# Topics in High-Dimensional Statistics

Lecture 3: Empirical Risk Minimization I
*Introduction*

**Contents**

This is the first of a series of lectures on Empirical Risk Minimization (ERM) for Statistical Learning. We start by describing the problem of Statistical Learning in an abstract fashion and detail a number of examples from supervised and unsupervised learning.

## 1 Statistical learning

The general problem of statistical learning can be described as follows. Consider a **parameter set** $\Theta$ (also called **model** or **decision space**), an **outcome set** $\mathcal{Z}$ and a **cost function**[1]

$$\gamma : \Theta \times \mathcal{Z} \to \mathbb{R}.$$

Consider a generic $\mathcal{Z}$-valued random variable $Z$ with distribution $P$. In this setup, the goal of statistical learning is to solve the optimization problem

$$\min_{\theta \in \Theta} \mathbb{E}[\gamma(\theta, Z)],$$

with only sample access to $P$, i.e., given only an independent data set, or **learning sample**,

$$\mathcal{D}_n := \{Z_i\}_{i=1}^n,$$

composed of $\mathcal{Z}$-valued and i.i.d. random variables with distribution $P$, and supposed independent from $Z$. The goal is to build a $\Theta$-valued $\theta_n$, based on the learning sample, for which the **risk**, defined by

$$R(\theta_n) := \mathbb{E}[\gamma(\theta_n, Z)|\mathcal{D}_n], \tag{1.1}$$

is as close as possible, which high-probability or in expectation to the **optimal risk**

$$R^* := \inf_{\theta \in \Theta} \mathbb{E}[\gamma(\theta, Z)].$$

The (random) positive quantity

$$\mathcal{E}(\theta_n) := R(\theta_n) - R^*,$$

is called **excess risk** in the sequel.

[1] The cost function is also referred to as a contrast function. At this level of generality, we deliberately do not refer to $\gamma$ as to a loss function, a terminology dedicated to another object that typically appears in supervised learning as seen later on.

## 2 On the definition of the risk

For any data dependent $\theta_n$, the definition of the risk

$$R(\theta_n) = \mathbb{E}[\gamma(\theta_n, Z)|\mathcal{D}_n] = \mathbb{E}[\gamma(\theta_n, Z)|Z_1, \ldots, Z_n], \tag{2.1}$$

introduced in (1.1), involves a conditional expectation. In particular the risk $R(\theta_n)$ is a random variable which, by construction of conditional expectation, is a measurable function of the learning sample. To clarify the definition, and its uses in the sequel, let us comment further on this definition. To avoid confusion in manipulating the risk in the sequel, the best possible definition of $R(\theta_n)$ is as follows. First, define the deterministic function $\varphi : \Theta \to \mathbb{R}$ by

$$\varphi(\theta) := \int_{\mathcal{Z}} \gamma(\theta, z)\, \mathrm{d}P(z).$$

Then, for any $\theta_n$, possibly depending on the independent data $\mathcal{D}_n$, we set

$$R(\theta_n) := \varphi(\theta_n). \tag{2.2}$$

The equivalence between representations (2.1) and (2.2) follows from the following result.

**Lemma 2.1.** *Let $\phi : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ be a measurable map. Let $U$ and $V$ be two independent random variables, defined on probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values respectively in $\mathcal{U}$ and $\mathcal{V}$. Then, provided $\phi$ is positive or $\mathbb{E}[|\phi(U, V)|] < +\infty$, and defining $\varphi : \mathcal{U} \to \mathbb{R}$ by*

$$\varphi(u) := \int \phi(u, v)\, \mathrm{d}\mathbb{P}_V(v),$$

*we have*

$$\mathbb{E}[\phi(U, V)|U] = \varphi(U).$$

**Remark 2.2.** *An alternative definition used for the risk, in many textbooks, is*

$$R(\theta_n) = \mathbb{E}_{Z\sim P}[\gamma(\theta_n, Z)].$$

*This definition and notation can be the source of many mistakes if not used properly It is usually fine if, as in our context, the generic random variable $Z$ is independent from the learning sample $\{Z_i\}_{i=1}^n$ but can lead to trouble in different settings. For this reason, we'll refrain from using this improper notation.*

## 3 Empirical risk minimization

The construction of the learners decision $\theta_n$ can usually be represented formally as

$$\theta_n = A_n(Z_1, \ldots, Z_n),$$

where $\mathcal{A} = (A_m)_{m \geq 1}$ denotes a **statistical learning algorithm**, i.e., a sequence of functions

$$A_m : \mathcal{Z}^m \to \Theta.$$

The most natural statistical learning algorithm in this context is known as **Empirical Risk Minimization** (abbreviated ERM). This algorithm is defined, for all $m \geq 1$ and all $\{z_i\}_{i=1}^m \in \mathcal{Z}^m$, by

$$A_m(z_1, \ldots, z_m) \in \arg\min_{\theta \in \Theta} \frac{1}{m} \sum_{i=1}^{m} \gamma(\theta, z_i),$$

leading to

$$\theta_n^{\mathrm{ERM}} \in \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \gamma(\theta, Z_i).$$

The goal of this lecture is precisely to study the performance of this algorithm by establishing upper bounds on the excess risk

$$\mathcal{E}(\theta_n^{\mathrm{ERM}}),$$

in relation to three aspects of the problem:
- the sample size $n$,
- the complexity of the parameter set $\Theta$,
- the analytical properties of the cost function $\gamma$.

## 4 Example: Supervised learning

A standard statistical learning setting is known as **supervised learning**. In this setting, the parameter space $\Theta$ is typically a set of functions from an input space $\mathcal{X}$ to an output space $\mathcal{Y}$, and the outcome space is the product space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The available data set, or learning sample, take the form of i.i.d. random pairs, i.e.,

$$\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n,$$

with same (and unknown) distribution $P$ as a generic random pair $(X, Y)$, taking values in $\mathcal{X} \times \mathcal{Y}$.

The variable $X$ usually comes in the form of a vector of measurements (i.e., $\mathcal{X} \subset \mathbb{R}^d$ for some $d \geq 2$), associated to a certain phenomenon (e.g., parameters of a certain incoming email) and the variable $Y$, called the **label** of $X$, is usually a real number (i.e., $\mathcal{Y} \subset \mathbb{R}$) that summarizes some important information encoded in vector $X$ (e.g., $Y = 1$ if the incoming email is a spam and 0 otherwise). In this setup, the learner's objective is to learn a functional link between $X$ and $Y$, based on a learning sample $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$. This is usually formalized as follows: given a **loss function**

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R},$$

and the set $\Theta$ of functions $\theta : \mathcal{X} \to \mathcal{Y}$, the learner has to choose $\theta_n \in \Theta$ based on the learning sample such that the risk

$$R(\theta_n) := \mathbb{E}[\ell(Y, \theta_n(X))|\mathcal{D}_n],$$

is as small as possible.

Note that this corresponds exactly to the general setup introduced above with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and cost function $\gamma : \Theta \times \mathcal{Z} \to \mathbb{R}$ defined by

$$\gamma(\theta, (x, y)) := \ell(y, \theta(x)).$$

Common examples of supervised learning problems include least-squares regression and binary classification.

### Least-squares regression

Least-squares regression refers to the supervised learning problem when the chosen loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is

$$\ell(y, u) := (y - u)^2.$$

It is usually considered when the label is of continuous nature (i.e., $\mathcal{Y}$ is an interval).

### Binary classification

Binary classification (also called pattern recognition) refers to the case where the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is[2]

$$\ell(y, u) := \mathbf{1}\{y \neq u\},$$

[2]Where it is understood that $\mathbf{1}\{...\}$ equal 1 when ... is true and 0 otherwise.

and is considered when the label is of binary nature (i.e., $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{-1, 1\}$).

**Large margin classification**

Large margin classification is an important alternative to binary classification as it provides a convex formulation of the problem. It refers to the case where the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is of the form

$$\ell(y, u) := \phi(-yu),$$

for an auxiliary convex function $\phi : \mathbb{R} \to \mathbb{R}$ and it is considered when the label is of binary nature (i.e., $\mathcal{Y} = \{-1, 1\}$).

## 5 Example: $k$-means clustering

The **$k$-means clustering** problem is a classical problem of "unsupervised" learning (i.e., statistical learning with unlabeled data). In $k$-means clustering, the learner is given:

- an integer $k \geq 2$,
- a learning sample $\mathcal{D}_n = \{X_i\}_{i=1}^n$ of i.i.d. random variables (independent and with same distribution as a generic and independent random variable $X$) taking values in $\mathcal{X} = \mathbb{R}^d$ (for example) with $n \geq k$,

and is asked to chose a collection $\theta_n$ of $k$ points in $\mathbb{R}^d$, i.e.,

$$\theta_n = (\theta_n^1, \ldots, \theta_n^k) \in \Theta := (\mathbb{R}^d)^k,$$

for which the risk

$$R(\theta_n) := \mathbb{E}[\min_{1 \leq j \leq k} \|X - \theta_n^j\|^2 | \mathcal{D}_n],$$

is as small as possible. Once the centers $\theta_n^1, \ldots, \theta_n^k$ are chosen, the associated data clusters are defined by

$$\mathcal{C}_j := \{X_i\}_{i=1}^n \cap V(\theta_n^j),$$

where the sets $V(\theta_n^j)$, called Voronoi cells, are defined by

$$V(\theta_n^j) := \{x \in \mathbb{R}^d : \|x - \theta_n^j\| = \min_{1 \leq s \leq k} \|x - \theta_n^s\|\},$$

where ties are broken arbitrarily.

The $k$-means clustering problem can be seen as a special case of the general statistical learning problem where $\Theta = (\mathbb{R}^d)^k$, $\mathcal{Z} = \mathbb{R}^d$ and the cost function $\gamma : \Theta \times \mathcal{Z} \to \mathbb{R}$ is defined by

$$\gamma((\theta^1, \ldots, \theta^k), x) := \min_{1 \leq j \leq k} \|x - \theta^j\|^2.$$

## 6 Example: Density estimation via maximum likelihood

Consider a sample $\mathcal{D}_n = \{X_i\}_{i=1}^n$ of i.i.d. $\mathbb{R}^d$-valued random variables which same distribution $P$ as (and independent from) a generic random variable $X$. Suppose $X$ has (unknown) density $\theta^* : \mathbb{R}^d \to \mathbb{R}_+$ with respect to a fixed and known positive measure $\mu$ on $\mathbb{R}^d$, i.e., $\mathrm{d}P/\mathrm{d}\mu = \theta^*$. Suppose one wants to estimate $\theta^*$ based on $\mathcal{D}_n$. This problem, known as density estimation, is a classical problem in statistics. The following result provides the basic justification for the maximum likelihood method.

**Theorem 6.1.** *Let $\theta : \mathbb{R}^d \to \mathbb{R}_+$ be measurable and such that $\int_{\mathbb{R}^d} \theta(x)\,\mathrm{d}\mu(x) = 1$. Then,*

$$\mathbb{E}[\log \theta^*(X)] \geq \mathbb{E}[\log \theta(X)].$$

*Proof.* By concavity of the logarithm, we deduce from Jensen's inequality that

$$\begin{aligned}
\mathbb{E}\left[\log\left(\frac{\theta(X)}{\theta^*(X)}\right)\right] &\leq \log \mathbb{E}\left[\frac{\theta(X)}{\theta^*(X)}\right] \\
&= \log\left(\int_{\mathbb{R}^d} \frac{\theta(x)}{\theta^*(x)} \theta^*(x)\,\mathrm{d}\mu(x)\right) \\
&= \log\left(\int_{\mathbb{R}^d} \theta(x)\,\mathrm{d}\mu(x)\right) \\
&= 0,
\end{aligned}$$

which concludes the proof. □

The last result shows that the problem of density estimation can be seen as an example of the general statistical learning problem described above with $\Theta$ being any set of density functions with respect to $\mu$, i.e.,

$$\Theta \subset \{\theta : \mathbb{R}^d \to \mathbb{R}_+ | \int_{\mathbb{R}^d} \theta\,\mathrm{d}\mu = 1\},$$

with $\mathcal{Z} = \mathbb{R}^d$ and with cost function $\gamma : \Theta \times \mathcal{Z} \to \mathbb{R}$ defined by

$$\gamma(\theta, x) = -\log \theta(x).$$

In particular, it follows from the previous computations that if $\theta^* \in \Theta$, the excess risk of any density estimator $\theta_n$ is precisely the Kullback-Leibler divergence between $\theta^*$ and $\theta_n$, i.e.,

$$\mathcal{E}(\theta_n) = \int_{\mathbb{R}^d} \theta^* \log \frac{\theta_n}{\theta^*} \, \mathrm{d}\mu = \mathrm{KL}(\theta^* \| \theta_n).$$

Note finally that the empirical risk minimizer $\hat{\theta}_n^{\mathrm{ERM}}$ reduces in this case to the popular maximum likelihood estimator

$$\hat{\theta}_n^{\mathrm{ERM}} \in \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \gamma(\theta, X_i) = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ln \theta(X_i).$$

## 7 Example: Density estimation via square loss minimization

In the setting of the previous section, an alternative point of view is often considered based on the following observation. Consider

$$\bar{\theta} \in \arg\min_{\theta \in \Theta} \int_{\mathbb{R}^d} (\theta - \theta^*)^2 \, \mathrm{d}\mu,$$

i.e., the closest element in model $\Theta$ to the true density $\theta^*$ in $L^2(P)$-norm. Then, we have the following.

**Lemma 7.1.**

$$\bar{\theta} \in \arg\min_{\theta \in \Theta} \Big( \int_{\mathbb{R}^d} \theta^2 \, \mathrm{d}\mu - 2 \int_{\mathbb{R}^d} \theta \, \mathrm{d}P \Big).$$

*Proof.* Developing the square, we observe that for every $\theta \in \Theta$,

$$\begin{aligned} \int_{\mathbb{R}^d} (\theta - \theta^*)^2 \, \mathrm{d}\mu &= \int_{\mathbb{R}^d} \theta^2 \, \mathrm{d}\mu - 2 \int_{\mathbb{R}^d} \theta\theta^* \, \mathrm{d}\mu + \int_{\mathbb{R}^d} (\theta^*)^2 \, \mathrm{d}\mu \\ &= \int_{\mathbb{R}^d} \theta^2 \, \mathrm{d}\mu - 2 \int_{\mathbb{R}^d} \theta \, \mathrm{d}P + \int_{\mathbb{R}^d} (\theta^*)^2 \, \mathrm{d}\mu, \end{aligned}$$

which clearly implies that

$$\arg\min_{\theta \in \Theta} \int_{\mathbb{R}^d} (\theta - \theta^*)^2 \, \mathrm{d}\mu = \arg\min_{\theta \in \Theta} \Big( \int_{\mathbb{R}^d} \theta^2 \, \mathrm{d}\mu - 2 \int_{\mathbb{R}^d} \theta \, \mathrm{d}P \Big).$$

□

From this point of view, the problem of density estimation can be cast as a statistical learning problem with

$$\Theta \subset \{\theta : \mathbb{R}^d \to \mathbb{R}_+ | \int_{\mathbb{R}^d} \theta \, \mathrm{d}\mu = 1\}, \qquad \mathcal{Z} = \mathbb{R}^d,$$

and

$$\gamma(\theta, x) = \int_{\mathbb{R}^d} \theta^2 \, \mathrm{d}\mu - 2\theta(x).$$

The risk of a given density function $\theta \in \Theta$ is then

$$R(\theta) = \int_{\mathbb{R}^d} \theta^2 \, \mathrm{d}\mu - 2 \int_{\mathbb{R}^d} \theta \, \mathrm{d}P.$$

Note finally that the empirical risk minimizer $\theta_n^{\mathrm{ERM}}$ becomes in this case in this case

$$\theta_n^{\mathrm{ERM}} \in \arg\min_{\theta \in \Theta} \{ \int_{\mathbb{R}^d} \theta^2 \, \mathrm{d}\mu - \frac{2}{n} \sum_{i=1}^{n} \theta(X_i) \}.$$