{if } f(x) \neq y \\ 0 & \text{else} \end{cases}$$

Measuring globally the quality of the prediction

Risk, defined as the expected loss

$$\mathcal{R}(f) = \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(x)) \mathrm{dP}(x, y) = \mathrm{E}\left[L(Y, f(X))\right]$$

- ideally, search for the minimizer $f_0 = \mathrm{argmin}_{f\in\mathcal{H}} \mathcal{R}(f)$
- yet $\mathrm{P}(x, y)$ is unknown...

Empirical risk, defined as the empirical mean of the loss

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i))$$

- practically, search for the minimizer $f_n = \mathrm{argmin}_{f\in\mathcal{H}} \mathcal{R}_n(f)$
- notice that f_n **is a random quantity**
- this is called the ERM (empirical risk minimization)

Measuring globally the quality of the prediction

Risk, defined as the expected loss

$$\mathcal{R}(f) = \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(x)) \mathrm{dP}(x, y) = \mathrm{E}\left[L(Y, f(X))\right]$$

- ideally, search for the minimizer $f_0 = \mathrm{argmin}_{f\in\mathcal{H}} \mathcal{R}(f)$
- yet $\mathrm{P}(x, y)$ is unknown...

Empirical risk, defined as the empirical mean of the loss

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i))$$

- practically, search for the minimizer $f_n = \mathrm{argmin}_{f\in\mathcal{H}} \mathcal{R}_n(f)$
- notice that f_n **is a random quantity**
- this is called the ERM (empirical risk minimization)

Summary

Given

- $\mathcal{D} = \{(x_i, y_i)_{1 \leq i \leq n}\}$, sampled iid from unkown $\mathrm{P}(x, y)$,
- $\mathcal{H}$, chosen by the practitioner,
- L, chosen by the practitioner and depending on the problem at hand,

supervised learning computes f_n (instead of ideally f_0)

Questions

- does f_n converges to f_0?
- given n samples and $\mathcal{H}$, how close is f_n to f_0?

Summary

Given

- $\mathcal{D} = \{(x_i, y_i)_{1 \leq i \leq n}\}$, sampled iid from unkown $\mathrm{P}(x, y)$,
- $\mathcal{H}$, chosen by the practitioner,
- L, chosen by the practitioner and depending on the problem at hand,

supervised learning computes f_n (instead of ideally f_0)

Questions

- does f_n converges to f_0?
- given n samples and $\mathcal{H}$, how close is f_n to f_0?

Write f_* the unconstrained minimizer of the risk:

$$f_* = \operatorname*{argmin}_{f \in \mathcal{Y}^{\mathcal{X}}} \mathcal{R}(f), \quad \mathcal{R}_* = \mathcal{R}(f_*).$$

Bias-variance decomposition

$$\underbrace{\mathcal{R}(f_n) - \mathcal{R}_*}_{\text{error}} = \underbrace{\mathcal{R}(f_0) - \mathcal{R}_*}_{\text{bias}} + \underbrace{\mathcal{R}(f_n) - \mathcal{R}(f_0)}_{\text{variance}}.$$

- bias: how well chosen is $\mathcal{H}$, does not depend on data
- variance: random quantity that measures how close is f_n to f_0

Write f_* the unconstrained minimizer of the risk:

$$f_* = \operatorname*{argmin}_{f \in \mathcal{Y}^{\mathcal{X}}} \mathcal{R}(f), \quad \mathcal{R}_* = \mathcal{R}(f_*).$$

Bias-variance decomposition

$$\underbrace{\mathcal{R}(f_n) - \mathcal{R}_*}_{\text{error}} = \underbrace{\mathcal{R}(f_0) - \mathcal{R}_*}_{\text{bias}} + \underbrace{\mathcal{R}(f_n) - \mathcal{R}(f_0)}_{\text{variance}}.$$

- bias: how well chosen is $\mathcal{H}$, does not depend on data
- variance: random quantity that measures how close is f_n to f_0

Abstraction

Write $z = (x, y)$ and $Q(z, f) = L(y, f(x))$, we have

$$\mathcal{R}(f) = \mathrm{E}\left[Q(Z, f)\right] \text{ and } \mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} Q(z_i, f).$$

In the probability theory, convergence of the empirical mean of iid random variables to the common expectation: law of large numbers

Let $f \in \mathcal{H}$ be a **fixed** function, by the weak law of large numbers,

$$\mathcal{R}_n(f) \xrightarrow[n\to\infty]{P} \mathcal{R}(f) \Leftrightarrow \forall \epsilon > 0, \mathrm{P}\left(|\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) \xrightarrow[n\to\infty]{} 0.$$

One can't replace f by f_n in the above Eq., as f_n depends on data

Abstraction

Write $z = (x, y)$ and $Q(z, f) = L(y, f(x))$, we have

$$\mathcal{R}(f) = \mathrm{E}\left[Q(Z, f)\right] \text{ and } \mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} Q(z_i, f).$$

In the probability theory, convergence of the empirical mean of iid random variables to the common expectation: law of large numbers

Let $f \in \mathcal{H}$ be a **fixed** function, by the weak law of large numbers,

$$\mathcal{R}_n(f) \xrightarrow[n\to\infty]{P} \mathcal{R}(f) \Leftrightarrow \forall \epsilon > 0, \mathrm{P}\left(|\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) \xrightarrow[n\to\infty]{} 0.$$

One can't replace f by f_n in the above Eq., as f_n depends on data

Abstraction

Write $z = (x, y)$ and $Q(z, f) = L(y, f(x))$, we have

$$\mathcal{R}(f) = \mathrm{E}\left[Q(Z, f)\right] \text{ and } \mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} Q(z_i, f).$$

In the probability theory, convergence of the empirical mean of iid random variables to the common expectation: law of large numbers

Let $f \in \mathcal{H}$ be a **fixed** function, by the weak law of large numbers,

$$\mathcal{R}_n(f) \xrightarrow[n\to\infty]{P} \mathcal{R}(f) \Leftrightarrow \forall \epsilon > 0, \mathrm{P}\left(|\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) \xrightarrow[n\to\infty]{} 0.$$

One can't replace f by f_n in the above Eq., as f_n depends on data

Abstraction

Write $z = (x, y)$ and $Q(z, f) = L(y, f(x))$, we have

$$\mathcal{R}(f) = \mathrm{E}\left[Q(Z, f)\right] \text{ and } \mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} Q(z_i, f).$$

In the probability theory, convergence of the empirical mean of iid random variables to the common expectation: law of large numbers

Let $f \in \mathcal{H}$ be a **fixed** function, by the weak law of large numbers,

$$\mathcal{R}_n(f) \xrightarrow[n\to\infty]{P} \mathcal{R}(f) \Leftrightarrow \forall \epsilon > 0, \mathrm{P}\left(|\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) \xrightarrow[n\to\infty]{} 0.$$

One can't replace f by f_n in the above Eq., as f_n depends on data

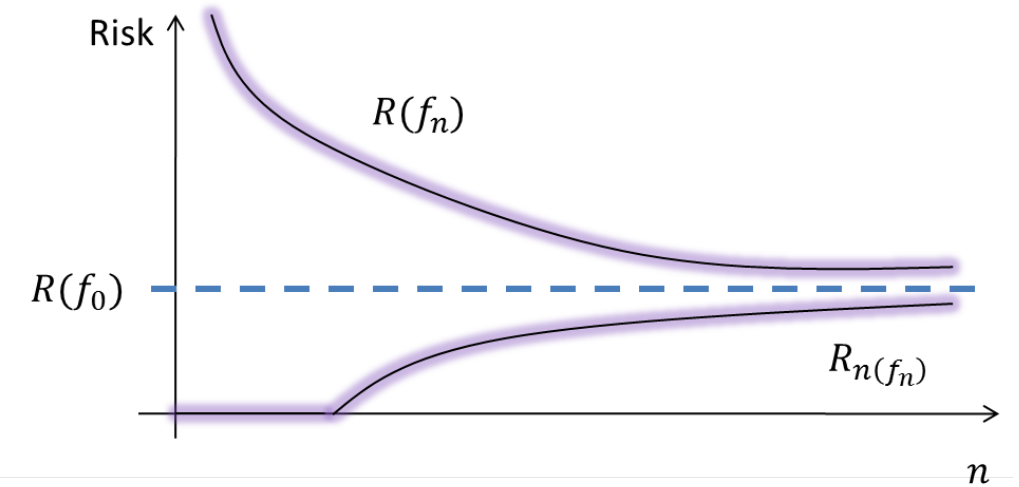
Classic consistency

Definition (Classic consistency of the ERM principle)

We say the the ERM principle is consistent for the set of functions $Q(z, f)$, $f \in \mathcal{H}$, and for the distribution $\mathrm{P}(z)$ if the following convergences occur:

$$\mathcal{R}(f_n) \xrightarrow[n\to\infty]{P} \mathcal{R}(f_0) \text{ and } \mathcal{R}_n(f_n) \xrightarrow[n\to\infty]{P} \mathcal{R}(f_0).$$

Classic consistency

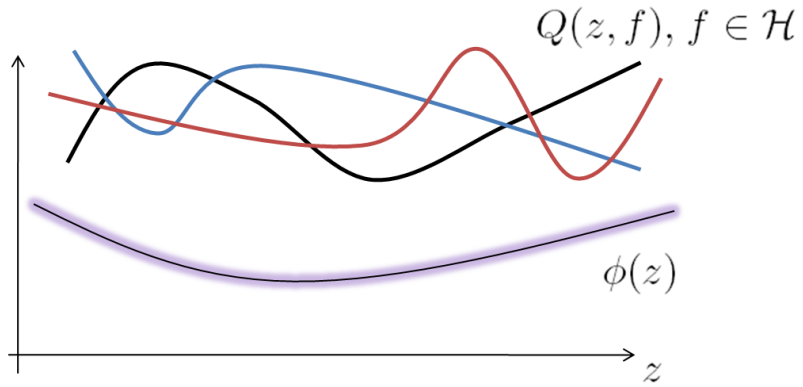


Definition (Classic consistency of the ERM principle)

We say the the ERM principle is consistent for the set of functions $Q(z, f)$, $f \in \mathcal{H}$, and for the distribution $\mathrm{P}(z)$ if the following convergences occur:

$$\mathcal{R}(f_n) \xrightarrow[n\to\infty]{P} \mathcal{R}(f_0) \text{ and } \mathcal{R}_n(f_n) \xrightarrow[n\to\infty]{P} \mathcal{R}(f_0).$$

Strict consistency



Definition (Strict consistency of the ERM principle)

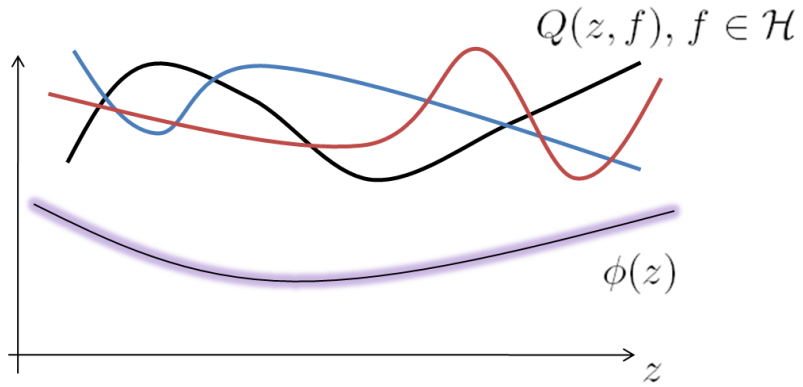
Let $Q(z, f),\ f \in \mathcal{H}$ be a set of function and $P(z)$ a distribution. For $c \in \mathbb{R}$, define $\mathcal{H}(c)$ the set

$$\mathcal{H}(c) = \{f \in \mathcal{H} :\ \mathcal{R}(f) = \int Q(z, f) \mathrm{dP}(z) \geq c\}.$$

The principle of ERM is said to be strictly consistent (for the above set of functions and distribution) if for any $c \in \mathbb{R}$, we have

$$\inf_{f \in \mathcal{H}(c)} \mathcal{R}_n(f) \xrightarrow[n \to \infty]{P} \inf_{f \in \mathcal{H}(c)} \mathcal{R}(f).$$

Strict consistency



Definition (Strict consistency of the ERM principle)

Let $Q(z, f),\ f \in \mathcal{H}$ be a set of function and $P(z)$ a distribution.
For $c \in \mathbb{R}$, define $\mathcal{H}(c)$ the set
$\mathcal{H}(c) = \{f \in \mathcal{H} :\ \mathcal{R}(f) = \int Q(z, f)\mathrm{dP}(z) \geq c\}$.
The principle of ERM is said to be strictly consistent (for the above set of functions and distribution) if for any $c \in \mathbb{R}$, we have

$$\inf_{f \in \mathcal{H}(c)} \mathcal{R}_n(f) \xrightarrow[n \to \infty]{P} \inf_{f \in \mathcal{H}(c)} \mathcal{R}(f).$$

Uniform convergence

Recall the weak law of large number: for $f \in \mathcal{H}$,

$$\mathcal{R}_n(f) \xrightarrow[n\to\infty]{P} \mathcal{R}(f).$$

It is not sufficient to ensure consistency (can't put f_n instead of f).
Consider a uniform convergence in probabilities:

$$\sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| \xrightarrow[n\to\infty]{P} 0$$

$$\Leftrightarrow \forall \epsilon > 0, \mathrm{P}\left(\sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) \xrightarrow[n\to\infty]{} 0.$$

It is much stronger than the law of large numbers (worst case analysis), and sufficient for consistency, as

$$|\mathcal{R}(f_n) - \mathcal{R}_n(f_n)| \leq \sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)|.$$

Uniform convergence

Recall the weak law of large number: for $f \in \mathcal{H}$,

$$\mathcal{R}_n(f) \xrightarrow[n\to\infty]{P} \mathcal{R}(f).$$

It is not sufficient to ensure consistency (can't put f_n instead of f).
Consider a uniform convergence in probabilities:

$$\sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| \xrightarrow[n\to\infty]{P} 0$$

$$\Leftrightarrow \forall \epsilon > 0, \mathrm{P}\left(\sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) \xrightarrow[n\to\infty]{} 0.$$

It is much stronger than the law of large numbers (worst case analysis), and sufficient for consistency, as

$$|\mathcal{R}(f_n) - \mathcal{R}_n(f_n)| \leq \sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)|.$$

Uniform convergence

Recall the weak law of large number: for $f \in \mathcal{H}$,

$$\mathcal{R}_n(f) \xrightarrow[n\to\infty]{P} \mathcal{R}(f).$$

It is not sufficient to ensure consistency (can't put f_n instead of f). Consider a uniform convergence in probabilities:

$$\sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| \xrightarrow[n\to\infty]{P} 0$$

$$\Leftrightarrow \forall \epsilon > 0, \mathrm{P}\left(\sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) \xrightarrow[n\to\infty]{} 0.$$

It is much stronger than the law of large numbers (worst case analysis), and sufficient for consistency, as

$$|\mathcal{R}(f_n) - \mathcal{R}_n(f_n)| \leq \sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)|.$$

Theorem (Vapnik's key theorem)

Assume that there exists two constants a and A such that for any function $Q(z, f)$, $f \in \mathcal{H}$, and for a given distribution $\mathrm{P}(z)$, we have

$$a \leq \mathcal{R}(f) = \int Q(z, f) \mathrm{dP}(z) \leq A.$$

Then, the following assertion are equivalents:

1. *for the distribution $\mathrm{P}(z)$, the ERM principle is strictly consistent for the set of functions $Q(z, f)$, $f \in \mathcal{H}$;*
2. *for the distribution $\mathrm{P}(z)$ there is a one-sided uniform convergence over the set of functions $Q(z, f)$, $f \in \mathcal{H}$,*

$$\forall \epsilon > 0, \mathrm{P}\left(\sup_{f \in \mathcal{H}} (\mathcal{R}(f) - \mathcal{R}_n(f))_+ > \epsilon\right) \xrightarrow[n \to \infty]{} 0,$$

where $(x)_+ = \max(x, 0)$.

Any analysis is a worst case analysis

Theorem (Vapnik's key theorem)

Assume that there exists two constants a and A such that for any function $Q(z, f)$, $f \in \mathcal{H}$, and for a given distribution $\mathrm{P}(z)$, we have

$$a \leq \mathcal{R}(f) = \int Q(z, f) \mathrm{dP}(z) \leq A.$$

Then, the following assertion are equivalents:

1. *for the distribution $\mathrm{P}(z)$, the ERM principle is strictly consistent for the set of functions $Q(z, f)$, $f \in \mathcal{H}$;*
2. *for the distribution $\mathrm{P}(z)$ there is a one-sided uniform convergence over the set of functions $Q(z, f)$, $f \in \mathcal{H}$,*

$$\forall \epsilon > 0, \mathrm{P}\left(\sup_{f \in \mathcal{H}} (\mathcal{R}(f) - \mathcal{R}_n(f))_+ > \epsilon\right) \xrightarrow[n \to \infty]{} 0,$$

where $(x)_+ = \max(x, 0)$.

Any analysis is a worst case analysis

Quantitative law of large numbers

For now, assume that for any z and any f, $Q(z, f) \in \{0, 1\}$ (binary classification).

Theorem (Hoeffding's inequality)

Let $X_1, \ldots, X_n$ be i.i.d. random variables, bounded in (0,1) and of common mean $\mu = \mathrm{E}[X_1]$. Then:

$$\forall \epsilon > 0, \quad \mathrm{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \epsilon \right) \leq 2e^{-2n\epsilon^2}.$$

PAC (Probably Approximately Correct) form:

- with probability at least $1 - \delta$, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}$$

Quantitative law of large numbers

For now, assume that for any z and any f, $Q(z, f) \in \{0, 1\}$ (binary classification).

Theorem (Hoeffding's inequality)

Let $X_1, \ldots, X_n$ be i.i.d. random variables, bounded in (0,1) and of common mean $\mu = \mathrm{E}[X_1]$. Then:

$$\forall \epsilon > 0, \quad \mathrm{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \epsilon \right) \leq 2e^{-2n\epsilon^2}.$$

PAC (Probably Approximately Correct) form:

- with probability at least $1 - \delta$, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}$$

Quantitative law of large numbers

For now, assume that for any z and any f, $Q(z, f) \in \{0, 1\}$ (binary classification).

Theorem (Hoeffding's inequality)

Let $X_1, \ldots, X_n$ be i.i.d. random variables, bounded in (0,1) and of common mean $\mu = \mathrm{E}[X_1]$. Then:

$$\forall \epsilon > 0, \quad \mathrm{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \epsilon \right) \leq 2e^{-2n\epsilon^2}.$$

PAC (Probably Approximately Correct) form:

- with probability at least $1 - \delta$, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}$$

Back to risk control

Let $f \in \mathcal{H}$ be fixed, w.p. at least $1 - \delta$ we have

$$|\mathcal{R}(f) - \mathcal{R}_n(f)| \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}.$$

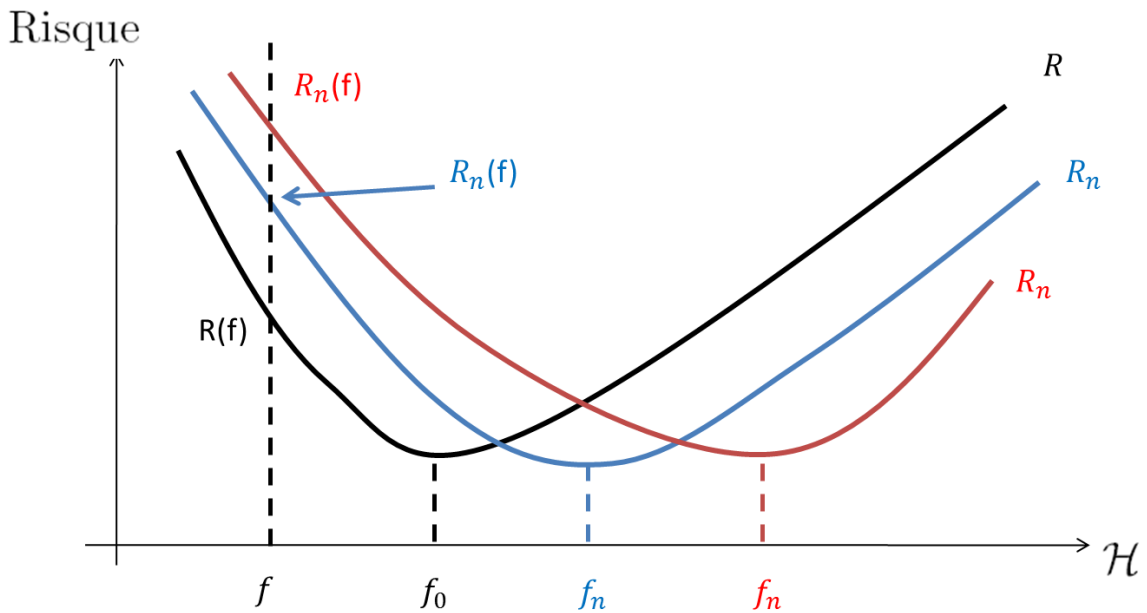
This is not enough for uniform convergence...

Back to risk control

Let $f \in \mathcal{H}$ be fixed, w.p. at least $1 - \delta$ we have

$$|\mathcal{R}(f) - \mathcal{R}_n(f)| \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}.$$

This is not enough for uniform convergence...



Finite hypothesis space

Assume that $\mathcal{H}$ is a finite set, such that Card $\mathcal{H} = h$. We can write

$$\begin{aligned} \mathrm{P}\left(\sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) &= \mathrm{P}\left(\bigcup_{f\in\mathcal{H}} \{|\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\}\right) \\ &\leq \sum_{f\in\mathcal{H}} \mathrm{P}\left(|\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) \\ &\leq 2he^{-2n\epsilon^2}. \end{aligned}$$

In other words, w.p. at least $1-\delta$, we have

$$|\mathcal{R}(f_n) - \mathcal{R}_n(f_n)| \leq \sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| \leq \sqrt{\frac{\ln \frac{2h}{\delta}}{2n}}.$$

finite $\mathcal{H}$ is a too strong assumption

Finite hypothesis space

Assume that $\mathcal{H}$ is a finite set, such that Card $\mathcal{H} = h$. We can write

$$\begin{aligned} \mathrm{P}\left(\sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) &= \mathrm{P}\left(\bigcup_{f\in\mathcal{H}} \{|\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\}\right) \\ &\leq \sum_{f\in\mathcal{H}} \mathrm{P}\left(|\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) \\ &\leq 2he^{-2n\epsilon^2}. \end{aligned}$$

In other words, w.p. at least $1-\delta$, we have

$$|\mathcal{R}(f_n) - \mathcal{R}_n(f_n)| \leq \sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| \leq \sqrt{\frac{\ln \frac{2h}{\delta}}{2n}}.$$

finite $\mathcal{H}$ is a too strong assumption

Finite hypothesis space

Assume that $\mathcal{H}$ is a finite set, such that Card $\mathcal{H} = h$. We can write

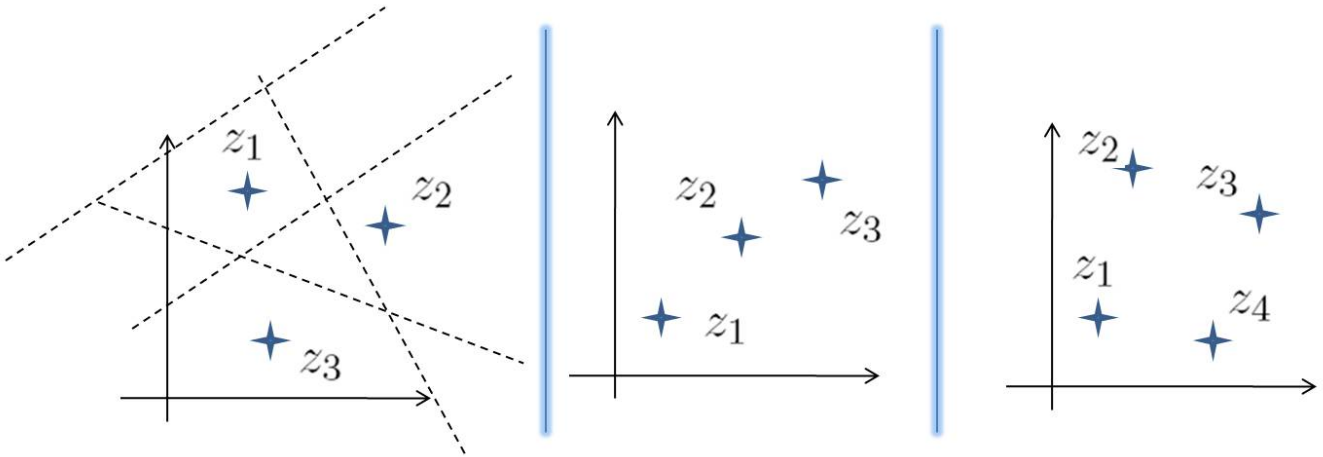
$$\begin{aligned} \mathrm{P}\left(\sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) &= \mathrm{P}\left(\bigcup_{f\in\mathcal{H}} \{|\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\}\right) \\ &\leq \sum_{f\in\mathcal{H}} \mathrm{P}\left(|\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) \\ &\leq 2he^{-2n\epsilon^2}. \end{aligned}$$

In other words, w.p. at least $1-\delta$, we have

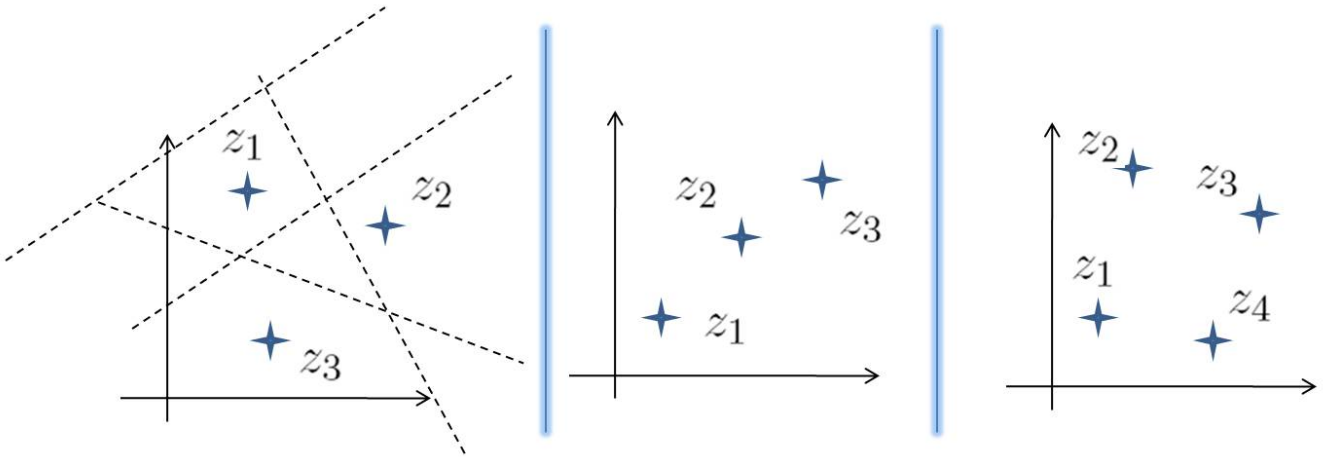
$$|\mathcal{R}(f_n) - \mathcal{R}_n(f_n)| \leq \sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| \leq \sqrt{\frac{\ln \frac{2h}{\delta}}{2n}}.$$

finite $\mathcal{H}$ is a too strong assumption

Counting smartly the number of functions



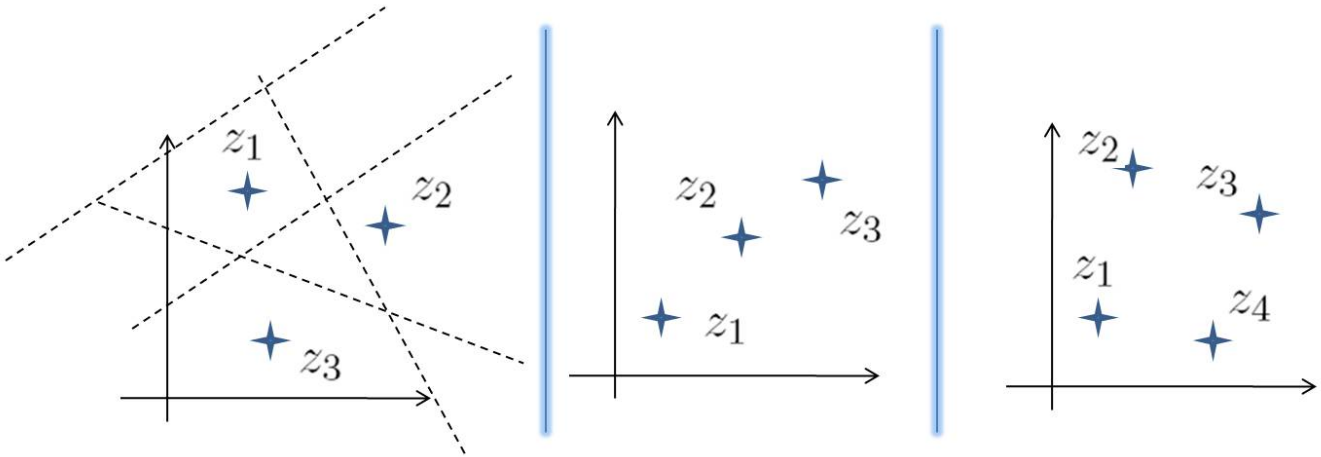
Counting smartly the number of functions



- consider $\sup_{f \in \mathcal{H}} \mathcal{R}_n(f)$
- $\mathcal{H}$ is uncountable, but $\mathcal{R}_n(f)$ takes less than 2^n values
- yet we're interested in

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

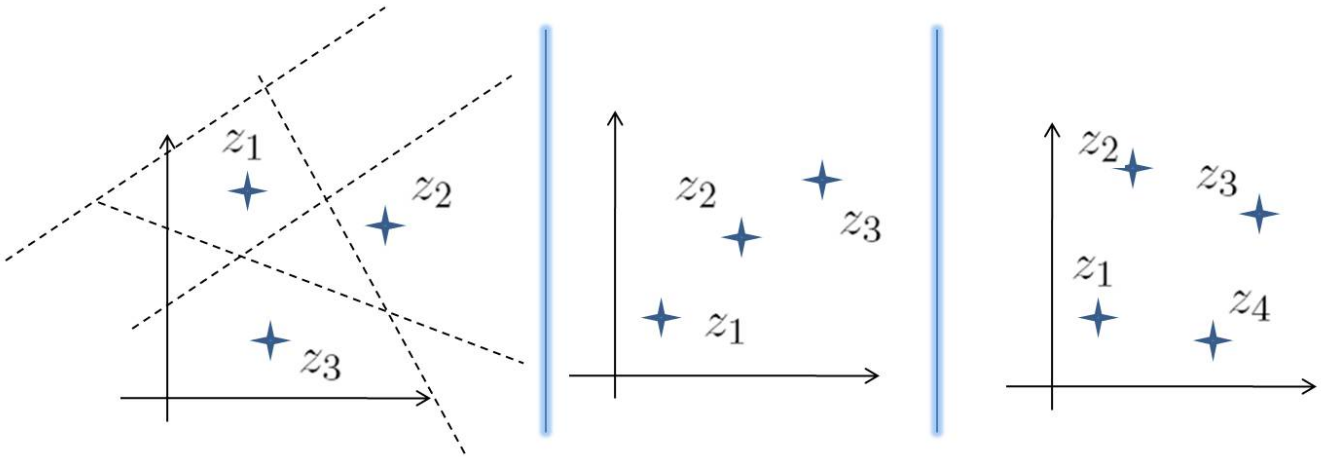
Counting smartly the number of functions



- consider $\sup_{f \in \mathcal{H}} \mathcal{R}_n(f)$
- $\mathcal{H}$ is uncountable, but $\mathcal{R}_n(f)$ takes less than 2^n values
- yet we're interested in

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

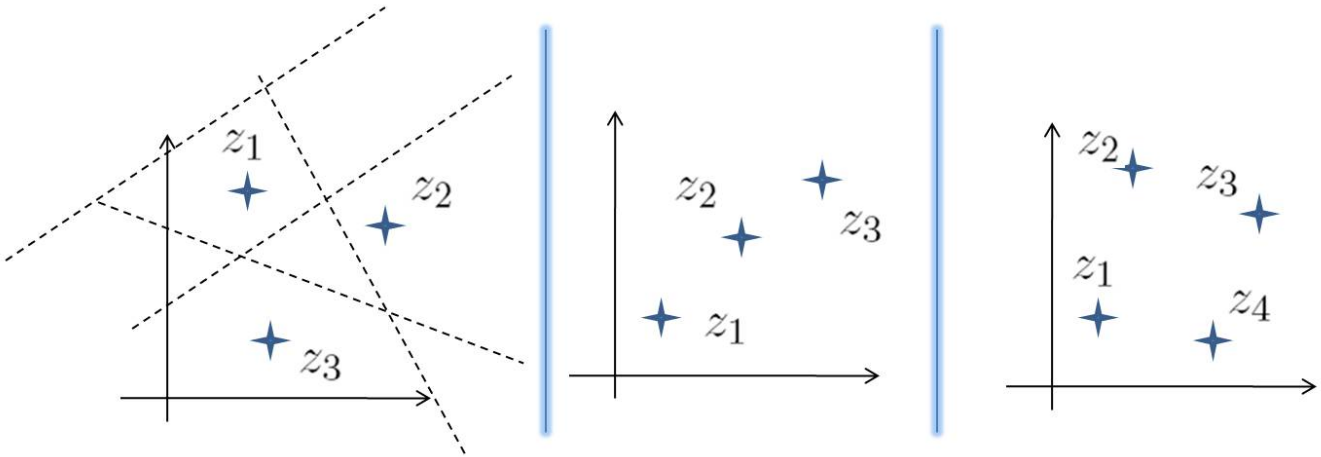
Counting smartly the number of functions



- consider $\sup_{f \in \mathcal{H}} \mathcal{R}_n(f)$
- $\mathcal{H}$ is uncountable, but $\mathcal{R}_n(f)$ takes less than 2^n values
- yet we're interested in

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

Counting smartly the number of functions



Theorem (Ghost sample)

Let $z'_1, \ldots, z'_n$ be a ghost sample, independent from the data $z_1, \ldots, z_n$. Write $\mathcal{R}'_n(f) = \frac{1}{n}\sum_{i=1}^{n} Q(z'_i, f)$ the associated empirical risk. Then, for any $\epsilon > 0$ such that $n\epsilon^2 \geq 2$, we have

$$\mathrm{P}\left(\sup_{f\in\mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| > \epsilon\right) \leq 2\mathrm{P}\left(\sup_{f\in\mathcal{H}} |\mathcal{R}'_n(f) - \mathcal{R}_n(f)| > \frac{\epsilon}{2}\right)$$

Theorem (Vapnik & Chervonenkis, Sauer & Shelah)

Define $G^{\mathcal{H}}(n)$ *the growth function of a set of functions* $Q(z, f)$, $f \in \mathcal{H}$, *as*

$$G^{\mathcal{H}}(n) = \ln\left(\sup_{z_1,\dots,z_n \in \mathcal{Z}} \left(N^{\mathcal{H}}(z_1, \dots, z_n)\right)\right)$$

$$\text{with } N^{\mathcal{H}}(z_1, \dots, z_n) = \text{Card}(Q_{z_1,\dots,z_n})$$

$$\text{and } Q_{z_1,\dots,z_n} = \left\{\begin{pmatrix} Q(z_1, f) & \dots & Q(z_n, f)\end{pmatrix}^\top : f \in \mathcal{H}\right\}.$$

The growth function satisfies one of the two properties:

1. *either* $G^{\mathcal{H}}(n)$ *is linear,* $G^{\mathcal{H}}(n) = n \ln 2, \forall n \in \mathbb{N}^*$;
2. *or* $G^{\mathcal{H}}(n)$ *is sub logarithmic after a given rank,*
   $G^{\mathcal{H}}(n) \begin{cases} = n \ln 2 & \text{if } n \leq h \\ \leq h(1 + \ln \frac{n}{h}) & \text{if } n > h \end{cases}$, *with* h *the greater integer such that* $G^{\mathcal{H}}(n) = n \ln 2$.

In the first case, the Vapnik-Chervonenkis dimension is infinite, $d_{VC}(\mathcal{H}) = \infty$. *In the second case, we have* $d_{VC}(\mathcal{H}) = h$.

A bound on the risk

Theorem (Bound on the risk)

Let $\delta \in (0,1)$ and $n > d_{VC}(\mathcal{H})$. W.p. at least $1-\delta$, we have

$$\forall f \in \mathcal{H}, \quad \mathcal{R}(f) \leq \mathcal{R}_n(f) + 2\sqrt{\frac{2}{n}\left(d_{VC}(\mathcal{H}) \ln \frac{2en}{d_{VC}(\mathcal{H})} + \ln \frac{2}{\delta}\right)},$$

A bound on the risk

Theorem (Bound on the risk)

Let $\delta \in (0,1)$ and $n > d_{VC}(\mathcal{H})$. W.p. at least $1-\delta$, we have

$$\forall f \in \mathcal{H}, \quad \mathcal{R}(f) \leq \mathcal{R}_n(f) + 2\sqrt{\frac{2}{n}\left(d_{VC}(\mathcal{H}) \ln \frac{2en}{d_{VC}(\mathcal{H})} + \ln \frac{2}{\delta}\right)},$$

- depends on $\mathcal{H}$, not on the unknown distribution $\mathrm{P}(z)$
- strict consistency as long as $d_{VC}(\mathcal{H}) < \infty$
- should have $n \gg d_{VC}(\mathcal{H})$ to avoid over-fitting
- how close is f_n to f_0?

A bound on the risk

Theorem (Bound on the risk)

Let $\delta \in (0,1)$ and $n > d_{VC}(\mathcal{H})$. W.p. at least $1-\delta$, we have

$$\forall f \in \mathcal{H}, \quad \mathcal{R}(f) \leq \mathcal{R}_n(f) + 2\sqrt{\frac{2}{n}\left(d_{VC}(\mathcal{H}) \ln \frac{2en}{d_{VC}(\mathcal{H})} + \ln \frac{2}{\delta}\right)},$$

- depends on $\mathcal{H}$, not on the unknown distribution $\mathrm{P}(z)$
- strict consistency as long as $d_{VC}(\mathcal{H}) < \infty$
- should have $n \gg d_{VC}(\mathcal{H})$ to avoid over-fitting
- how close is f_n to f_0?

A bound on the risk

Theorem (Bound on the risk)

Let $\delta \in (0,1)$ and $n > d_{VC}(\mathcal{H})$. W.p. at least $1-\delta$, we have

$$\forall f \in \mathcal{H}, \quad \mathcal{R}(f) \leq \mathcal{R}_n(f) + 2\sqrt{\frac{2}{n}\left(d_{VC}(\mathcal{H}) \ln \frac{2en}{d_{VC}(\mathcal{H})} + \ln \frac{2}{\delta}\right)},$$

Theorem

Let $\delta \in (0,1)$ and $\epsilon > 0$. We have

$$\mathrm{P}\left(\sup_{f \in \mathcal{H}} |\mathcal{R}(f) - \mathcal{R}_n(f)| \leq \epsilon\right) \geq 1 - \delta$$

$$\Rightarrow \mathrm{P}\left(\mathcal{R}(f_n) - \mathcal{R}(f_0) \leq 2\epsilon\right) \geq 1 - \delta$$

Focus on classification

- usually, $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{Y} = \{1, \ldots, K\}$
- binary loss is the natural loss, $L(y, f(x)) = \mathbb{I}_{\{y \neq f(x)\}}$, associated risk

$$\mathcal{R}(f) = \mathrm{E}\left[L(Y, f(X))\right] = \mathrm{E}\left[\mathbb{I}_{\{y \neq f(x)\}}\right] = \mathrm{P}\left(Y \neq f(X)\right)$$

- associated empirical risk

$$\mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{y_i \neq f(x_i)\}}$$

Problems

1. hard to design $\mathcal{H} \subset \{1, \ldots, K\}^{\mathcal{X}}$
2. hard to optimize the resulting risk (not smooth, not convex)

Focus on classification

- usually, $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{Y} = \{1, \ldots, K\}$
- binary loss is the natural loss, $L(y, f(x)) = \mathbb{I}_{\{y \neq f(x)\}}$, associated risk

$$\mathcal{R}(f) = \mathrm{E}\left[L(Y, f(X))\right] = \mathrm{E}\left[\mathbb{I}_{\{y \neq f(x)\}}\right] = \mathrm{P}\left(Y \neq f(X)\right)$$

- associated empirical risk

$$\mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{y_i \neq f(x_i)\}}$$

Problems

1. hard to design $\mathcal{H} \subset \{1, \ldots, K\}^{\mathcal{X}}$
2. hard to optimize the resulting risk (not smooth, not convex)

Designing $\mathcal{H}$

- w.l.o., assume that $\mathcal{Y} = \{-1, 1\}$
- recall the linear parameterization

$$\mathcal{H} = \left\{ f_\alpha : x \to \alpha^\top \phi(x), \alpha \in \mathbb{R}^d \right\}$$

- define, for classification

$$\mathcal{G} = \{ g_\alpha : x \to \text{sgn}(f_\alpha(x)), f_\alpha \in \mathcal{H} \} \subset \{-1, 1\}^{\mathcal{X}}$$

- the resulting optimization problem is not simple

$$\min_{\alpha \in \mathbb{R}^d} \mathcal{R}_n(g_\alpha) = \min_{\alpha \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{y_i \neq \text{sgn}(f_\alpha(x))\}}$$

Designing $\mathcal{H}$

- w.l.o., assume that $\mathcal{Y} = \{-1, 1\}$
- recall the linear parameterization

$$\mathcal{H} = \left\{ f_\alpha : x \to \alpha^\top \phi(x), \alpha \in \mathbb{R}^d \right\}$$

- define, for classification

$$\mathcal{G} = \{ g_\alpha : x \to \text{sgn}(f_\alpha(x)), f_\alpha \in \mathcal{H} \} \subset \{-1, 1\}^{\mathcal{X}}$$

- the resulting optimization problem is not simple

$$\min_{\alpha \in \mathbb{R}^d} \mathcal{R}_n(g_\alpha) = \min_{\alpha \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{y_i \neq \text{sgn}(f_\alpha(x))\}}$$

Designing $\mathcal{H}$

- w.l.o., assume that $\mathcal{Y} = \{-1, 1\}$
- recall the linear parameterization

$$\mathcal{H} = \left\{ f_\alpha : x \to \alpha^\top \phi(x), \alpha \in \mathbb{R}^d \right\}$$

- define, for classification

$$\mathcal{G} = \{ g_\alpha : x \to \text{sgn}(f_\alpha(x)), f_\alpha \in \mathcal{H} \} \subset \{-1, 1\}^{\mathcal{X}}$$

- the resulting optimization problem is not simple

$$\min_{\alpha \in \mathbb{R}^d} \mathcal{R}_n(g_\alpha) = \min_{\alpha \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{y_i \neq \text{sgn}(f_\alpha(x))\}}$$

Designing $\mathcal{H}$

- w.l.o., assume that $\mathcal{Y} = \{-1, 1\}$
- recall the linear parameterization

$$\mathcal{H} = \left\{ f_\alpha : x \to \alpha^\top \phi(x), \alpha \in \mathbb{R}^d \right\}$$

- define, for classification

$$\mathcal{G} = \{ g_\alpha : x \to \text{sgn}(f_\alpha(x)), f_\alpha \in \mathcal{H} \} \subset \{-1, 1\}^{\mathcal{X}}$$

- the resulting optimization problem is not simple

$$\min_{\alpha \in \mathbb{R}^d} \mathcal{R}_n(g_\alpha) = \min_{\alpha \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{y_i \neq \text{sgn}(f_\alpha(x))\}}$$

Surrogate

	$\varphi(t)$ for $t \in \mathbb{R}$
hinge loss	$\max(0, 1-t)$
truncated least-squares	$(\max(0, 1-t))^2$
least-squares	$(1-t)^2$
exponential loss	e^{-t}
sigmoid loss	$1 - \tanh(t)$
logistic loss	$\ln(1+e^{-t})$

- original problem

$$\mathcal{R}_n(g_\alpha) = \min_{\alpha \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{y_i \neq \text{sgn}(f_\alpha(x))\}}$$

- exponential surrogate

$$\mathcal{R}(f_\alpha) = \mathrm{E}\left[e^{-Yf_\alpha(X)}\right] \text{ and } \mathcal{R}_n(f_\alpha) = \frac{1}{n}\sum_{i=1}^{n} e^{-y_i f_\alpha(x_i)}.$$

Surrogate

	$\varphi(t)$ for $t \in \mathbb{R}$
hinge loss	$\max(0, 1-t)$
truncated least-squares	$(\max(0, 1-t))^2$
least-squares	$(1-t)^2$
exponential loss	e^{-t}
sigmoid loss	$1 - \tanh(t)$
logistic loss	$\ln(1+e^{-t})$

- original problem

$$\mathcal{R}_n(g_\alpha) = \min_{\alpha \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{y_i \neq \text{sgn}(f_\alpha(x))\}}$$

- exponential surrogate

$$\mathcal{R}(f_\alpha) = \mathrm{E}\left[e^{-Yf_\alpha(X)}\right] \text{ and } \mathcal{R}_n(f_\alpha) = \frac{1}{n} \sum_{i=1}^{n} e^{-y_i f_\alpha(x_i)}.$$

Surrogate

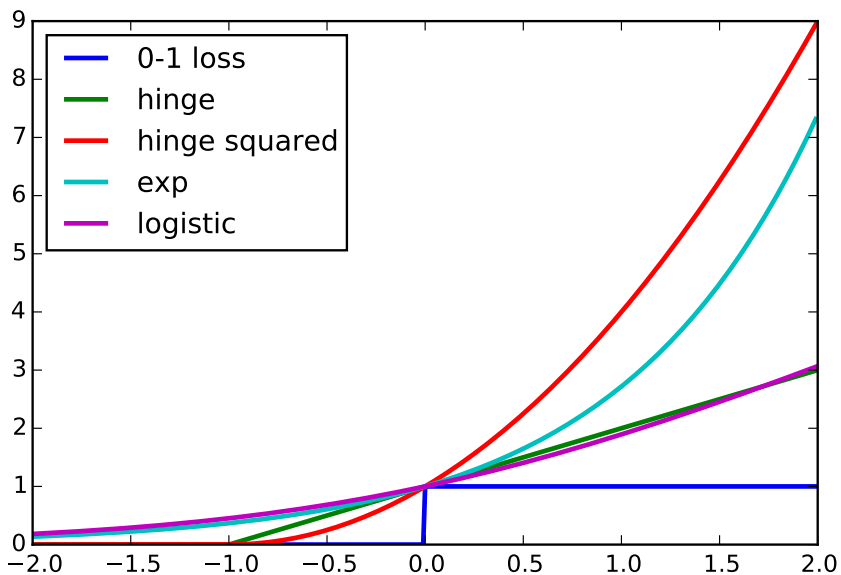
	$\varphi(t)$ for $t \in \mathbb{R}$
hinge loss	$\max(0, 1-t)$
truncated least-squares	$(\max(0, 1-t))^2$
least-squares	$(1-t)^2$
exponential loss	e^{-t}
sigmoid loss	$1-\tanh(t)$
logistic loss	$\ln(1+e^{-t})$

- more generally

$$\mathcal{R}(f_\alpha) = \mathrm{E}\left[\varphi(Yf_\alpha(X))\right] \text{ and } \mathcal{R}_n(f_\alpha) = \frac{1}{n}\sum_{i=1}^{n}\varphi(y_i f_\alpha(x_i))$$

Surrogate

	$\varphi(t)$ for $t \in \mathbb{R}$
hinge loss	$\max(0, 1-t)$
truncated least-squares	$(\max(0, 1-t))^2$
least-squares	$(1-t)^2$
exponential loss	e^{-t}
sigmoid loss	$1-\tanh(t)$
logistic loss	$\ln(1+e^{-t})$



A general problem

- $\mathcal{G} \subset \{1, .., K\}^{\mathcal{X}}$
- let $c(x, g(x), y)$ be the cost of assigning label $g(x)$ to input x when the oracle provide the response y
  (binary loss: $c(x, g(x), y) = \mathbb{I}_{\{y \neq g(x)\}}$)
- the (empirical) risk is defined as

$$\mathcal{R}(g) = \mathrm{E}\left[c(X, g(X), Y)\right]$$

$$\mathcal{R}_n(g) = \frac{1}{n} \sum_{i=1}^{n} c(x_i, g(x_i), y_i)$$

A general problem

- $\mathcal{G} \subset \{1, .., K\}^{\mathcal{X}}$
- let $c(x, g(x), y)$ be the cost of assigning label $g(x)$ to input x when the oracle provide the response y
  (binary loss: $c(x, g(x), y) = \mathbb{I}_{\{y \neq g(x)\}}$)
- the (empirical) risk is defined as

$$\mathcal{R}(g) = \mathrm{E}\left[c(X, g(X), Y)\right]$$

$$\mathcal{R}_n(g) = \frac{1}{n} \sum_{i=1}^{n} c(x_i, g(x_i), y_i)$$

A general problem

- $\mathcal{G} \subset \{1, .., K\}^{\mathcal{X}}$
- let $c(x, g(x), y)$ be the cost of assigning label $g(x)$ to input x when the oracle provide the response y
  (binary loss: $c(x, g(x), y) = \mathbb{I}_{\{y \neq g(x)\}}$)
- the (empirical) risk is defined as

$$\mathcal{R}(g) = \mathrm{E}\left[c(X, g(X), Y)\right]$$

$$\mathcal{R}_n(g) = \frac{1}{n} \sum_{i=1}^{n} c(x_i, g(x_i), y_i)$$

Designing $\mathcal{H}$

- recall the linear parameterization

$$\mathcal{H} = \left\{ f_\alpha : x \to \alpha^\top \phi(x), \alpha \in \mathbb{R}^d \right\}$$

- Write $f_{\alpha_1,\ldots,\alpha_K} : \mathcal{X} \to \mathbb{R}^K$ the function

$$f_{\alpha_1,\ldots,\alpha_K}(x) = \begin{pmatrix} f_{\alpha_1}(x) & \ldots & f_{\alpha_K}(x) \end{pmatrix}^\top$$

- Define

$$\mathcal{G} = \left\{ g_{\alpha_1,\ldots,\alpha_K} : x \to \operatorname*{argmax}_{1 \leq k \leq K} f_{\alpha_k}(x), \forall 1 \leq k \leq K : f_{\alpha_k} \in \mathcal{H} \right\}$$

- usually, the constraint $\sum_{k=1}^K f_{\alpha_k} = 0$ is added

Designing $\mathcal{H}$

- recall the linear parameterization

$$\mathcal{H} = \left\{ f_\alpha : x \to \alpha^\top \phi(x), \alpha \in \mathbb{R}^d \right\}$$

- Write $f_{\alpha_1,\ldots,\alpha_K} : \mathcal{X} \to \mathbb{R}^K$ the function

$$f_{\alpha_1,\ldots,\alpha_K}(x) = \begin{pmatrix} f_{\alpha_1}(x) & \ldots & f_{\alpha_K}(x) \end{pmatrix}^\top$$

- Define

$$\mathcal{G} = \left\{ g_{\alpha_1,\ldots,\alpha_K} : x \to \operatorname*{argmax}_{1 \leq k \leq K} f_{\alpha_k}(x), \forall 1 \leq k \leq K : f_{\alpha_k} \in \mathcal{H} \right\}$$

- usually, the constraint $\sum_{k=1}^K f_{\alpha_k} = 0$ is added

Designing $\mathcal{H}$

- recall the linear parameterization

$$\mathcal{H} = \left\{ f_\alpha : x \to \alpha^\top \phi(x), \alpha \in \mathbb{R}^d \right\}$$

- Write $f_{\alpha_1,\ldots,\alpha_K} : \mathcal{X} \to \mathbb{R}^K$ the function

$$f_{\alpha_1,\ldots,\alpha_K}(x) = \begin{pmatrix} f_{\alpha_1}(x) & \ldots & f_{\alpha_K}(x) \end{pmatrix}^\top$$

- Define

$$\mathcal{G} = \left\{ g_{\alpha_1,\ldots,\alpha_K} : x \to \operatorname*{argmax}_{1 \leq k \leq K} f_{\alpha_k}(x), \forall 1 \leq k \leq K : f_{\alpha_k} \in \mathcal{H} \right\}$$

- usually, the constraint $\sum_{k=1}^K f_{\alpha_k} = 0$ is added

Designing $\mathcal{H}$

- recall the linear parameterization

$$\mathcal{H} = \left\{ f_\alpha : x \to \alpha^\top \phi(x), \alpha \in \mathbb{R}^d \right\}$$

- Write $f_{\alpha_1,\ldots,\alpha_K} : \mathcal{X} \to \mathbb{R}^K$ the function

$$f_{\alpha_1,\ldots,\alpha_K}(x) = \begin{pmatrix} f_{\alpha_1}(x) & \ldots & f_{\alpha_K}(x) \end{pmatrix}^\top$$

- Define

$$\mathcal{G} = \left\{ g_{\alpha_1,\ldots,\alpha_K} : x \to \operatorname*{argmax}_{1 \leq k \leq K} f_{\alpha_k}(x), \forall 1 \leq k \leq K : f_{\alpha_k} \in \mathcal{H} \right\}$$

- usually, the constraint $\sum_{k=1}^K f_{\alpha_k} = 0$ is added

surrogate

	$\psi(s)$ for $s \in \mathbb{R}$
hinge loss	$\max(0, 1+s)$
truncated least-squares	$(\max(0, 1+s))^2$
least-squares	$(1+s)^2$
exponential loss	e^s
logistic loss	$\ln(1+e^s)$

- a classic surrogate

$$\mathcal{R}_n(f_{\alpha_1,\ldots,\alpha_n}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} c(x_i, k, y_i)\psi(f_{\alpha_k}(x_i))$$

- another possible surrogate

$$\mathcal{R}_n(f_{\alpha_1,\ldots,\alpha_n}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} c(x_i, k, y_i)e^{f_{\alpha_k}(x_i)-f_{\alpha_{y_i}}(x_i)}$$

- knowing what proxy to use is an open question

surrogate

	$\psi(s)$ for $s \in \mathbb{R}$
hinge loss	$\max(0, 1+s)$
truncated least-squares	$(\max(0, 1+s))^2$
least-squares	$(1+s)^2$
exponential loss	e^s
logistic loss	$\ln(1+e^s)$

- a classic surrogate

$$\mathcal{R}_n(f_{\alpha_1,\ldots,\alpha_n}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} c(x_i, k, y_i)\psi(f_{\alpha_k}(x_i))$$

- another possible surrogate

$$\mathcal{R}_n(f_{\alpha_1,\ldots,\alpha_n}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} c(x_i, k, y_i)e^{f_{\alpha_k}(x_i)-f_{\alpha_{y_i}}(x_i)}$$

- knowing what proxy to use is an open question

surrogate

	$\psi(s)$ for $s \in \mathbb{R}$
hinge loss	$\max(0, 1+s)$
truncated least-squares	$(\max(0, 1+s))^2$
least-squares	$(1+s)^2$
exponential loss	e^s
logistic loss	$\ln(1+e^s)$

- a classic surrogate

$$\mathcal{R}_n(f_{\alpha_1,\ldots,\alpha_n}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} c(x_i, k, y_i)\psi(f_{\alpha_k}(x_i))$$

- another possible surrogate

$$\mathcal{R}_n(f_{\alpha_1,\ldots,\alpha_n}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} c(x_i, k, y_i)e^{f_{\alpha_k}(x_i)-f_{\alpha_{y_i}}(x_i)}$$

- knowing what proxy to use is an open question

- Does optimizing the proxy makes sense?
- Calibration:

$$\mathcal{R}_\varphi(f) \leq \delta(\epsilon) \Rightarrow \mathcal{R}(f) \leq \epsilon$$

- Does optimizing the proxy makes sense?
- Calibration:

$$\mathcal{R}_\varphi(f) \leq \delta(\epsilon) \Rightarrow \mathcal{R}(f) \leq \epsilon$$

- Choosing the hypothesis space $\mathcal{H}$
  - small space: low bias but high variance
  - large space: low variance but high bias
- There's a trade-off, that can be handled through regularization
- let $\Omega : \mathcal{H} \to \mathbb{R}$ be a function that penalizes the complexity of a candidate solution

$$J_n(f) = \mathcal{R}_n(f) + \lambda\Omega(f)$$

- alternative viewpoint

$$\min_{f \in \mathcal{H}:\Omega(f) \leq \eta} \mathcal{R}_n(f)$$

- Choosing the hypothesis space $\mathcal{H}$
  - small space: low bias but high variance
  - large space: low variance but high bias
- There's a trade-off, that can be handled through regularization
- let $\Omega : \mathcal{H} \to \mathbb{R}$ be a function that penalizes the complexity of a candidate solution

$$J_n(f) = \mathcal{R}_n(f) + \lambda\Omega(f)$$

- alternative viewpoint

$$\min_{f \in \mathcal{H}:\Omega(f) \leq \eta} \mathcal{R}_n(f)$$

- Choosing the hypothesis space $\mathcal{H}$
  - small space: low bias but high variance
  - large space: low variance but high bias
- There's a trade-off, that can be handled through regularization
- let $\Omega : \mathcal{H} \to \mathbb{R}$ be a function that penalizes the complexity of a candidate solution

$$J_n(f) = \mathcal{R}_n(f) + \lambda\Omega(f)$$

- alternative viewpoint

$$\min_{f \in \mathcal{H}:\Omega(f) \leq \eta} \mathcal{R}_n(f)$$

- Choosing the hypothesis space $\mathcal{H}$
  - small space: low bias but high variance
  - large space: low variance but high bias
- There's a trade-off, that can be handled through regularization
- let $\Omega : \mathcal{H} \to \mathbb{R}$ be a function that penalizes the complexity of a candidate solution

$$J_n(f) = \mathcal{R}_n(f) + \lambda\Omega(f)$$

- alternative viewpoint

$$\min_{f \in \mathcal{H}:\Omega(f) \leq \eta} \mathcal{R}_n(f)$$

Assume here a space of parameterized functions, $\mathcal{H} = \{f_\alpha : \mathcal{X} \to \mathbb{R}, \alpha \in \mathbb{R}^d\}$

- ℓ_2-penalization

$$\Omega(f_\alpha) = \|\alpha\|_2^2 = \sum_{j=1}^{d} \alpha_j^2$$

- ℓ_1-penalization

$$\Omega(f_\alpha) = \|\alpha\|_0 = \text{Card}\left(\{j \in \{1, \ldots, d\} : \alpha_j \neq 0\}\right)$$

- ℓ_1-penalization

$$\Omega(f_\alpha) = \|\alpha\|_1 = \sum_{j=1}^{d} |\alpha_j|.$$

Assume here a space of parameterized functions, $\mathcal{H} = \{f_\alpha : \mathcal{X} \to \mathbb{R}, \alpha \in \mathbb{R}^d\}$

- ℓ_2-penalization

$$\Omega(f_\alpha) = \|\alpha\|_2^2 = \sum_{j=1}^{d} \alpha_j^2$$

- ℓ_1-penalization

$$\Omega(f_\alpha) = \|\alpha\|_0 = \text{Card}\left(\{j \in \{1, \ldots, d\} : \alpha_j \neq 0\}\right)$$

- ℓ_1-penalization

$$\Omega(f_\alpha) = \|\alpha\|_1 = \sum_{j=1}^{d} |\alpha_j|.$$

Assume here a space of parameterized functions, $\mathcal{H} = \{f_\alpha : \mathcal{X} \to \mathbb{R}, \alpha \in \mathbb{R}^d\}$

- ℓ_2-penalization

$$\Omega(f_\alpha) = \|\alpha\|_2^2 = \sum_{j=1}^{d} \alpha_j^2$$

- ℓ_1-penalization

$$\Omega(f_\alpha) = \|\alpha\|_0 = \text{Card}\left(\{j \in \{1, \ldots, d\} : \alpha_j \neq 0\}\right)$$

- ℓ_1-penalization

$$\Omega(f_\alpha) = \|\alpha\|_1 = \sum_{j=1}^{d} |\alpha_j|.$$

Assume here a space of parameterized functions, $\mathcal{H} = \{f_\alpha : \mathcal{X} \to \mathbb{R}, \alpha \in \mathbb{R}^d\}$

- ℓ_2-penalization

$$\Omega(f_\alpha) = \|\alpha\|_2^2 = \sum_{j=1}^{d} \alpha_j^2$$

- ℓ_1-penalization

$$\Omega(f_\alpha) = \|\alpha\|_0 = \text{Card}\left(\{j \in \{1, \ldots, d\} : \alpha_j \neq 0\}\right)$$

- ℓ_1-penalization

$$\Omega(f_\alpha) = \|\alpha\|_1 = \sum_{j=1}^{d} |\alpha_j|.$$