# Geometric Aspects of Statistical Learning Theory

Statistical learning theory plays a central role in modern data science, and the question we focus on in this course has been the key question in the area since the late 60s.

To describe the problem, let $F$ be a class of functions defined on a probability space $(\Omega, \mu)$, and consider a random variable $Y$. The goal is to find some function that is almost as close to $Y$ as the best approximation to $Y$ in $F$. The twist is the limited information at one's disposal: while the class $F$ is given, $Y$ is not known, nor is the underlying measure $\mu$. Instead, if $X$ is distributed according to $\mu$ and $(X,Y)$ is the joint distribution of $X$ and $Y$, one is given $N$ independent sample points $\{(X_1,Y_1),...,(X_N,Y_N)\}$, selected according to $(X,Y)$. Using the sample and the identity of $F$, one has to generate a (data-dependent) function $\hat{f}$ that approximates $Y$. The success of the choice is determined by the accuracy-confidence tradeoff: one has to ensure that with probability $1-\delta$ with respect to $(X_i,Y_i)_{i=1}^N$,

$$\mathbb{E}\left((\hat{f}(X)-Y)^2|(X_i,Y_i)_{i=1}^N\right) \leq \inf_{f\in F}\mathbb{E}(f(X)-Y)^2+\varepsilon. \tag{1}$$

Naturally, high accuracy and high confidence are conflicting requirements: the higher the wanted accuracy, the more difficult it is to ensure that (1) holds with high confidence.

**Question:** Given a class $F$, a distribution $(X,Y)$, and a sample size $N$, what is the optimal tradeoff between the wanted accuracy $\varepsilon$ and the confidence $1-\delta$? And, what is the right choice of $\hat{f}$ that attains the optimal tradeoff?

The aim of this course is to show that this question has a strong geometric flavour and to highlight some of the ideas in empirical processes theory and in asymptotic geometric analysis that have led to its solution—under minimal assumptions on the class $F$ and on $(X,Y)$.

## The plan

(1) Why is learning possible? The definition of a learning problem; what can we hope for; the quadratic and multiplier processes; complexity measures of classes of functions (2 hours).

(2) The small-ball method and (some of) its applications (4 hours).

(3) Median-of-means tournaments and the solution for convex classes (2 hours).

(4) Complexity measures of classes revisited: chaining methods for Bernoulli and gaussian processes; combinatorial dimension and metric entropy (4 hours).

## Prerequisites

The course will require the knowledge of (graduate level) probability/measure theory and functional analysis, as well as some mathematical maturity. Most of the material I will cover can be found in the course's lecture notes. Because of the nature of the course, some of the details will be left for the students.

The original articles containing (some of) the results I will present in the course are

S. Mendelson, R. Vershynin, Entropy and the combinatorial dimension, Inventiones Mathematicae, 152(1), 37-55, 2003.

S. Mendelson, S. Mendelson, Learning without concentration, Journal of the ACM, 62(3), Article No. 21, 1-25, 2015.

V. Koltchinskii, S. Mendelson, Bounding the smallest singular value of a random matrix without concentration, International Mathematics Research Notices, Vol. 2015 (23), 12991–13008, 2015.

G. Lugosi, S. Mendelson, Risk minimization by median-of-means tournaments, Journal of the EMS, https://arxiv.org/abs/1608.00757.

S. Mendelson, An optimal unrestricted learning procedure, https://arxiv.org/abs/1707.05342.

Additional material can be found in the books:

M. Ledoux, M. Talagrand, Probability in Banach spaces.

A. Van der Vaart, J. Wellner, Weak Convergence and Empirical Processes.

R. Vershynin, High dimensional probability — an introduction with applications in Data Science.