Introduction to Statistical Learning Theory

Olivier Bousquet[1], Stéphane Boucheron[2], and Gábor Lugosi[3]

[1] Max-Planck Institute for Biological Cybernetics
Spemannstr. 38, D-72076 Tübingen, Germany
olivier.bousquet@m4x.org
WWW home page: http://www.kyb.mpg.de/~bousquet

[2] Université de Paris-Sud, Laboratoire d'Informatique
Bâtiment 490, F-91405 Orsay Cedex, France
stephane.boucheron@lri.fr
WWW home page: http://www.lri.fr/~bouchero

[3] Department of Economics, Pompeu Fabra University
Ramon Trias Fargas 25-27, 08005 Barcelona, Spain
lugosi@upf.es
WWW home page: http://www.econ.upf.es/~lugosi

Abstract. The goal of statistical learning theory is to study, in a statistical framework, the properties of learning algorithms. In particular, most results take the form of so-called error bounds. This tutorial introduces the techniques that are used to obtain such results.

1 Introduction

The main goal of statistical learning theory is to provide a framework for studying the problem of inference, that is of gaining knowledge, making predictions, making decisions or constructing models from a set of data. This is studied in a statistical framework, that is there are assumptions of statistical nature about the underlying phenomena (in the way the data is generated).

As a motivation for the need of such a theory, let us just quote V. Vapnik:

> (Vapnik, [1]) Nothing is more practical than a good theory.

Indeed, a theory of inference should be able to give a formal definition of words like learning, generalization, overfitting, and also to characterize the performance of learning algorithms so that, ultimately, it may help design better learning algorithms.
There are thus two goals: make things more precise and derive new or improved algorithms.

1.1 Learning and Inference

What is under study here is the process of inductive inference which can roughly be summarized as the following steps:

1. Observe a phenomenon
2. Construct a model of that phenomenon
3. Make predictions using this model

Of course, this definition is very general and could be taken more or less as the goal of Natural Sciences. The goal of Machine Learning is to actually *automate* this process and the goal of Learning Theory is to *formalize* it.

In this tutorial we consider a special case of the above process which is the supervised learning framework for pattern recognition. In this framework, the data consists of instance-label pairs, where the label is either $+1$ or -1. Given a set of such pairs, a learning algorithm constructs a function mapping instances to labels. This function should be such that it makes few mistakes when predicting the label of unseen instances.

Of course, given some training data, it is always possible to build a function that fits exactly the data. But, in the presence of noise, this may not be the best thing to do as it would lead to a poor performance on unseen instances (this is usually referred to as overfitting). The general idea behind the design of

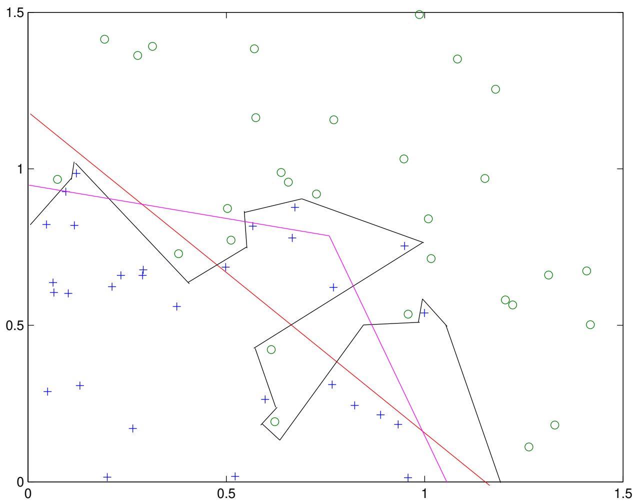

Fig. 1. Trade-off between fit and complexity.

learning algorithms is thus to look for *regularities* (in a sense to be defined later) in the observed phenomenon (i.e. training data). These can then be *generalized* from the observed past to the future. Typically, one would look, in a collection of possible models, for one which fits well the data, but at the same time is as simple as possible (see Figure 1). This immediately raises the question of how to measure and quantify simplicity of a model (i.e. a $\{-1, +1\}$-valued function).

It turns out that there are many ways to do so, but no best one. For example in Physics, people tend to prefer models which have a small number of constants and that correspond to simple mathematical formulas. Often, the length of description of a model in a coding language can be an indication of its complexity. In classical statistics, the number of free parameters of a model is usually a measure of its complexity. Surprisingly as it may seem, there is no universal way of measuring simplicity (or its counterpart complexity) and the choice of a specific measure inherently depends on the problem at hand. It is actually in this choice that the designer of the learning algorithm introduces knowledge about the specific phenomenon under study.

This lack of universally best choice can actually be formalized in what is called the *No Free Lunch* theorem, which in essence says that, if there is no assumption on how the past (i.e. training data) is related to the future (i.e. test data), prediction is impossible. Even more, if there is no a priori restriction on the possible phenomena that are expected, it is impossible to generalize and there is thus no better algorithm (any algorithm would be beaten by another one on some phenomenon).

Hence the need to make assumptions, like the fact that the phenomenon we observe can be explained by a simple model. However, as we said, simplicity is not an absolute notion, and this leads to the statement that data cannot replace knowledge, or in pseudo-mathematical terms:

$$\text{Generalization} = \text{Data} + \text{Knowledge}$$

1.2 Assumptions

We now make more precise the assumptions that are made by the Statistical Learning Theory framework. Indeed, as we said before we need to assume that the future (i.e. test) observations are related to the past (i.e. training) ones, so that the phenomenon is somewhat stationary.

At the core of the theory is a probabilistic model of the phenomenon (or data generation process). Within this model, the relationship between past and future observations is that they both are sampled independently from the same distribution (i.i.d.). The independence assumption means that each new observation yields maximum information. The identical distribution means that the observations give information about the underlying phenomenon (here a probability distribution).

An immediate consequence of this very general setting is that one can construct algorithms (e.g. k-nearest neighbors with appropriate k) that are *consistent*, which means that, as one gets more and more data, the predictions of the algorithm are closer and closer to the optimal ones. So this seems to indicate that we can have some sort of universal algorithm. Unfortunately, any (consistent) algorithm can have an arbitrarily bad behavior when given a finite training set. These notions are formalized in Appendix B.

Again, this discussion indicates that generalization can only come when one adds specific knowledge to the data. Each learning algorithm encodes specific knowledge (or a specific assumption about how the optimal classifier looks like), and works best when this assumption is satisfied by the problem to which it is applied.

Bibliographical remarks. Several textbooks, surveys, and research monographs have been written on pattern classification and statistical learning theory. A partial list includes Anthony and Bartlett [2], Breiman, Friedman, Olshen, and Stone [3], Devroye, Györfi, and Lugosi [4], Duda and Hart [5], Fukunaga [6], Kearns and Vazirani [7], Kulkarni, Lugosi, and Venkatesh [8], Lugosi [9], McLachlan [10], Mendelson [11], Natarajan [12], Vapnik [13, 14, 1], and Vapnik and Chervonenkis [15].

2 Formalization

We consider an input space $\mathcal{X}$ and output space $\mathcal{Y}$. Since we restrict ourselves to binary classification, we choose $\mathcal{Y} = \{-1, 1\}$. Formally, we assume that the pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ are random variables distributed according to an *unknown* distribution P. We observe a sequence of n i.i.d. pairs (X_i, Y_i) sampled according to P and the goal is to construct a function $g : \mathcal{X} \to \mathcal{Y}$ which *predicts* Y from X.

We need a criterion to choose this function g. This criterion is a low probability of error $P(g(X) \neq Y)$. We thus define the *risk* of g as

$$R(g) = P(g(X) \neq Y) = \mathbb{E}\left[\mathbb{1}_{g(X)\neq Y}\right] .$$

Notice that P can be decomposed as $P_X \times P(Y|X)$. We introduce the *regression function* $\eta(x) = \mathbb{E}\left[Y|X = x\right] = 2\mathbb{P}\left[Y = 1|X = x\right] - 1$ and the *target function* (or Bayes classifier) $t(x) = \operatorname{sgn} \eta(x)$. This function achieves the minimum risk over all possible measurable functions:

$$R(t) = \inf_g R(g) .$$

We will denote the value $R(t)$ by R^*, called the Bayes risk. In the deterministic case, one has $Y = t(X)$ almost surely ($\mathbb{P}\left[Y = 1|X\right] \in \{0, 1\}$) and $R^* = 0$. In the general case we can define the *noise level* as $s(x) = \min(\mathbb{P}\left[Y = 1|X = x\right], 1 - \mathbb{P}\left[Y = 1|X = x\right]) = (1 - \eta(x))/2$ ($s(X) = 0$ almost surely in the deterministic case) and this gives $R^* = \mathbb{E}s(X)$.

Our goal is thus to identify this function t, but since P is unknown we cannot directly measure the risk and we also cannot know directly the value of t at the data points. We can only measure the agreement of a candidate function with the data. This is called the *empirical risk*:

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{g(X_i)\neq Y_i} .$$

It is common to use this quantity as a criterion to select an estimate of t.

2.1 Algorithms

Now that the goal is clearly specified, we review the common strategies to (approximately) achieve it. We denote by g_n the function returned by the algorithm.

Because one cannot compute $R(g)$ but only approximate it by $R_n(g)$, it would be unreasonable to look for the function minimizing $R_n(g)$ among all possible functions. Indeed, when the input space is infinite, one can always construct a function g_n which perfectly predicts the labels of the training data (i.e. $g_n(X_i) = Y_i$, and $R_n(g_n) = 0$), but behaves on the other points as the opposite of the target function t, i.e. $g_n(X) = -Y$ so that $R(g_n) = 1$[4]. So one would have minimum empirical risk but maximum risk.

It is thus necessary to prevent this overfitting situation. There are essentially two ways to do this (which can be combined). The first one is to restrict the class of functions in which the minimization is performed, and the second is to modify the criterion to be minimized (e.g. adding a penalty for 'complicated' functions).

Empirical Risk Minimization. This algorithm is one of the most straightforward, yet it is usually efficient. The idea is to choose a *model* $\mathcal{G}$ of possible functions and to minimize the empirical risk in that model:

$$g_n = \arg \min_{g \in \mathcal{G}} R_n(g) .$$

Of course, this will work best when the target function belongs to $\mathcal{G}$. However, it is rare to be able to make such an assumption, so one may want to enlarge the model as much as possible, while preventing overfitting.

Structural Risk Minimization. The idea here is to choose an infinite sequence $\{\mathcal{G}_d : d = 1, 2, \ldots\}$ of models of increasing size and to minimize the empirical risk in each model with an added penalty for the size of the model:

$$g_n = \arg \min_{g \in \mathcal{G}_d, d \in \mathbb{N}} R_n(g) + \mathrm{pen}(d, n) .$$

The penalty $\mathrm{pen}(d, n)$ gives preference to models where estimation error is small and measures the size or *capacity* of the model.

Regularization. Another, usually easier to implement approach consists in choosing a large model $\mathcal{G}$ (possibly dense in the continuous functions for example) and to define on $\mathcal{G}$ a *regularizer*, typically a norm $\|g\|$. Then one has to minimize the regularized empirical risk:

$$g_n = \arg \min_{g \in \mathcal{G}} R_n(g) + \lambda \|g\|^2 .$$

[4] Strictly speaking this is only possible if the probability distribution satisfies some mild conditions (e.g. has no atoms). Otherwise, it may not be possible to achieve $R(g_n) = 1$ but even in this case, provided the support of P contains infinitely many points, a similar phenomenon occurs.

Compared to SRM, there is here a free parameter λ, called the *regularization parameter* which allows to choose the right trade-off between fit and complexity. Tuning λ is usually a hard problem and most often, one uses extra validation data for this task.

Most existing (and successful) methods can be thought of as regularization methods.

Normalized Regularization. There are other possible approaches when the regularizer can, in some sense, be 'normalized', i.e. when it corresponds to some probability distribution over $\mathcal{G}$.
Given a probability distribution π defined on $\mathcal{G}$ (usually called a prior), one can use as a regularizer $-\log \pi(g)$[5]. Reciprocally, from a regularizer of the form $\|g\|^2$, if there exists a measure μ on $\mathcal{G}$ such that $\int e^{-\lambda\|g\|^2} d\mu(g) < \infty$ for some $\lambda > 0$, then one can construct a prior corresponding to this regularizer. For example, if $\mathcal{G}$ is the set of hyperplanes in $\mathbb{R}^d$ going through the origin, $\mathcal{G}$ can be identified with $\mathbb{R}^d$ and, taking μ as the Lebesgue measure, it is possible to go from the Euclidean norm regularizer to a spherical Gaussian measure on $\mathbb{R}^d$ as a prior[6].

This type of normalized regularizer, or prior, can be used to construct another probability distribution ρ on $\mathcal{G}$ (usually called posterior), as

$$\rho(g) = \frac{e^{-\gamma R_n(g)}}{Z(\gamma)} \pi(g) ,$$

where $\gamma \geq 0$ is a free parameter and $Z(\gamma)$ is a normalization factor.

There are several ways in which this ρ can be used. If we take the function maximizing it, we recover regularization as

$$\arg\max_{g \in \mathcal{G}} \rho(g) = \arg\min_{g \in \mathcal{G}} \gamma R_n(g) - \log \pi(g) ,$$

where the regularizer is $-\gamma^{-1} \log \pi(g)$[7].

Also, ρ can be used to *randomize* the predictions. In that case, before computing the predicted label for an input x, one samples a function g according to ρ and outputs $g(x)$. This procedure is usually called Gibbs classification.

Another way in which the distribution ρ constructed above can be used is by taking the expected prediction of the functions in $\mathcal{G}$:

$$g_n(x) = \operatorname{sgn}(\mathbb{E}_\rho(g(x))) .$$

[5] This is fine when $\mathcal{G}$ is countable. In the continuous case, one has to consider the density associated to π. We omit these details.

[6] Generalization to infinite dimensional Hilbert spaces can also be done but it requires more care. One can for example establish a correspondence between the norm of a reproducing kernel Hilbert space and a Gaussian process prior whose covariance function is the kernel of this space.

[7] Note that minimizing $\gamma R_n(g) - \log \pi(g)$ is equivalent to minimizing $R_n(g) - \gamma^{-1} \log \pi(g)$.

This is typically called Bayesian averaging.

At this point we have to insist again on the fact that the choice of the class $\mathcal{G}$ and of the associated regularizer or prior, has to come from *a priori* knowledge about the task at hand, and there is no universally best choice.

2.2 Bounds

We have presented the framework of the theory and the type of algorithms that it studies, we now introduce the kind of results that it aims at. The overall goal is to characterize the risk that some algorithm may have in a given situation. More precisely, a learning algorithm takes as input the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and produces a function g_n which depends on this data. We want to estimate the risk of g_n. However, $R(g_n)$ is a random variable (since it depends on the data) and it cannot be computed from the data (since it also depends on the unknown P). Estimates of $R(g_n)$ thus usually take the form of probabilistic bounds.

Notice that when the algorithm chooses its output from a model $\mathcal{G}$, it is possible, by introducing the best function g^* in $\mathcal{G}$, with $R(g^*) = \inf_{g\in\mathcal{G}} R(g)$, to write

$$R(g_n) - R^* = [R(g^*) - R^*] + [R(g_n) - R(g^*)] .$$

The first term on the right hand side is usually called the approximation error, and measures how well can functions in $\mathcal{G}$ approach the target (it would be zero if $t \in \mathcal{G}$). The second term, called estimation error is a random quantity (it depends on the data) and measures how close is g_n to the best possible choice in $\mathcal{G}$.

Estimating the approximation error is usually hard since it requires knowledge about the target. Classically, in Statistical Learning Theory it is preferable to avoid making specific assumptions about the target (such as its belonging to some model), but the assumptions are rather on the value of R^*, or on the noise function s.

It is also known that for any (consistent) algorithm, the rate of convergence to zero of the approximation error[8] can be arbitrarily slow if one does not make assumptions about the regularity of the target, while the rate of convergence of the estimation error can be computed without any such assumption. We will thus focus on the estimation error.

Another possible decomposition of the risk is the following:

$$R(g_n) = R_n(g_n) + [R(g_n) - R_n(g_n)] .$$

In this case, one estimates the risk by its empirical counterpart, and some quantity which approximates (or upper bounds) $R(g_n) - R_n(g_n)$.

To summarize, we write the three type of results we may be interested in.

[8] For this converge to mean anything, one has to consider algorithms which choose functions from a class which grows with the sample size. This is the case for example of Structural Risk Minimization or Regularization based algorithms.

- *Error bound*: $R(g_n) \leq R_n(g_n) + B(n, \mathcal{G})$. This corresponds to the estimation of the risk from an empirical quantity.
- *Error bound relative to the best in the class*: $R(g_n) \leq R(g^*) + B(n, \mathcal{G})$. This tells how "optimal" is the algorithm given the model it uses.
- *Error bound relative to the Bayes risk*: $R(g_n) \leq R^* + B(n, \mathcal{G})$. This gives theoretical guarantees on the convergence to the Bayes risk.

3 Basic Bounds

In this section we show how to obtain simple error bounds (also called generalization bounds). The elementary material from probability theory that is needed here and in the later sections is summarized in Appendix A.

3.1 Relationship to Empirical Processes

Recall that we want to estimate the risk $R(g_n) = \mathbb{E}\left[\mathbb{1}_{g_n(X) \neq Y}\right]$ of the function g_n returned by the algorithm after seeing the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. This quantity cannot be observed (P is unknown) and is a random variable (since it depends on the data). Hence one way to make a statement about this quantity is to say how it relates to an estimate such as the empirical risk $R_n(g_n)$. This relationship can take the form of upper and lower bounds for

$$\mathbb{P}\left[R(g_n) - R_n(g_n) > \varepsilon\right] .$$

For convenience, let $Z_i = (X_i, Y_i)$ and $Z = (X, Y)$. Given $\mathcal{G}$ define the *loss class*

$$\mathcal{F} = \{f : (x, y) \mapsto \mathbb{1}_{g(x) \neq y} : g \in \mathcal{G}\} . \tag{1}$$

Notice that $\mathcal{G}$ contains functions with range in $\{-1, 1\}$ while $\mathcal{F}$ contains non-negative functions with range in $\{0, 1\}$. In the remainder of the tutorial, we will go back and forth between $\mathcal{F}$ and $\mathcal{G}$ (as there is a bijection between them), sometimes stating the results in terms of functions in $\mathcal{F}$ and sometimes in terms of functions in $\mathcal{G}$. It will be clear from the context which classes $\mathcal{G}$ and $\mathcal{F}$ we refer to, and $\mathcal{F}$ will always be derived from the last mentioned class $\mathcal{G}$ in the way of (1).

We use the shorthand notation $Pf = \mathbb{E}\left[f(X, Y)\right]$ and $P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i, Y_i)$. P_n is usually called the *empirical measure* associated to the training sample. With this notation, the quantity of interest (difference between true and empirical risks) can be written as

$$Pf_n - P_n f_n . \tag{2}$$

An empirical process is a collection of random variables indexed by a class of functions, and such that each random variable is distributed as a sum of i.i.d. random variables (values taken by the function at the data):

$$\{Pf - P_n f\}_{f \in \mathcal{F}} .$$

One of the most studied quantity associated to empirical processes is their supremum:

$$\sup_{f \in \mathcal{F}} Pf - P_n f \, .$$

It is clear that if we know an upper bound on this quantity, it will be an upper bound on (2). This shows that the theory of empirical processes is a great source of tools and techniques for Statistical Learning Theory.

3.2 Hoeffding's Inequality

Let us rewrite again the quantity we are interested in as follows

$$R(g) - R_n(g) = \mathbb{E}\left[f(Z)\right] - \frac{1}{n}\sum_{i=1}^{n} f(Z_i) \, .$$

It is easy to recognize here the difference between the expectation and the empirical average of the random variable $f(Z)$. By the law of large numbers, we immediately obtain that

$$\mathbb{P}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}\left[f(Z)\right] = 0\right] = 1 \, .$$

This indicates that with enough samples, the empirical risk of a function is a good approximation to its true risk.
It turns out that there exists a quantitative version of the law of large numbers when the variables are bounded.

Theorem 1 (Hoeffding). *Let $Z_1, \ldots, Z_n$ be n i.i.d. random variables with $f(Z) \in [a, b]$. Then for all $\varepsilon > 0$, we have*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}\left[f(Z)\right]\right| > \varepsilon\right] \le 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) \, .$$

Let us rewrite the above formula to better understand its consequences. Denote the right hand side by δ. Then

$$\mathbb{P}\left[|P_n f - Pf| > (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\right] \le \delta \, ,$$

or (by inversion, see Appendix A) with probability at least $1-\delta$,

$$|P_n f - Pf| \le (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \, .$$

Applying this to $f(Z) = \mathbb{1}_{g(X) \neq Y}$ we get that for any g, and any $\delta > 0$, with probability at least $1 - \delta$

$$R(g) \leq R_n(g) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \,. \tag{3}$$

Notice that one has to consider a fixed function g and the probability is with respect to the sampling of the data. If the function depends on the data this does not apply!

3.3 Limitations

Although the above result seems very nice (since it applies to any class of bounded functions), it is actually severely limited. Indeed, what it essentially says is that for each (fixed) function $f \in \mathcal{F}$, there is a set S of samples for which $Pf - P_n f \leq \sqrt{\frac{\log \frac{2}{\delta}}{2n}}$ (and this set of samples has measure $\mathbb{P}[S] \geq 1 - \delta$). However, these sets S may be different for different functions. In other words, for the observed sample, only some of the functions in $\mathcal{F}$ will satisfy this inequality.

Another way to explain the limitation of Hoeffding's inequality is the following. If we take for $\mathcal{G}$ the class of all $\{-1, 1\}$-valued (measurable) functions, then for any fixed sample, there exists a function $f \in \mathcal{F}$ such that

$$Pf - P_n f = 1 \,.$$

To see this, take the function which is $f(X_i) = Y_i$ on the data and $f(X) = -Y$ everywhere else. This does not contradict Hoeffding's inequality but shows that it does not yield what we need.

Figure 2 illustrates the above argumentation. The horizontal axis corresponds

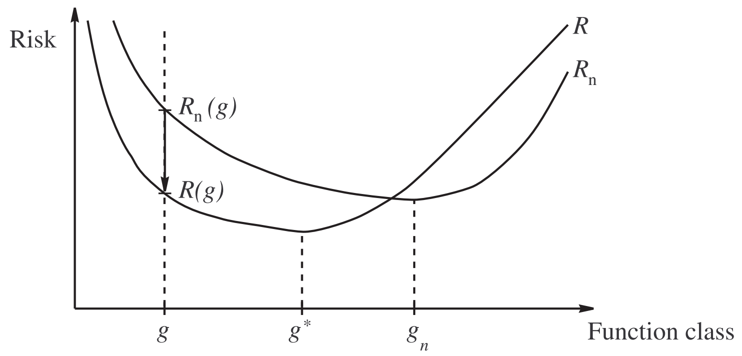

Fig. 2. Convergence of the empirical risk to the true risk over the class of functions.

to the functions in the class. The two curves represent the true risk and the empirical risk (for some training sample) of these functions. The true risk is fixed, while for each different sample, the empirical risk will be a different curve. If we observe a fixed function g and take several different samples, the point on the empirical curve will fluctuate around the true risk with fluctuations controlled by Hoeffding's inequality. However, for a fixed sample, if the class $\mathcal{G}$ is big enough, one can find somewhere along the axis, a function for which the difference between the two curves will be very large.

3.4 Uniform Deviations

Before seeing the data, we do not know which function the algorithm will choose. The idea is to consider *uniform* deviations

$$R(f_n) - R_n(f_n) \leq \sup_{f \in \mathcal{F}} (R(f) - R_n(f)) \tag{4}$$

In other words, if we can upper bound the supremum on the right, we are done. For this, we need a bound which holds simultaneously for all functions in a class.

Let us explain how one can construct such uniform bounds. Consider two functions f_1, f_2 and define

$$C_i = \{(x_1, y_1), \ldots, (x_n, y_n) : P f_i - P_n f_i > \varepsilon\} \; .$$

This set contains all the 'bad' samples, i.e. those for which the bound fails. From Hoeffding's inequality, for each i

$$\mathbb{P}\left[C_i\right] \leq \delta \, .$$

We want to measure how many samples are 'bad' for $i = 1$ or $i = 2$. For this we use (see Appendix A)

$$\mathbb{P}\left[C_1 \cup C_2\right] \leq \mathbb{P}\left[C_1\right] + \mathbb{P}\left[C_2\right] \leq 2\delta \, .$$

More generally, if we have N functions in our class, we can write

$$\mathbb{P}\left[C_1 \cup \ldots \cup C_N\right] \leq \sum_{i=1}^{N} \mathbb{P}\left[C_i\right]$$

As a result we obtain

$$\begin{aligned}
&\mathbb{P}\left[\exists f \in \{f_1, \ldots, f_N\} : Pf - P_n f > \varepsilon\right] \\
&\leq \sum_{i=1}^{N} \mathbb{P}\left[P f_i - P_n f_i > \varepsilon\right] \\
&\leq N \exp\left(-2n\varepsilon^2\right)
\end{aligned}$$

Hence, for $\mathcal{G} = \{g_1, \ldots, g_N\}$, for all $\delta > 0$ with probability at least $1 - \delta$,

$$\forall g \in \mathcal{G},\ R(g) \leq R_n(g) + \sqrt{\frac{\log N + \log \frac{1}{\delta}}{2n}}$$

This is an error bound. Indeed, if we know that our algorithm picks functions from $\mathcal{G}$, we can apply this result to g_n itself.

Notice that the main difference with Hoeffding's inequality is the extra $\log N$ term on the right hand side. This is the term which accounts for the fact that we want N bounds to hold simultaneously. Another interpretation of this term is as the number of bits one would require to specify one function in $\mathcal{G}$. It turns out that this kind of coding interpretation of generalization bounds is often possible and can be used to obtain error estimates [16].

3.5 Estimation Error

Using the same idea as before, and with no additional effort, we can also get a bound on the estimation error. We start from the inequality

$$R(g^*) \leq R_n(g^*) + \sup_{g \in \mathcal{G}} (R(g) - R_n(g)),$$

which we combine with (4) and with the fact that since g_n minimizes the empirical risk in $\mathcal{G}$,

$$R_n(g^*) - R_n(g_n) \geq 0$$

Thus we obtain

$$\begin{aligned} R(g_n) &= R(g_n) - R(g^*) + R(g^*) \\ &\leq R_n(g^*) - R_n(g_n) + R(g_n) - R(g^*) + R(g^*) \\ &\leq 2 \sup_{g \in \mathcal{G}} |R(g) - R_n(g)| + R(g^*) \end{aligned}$$

We obtain that with probability at least $1 - \delta$

$$R(g_n) \leq R(g^*) + 2\sqrt{\frac{\log N + \log \frac{2}{\delta}}{2n}}.$$

We notice that in the right hand side, both terms depend on the size of the class $\mathcal{G}$. If this size increases, the first term will decrease, while the second will increase.

3.6 Summary and Perspective

At this point, we can summarize what we have exposed so far.

– Inference requires to put assumptions on the process generating the data (data sampled i.i.d. from an unknown P), generalization requires knowledge (e.g. restriction, structure, or prior).

– The error bounds are valid with respect to the repeated sampling of training sets.
– For a fixed function g, for most of the samples

$$R(g) - R_n(g) \approx 1/\sqrt{n}$$

– For most of the samples if $|\mathcal{G}| = N$

$$\sup_{g \in \mathcal{G}} R(g) - R_n(g) \approx \sqrt{\log N/n}$$

The extra variability comes from the fact that the chosen g_n changes with the data.

So the result we have obtained so far is that with high probability, for a finite class of size N,

$$\sup_{g \in \mathcal{G}} (R(g) - R_n(g)) \leq \sqrt{\frac{\log N + \log \frac{1}{\delta}}{2n}} .$$

There are several things that can be improved:

– Hoeffding's inequality only uses the boundedness of the functions, not their variance.
– The union bound is as bad as if all the functions in the class were independent (i.e. if $f_1(Z)$ and $f_2(Z)$ were independent).
– The supremum over $\mathcal{G}$ of $R(g) - R_n(g)$ is not necessarily what the algorithm would choose, so that upper bounding $R(g_n) - R_n(g_n)$ by the supremum might be loose.

4 Infinite Case: Vapnik-Chervonenkis Theory

In this section we show how to extend the previous results to the case where the class $\mathcal{G}$ is infinite. This requires, in the non-countable case, the introduction of tools from Vapnik-Chervonenkis Theory.

4.1 Refined Union Bound and Countable Case

We first start with a simple refinement of the union bound that allows to extend the previous results to the (countably) infinite case.

Recall that by Hoeffding's inequality, for each $f \in \mathcal{F}$, for each $\delta > 0$ (possibly depending on f, which we write $\delta(f)$),

$$\mathbb{P}\left[Pf - P_n f > \sqrt{\frac{\log \frac{1}{\delta(f)}}{2n}} \right] \leq \delta(f) .$$

Hence, if we have a countable set $\mathcal{F}$, the union bound immediately yields

$$\mathbb{P}\left[\exists f \in \mathcal{F} : Pf - P_n f > \sqrt{\frac{\log \frac{1}{\delta(f)}}{2n}}\right] \leq \sum_{f \in \mathcal{F}} \delta(f)\,.$$

Choosing $\delta(f) = \delta p(f)$ with $\sum_{f \in \mathcal{F}} p(f) = 1$, this makes the right-hand side equal to δ and we get the following result. With probability at least $1 - \delta$,

$$\forall f \in \mathcal{F},\ Pf \leq P_n f + \sqrt{\frac{\log \frac{1}{p(f)} + \log \frac{1}{\delta}}{2n}}\,.$$

We notice that if $\mathcal{F}$ is finite (with size N), taking a uniform p gives the $\log N$ as before.

Using this approach, it is possible to put knowledge about the algorithm into $p(f)$, but p should be chosen before seeing the data, so it is not possible to 'cheat' by setting all the weight to the function returned by the algorithm after seeing the data (which would give the smallest possible bound). But, in general, if p is well-chosen, the bound will have a small value. Hence, the bound can be improved if one knows ahead of time the functions that the algorithm is likely to pick (i.e. knowledge improves the bound).

4.2 General Case

When the set $\mathcal{G}$ is uncountable, the previous approach does not directly work. The general idea is to look at the function class 'projected' on the sample. More precisely, given a sample $z_1, \ldots, z_n$, we consider

$$\mathcal{F}_{z_1,\ldots,z_n} = \{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\}$$

The size of this set is the number of possible ways in which the data $(z_1, \ldots, z_n)$ can be classified. Since the functions f can only take two values, this set will always be finite, no matter how big $\mathcal{F}$ is.

Definition 1 (Growth function). *The growth function is the maximum number of ways into which n points can be classified by the function class:*

$$S_{\mathcal{F}}(n) = \sup_{(z_1,\ldots,z_n)} |\mathcal{F}_{z_1,\ldots,z_n}|\,.$$

We have defined the growth function in terms of the loss class $\mathcal{F}$ but we can do the same with the initial class $\mathcal{G}$ and notice that $S_{\mathcal{F}}(n) = S_{\mathcal{G}}(n)$.

It turns out that this growth function can be used as a measure of the 'size' of a class of function as demonstrated by the following result.

Theorem 2 (Vapnik-Chervonenkis). *For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\forall g \in \mathcal{G},\ R(g) \leq R_n(g) + 2\sqrt{2\frac{\log S_{\mathcal{G}}(2n) + \log \frac{2}{\delta}}{n}}\,.$$

Notice that, in the finite case where $|\mathcal{G}| = N$, we have $S_{\mathcal{G}}(n) \leq N$ so that this bound is always better than the one we had before (except for the constants).

But the problem becomes now one of computing $S_{\mathcal{G}}(n)$.

4.3 VC Dimension

Since $g \in \{-1, 1\}$, it is clear that $S_{\mathcal{G}}(n) \leq 2^n$. If $S_{\mathcal{G}}(n) = 2^n$, there is a set of size n such that the class of functions can generate any classification on these points (we say that $\mathcal{G}$ *shatters* the set).

Definition 2 (VC dimension). *The* VC dimension *of a class $\mathcal{G}$ is the largest n such that*

$$S_{\mathcal{G}}(n) = 2^n .$$

In other words, the VC dimension of a class $\mathcal{G}$ is the size of the largest set that it can shatter.

In order to illustrate this definition, we give some examples. The first one is the set of half-planes in $\mathbb{R}^d$ (see Figure 3). In this case, as depicted for the case $d = 2$, one can shatter a set of $d + 1$ points but no set of $d + 2$ points, which means that the VC dimension is $d + 1$.

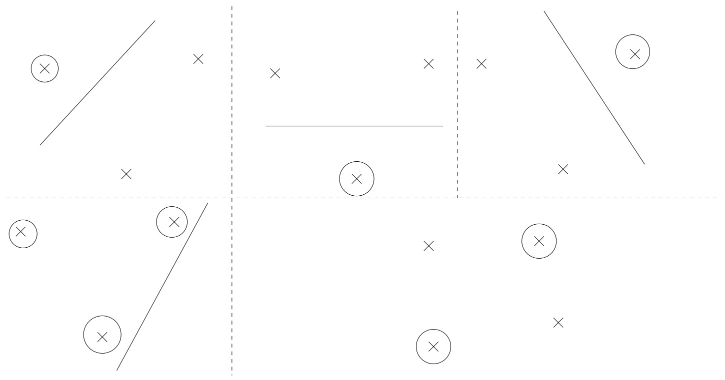

Fig. 3. Computing the VC dimension of hyperplanes in dimension 2: a set of 3 points can be shattered, but no set of four points.

It is interesting to notice that the number of parameters needed to define half-spaces in $\mathbb{R}^d$ is d, so that a natural question is whether the VC dimension is related to the number of parameters of the function class. The next example, depicted in Figure 4, is a family of functions with one parameter only:

$$\{\text{sgn}(\sin(tx)) : t \in \mathbb{R}\}$$

which actually has infinite VC dimension (this is an exercise left to the reader).

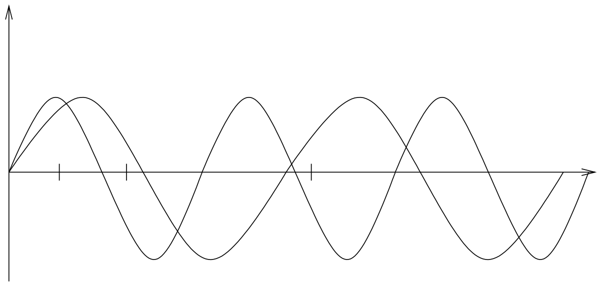

Fig. 4. VC dimension of sinusoids.

It remains to show how the notion of VC dimension can bring a solution to the problem of computing the growth function. Indeed, at first glance, if we know that a class has VC dimension h, it entails that for all $n \leq h$, $S_{\mathcal{G}}(n) = 2^n$ and $S_{\mathcal{G}}(n) < 2^n$ otherwise. This seems of little use, but actually, an intriguing phenomenon occurs for $n \geq h$ as depicted in Figure 5. The growth function

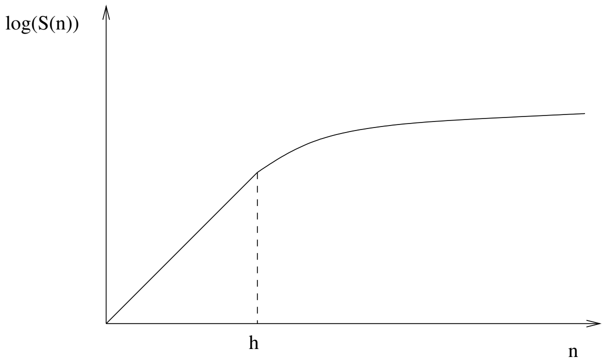

Fig. 5. Typical behavior of the log growth function.

which is exponential (its logarithm is linear) up until the VC dimension, becomes polynomial afterwards.

This behavior is captured in the following lemma.

Lemma 1 (Vapnik and Chervonenkis, Sauer, Shelah). *Let $\mathcal{G}$ be a class of functions with finite VC-dimension h. Then for all $n \in \mathbb{N}$,*

$$S_{\mathcal{G}}(n) \leq \sum_{i=0}^{h} \binom{n}{i},$$

and for all $n \geq h$,

$$S_{\mathcal{G}}(n) \leq \left(\frac{en}{h}\right)^h .$$

Using this lemma along with Theorem 2 we immediately obtain that if $\mathcal{G}$ has VC dimension h, with probability at least $1-\delta$,

$$\forall g \in \mathcal{G},\ R(g) \leq R_n(g) + 2\sqrt{2\frac{h\log\frac{2en}{h} + \log\frac{2}{\delta}}{n}} .$$

What is important to recall from this result, is that the difference between the true and empirical risk is at most of order

$$\sqrt{\frac{h\log n}{n}} .$$

An interpretation of VC dimension and growth functions is that they measure the *effective* size of the class, that is the size of the projection of the class onto finite samples. In addition, this measure does not just 'count' the number of functions in the class but depends on the geometry of the class (rather its projections). Finally, the finiteness of the VC dimension ensures that the empirical risk will converge uniformly over the class to the true risk.

4.4 Symmetrization

We now indicate how to prove Theorem 2. The key ingredient to the proof is the so-called *symmetrization* lemma. The idea is to replace the true risk by an estimate computed on an independent set of data. This is of course a mathematical technique and does not mean one needs to have more data to be able to apply the result. The extra data set is usually called 'virtual' or 'ghost sample'.

We will denote by $Z'_1, \ldots, Z'_n$ an independent (ghost) sample and by P'_n the corresponding empirical measure.

Lemma 2 (Symmetrization). *For any $t > 0$, such that $nt^2 \geq 2$,*

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}(P-P_n)f \geq t\right] \leq 2\mathbb{P}\left[\sup_{f\in\mathcal{F}}(P'_n-P_n)f \geq t/2\right] .$$

Proof. Let f_n be the function achieving the supremum (note that it depends on $Z_1, \ldots, Z_n$). One has (with $\wedge$ denoting the conjunction of two events),

$$\begin{aligned} \mathbb{1}_{(P-P_n)f_n>t}\mathbb{1}_{(P-P'_n)f_n<t/2} &= \mathbb{1}_{(P-P_n)f_n>t\,\wedge\,(P'_n-P)f_n\geq -t/2} \\ &\leq \mathbb{1}_{(P'_n-P_n)f_n>t/2} . \end{aligned}$$

Taking expectations with respect to the second sample gives

$$\mathbb{1}_{(P-P_n)f_n>t}\mathbb{P}'\left[(P-P'_n)f_n < t/2\right] \leq \mathbb{P}'\left[(P'_n-P_n)f_n > t/2\right] .$$

By Chebyshev's inequality (see Appendix A),

$$\mathbb{P}'\left[(P - P'_n)f_n \geq t/2\right] \leq \frac{4\mathsf{Var} f_n}{nt^2} \leq \frac{1}{nt^2}\,.$$

Indeed, a random variable with range in $[0,1]$ has variance less than $1/4$. Hence

$$\mathbb{1}_{(P-P_n)f_n > t}(1 - \frac{1}{nt^2}) \leq \mathbb{P}'\left[(P'_n - P_n)f_n > t/2\right]\,.$$

Taking expectation with respect to first sample gives the result. □

This lemma allows to replace the expectation Pf by an empirical average over the ghost sample. As a result, the right hand side only depends on the *projection* of the class $\mathcal{F}$ on the double sample:

$$\mathcal{F}_{Z_1,\ldots,Z_n,Z'_1,\ldots,Z'_n}\,,$$

which contains finitely many different vectors. One can thus use the simple union bound that was presented before in the finite case. The other ingredient that is needed to obtain Theorem 2 is again Hoeffding's inequality in the following form:

$$\mathbb{P}\left[P_n f - P'_n f > t\right] \leq e^{-nt^2/2}\,.$$

We now just have to put the pieces together:

$$\begin{aligned}
&\mathbb{P}\left[\sup_{f\in\mathcal{F}}(P - P_n)f \geq t\right] \\
&\leq 2\mathbb{P}\left[\sup_{f\in\mathcal{F}}(P'_n - P_n)f \geq t/2\right] \\
&= 2\mathbb{P}\left[\sup_{f\in\mathcal{F}_{Z_1,\ldots,Z_n,Z'_1,\ldots,Z'_n}}(P'_n - P_n)f \geq t/2\right] \\
&\leq 2S_{\mathcal{F}}(2n)\mathbb{P}\left[(P'_n - P_n)f \geq t/2\right] \\
&\leq 4S_{\mathcal{F}}(2n)e^{-nt^2/8}\,.
\end{aligned}$$

Using inversion finishes the proof of Theorem 2.

4.5 VC Entropy

One important aspect of the VC dimension is that it is *distribution independent*. Hence, it allows to get bounds that do not depend on the problem at hand: the same bound holds for any distribution. Although this may be seen as an advantage, it can also be a drawback since, as a result, the bound may be loose for most distributions.

We now show how to modify the proof above to get a distribution-dependent result. We use the following notation $N(\mathcal{F}, z_1^n) := |\mathcal{F}_{z_1,\ldots,z_n}|$.

Definition 3 (VC entropy). *The (annealed)* VC entropy *is defined as*

$$H_{\mathcal{F}}(n) = \log \mathbb{E}\left[N(\mathcal{F}, Z_1^n)\right]\,.$$

Theorem 3. *For any $\delta > 0$, with probability at least $1-\delta$,*

$$\forall g \in \mathcal{G},\ R(g) \leq R_n(g) + 2\sqrt{2\frac{H_{\mathcal{G}}(2n) + \log\frac{2}{\delta}}{n}}\,.$$

Proof. We again begin with the symmetrization lemma so that we have to upper bound the quantity

$$I = \mathbb{P}\left[\sup_{f \in \mathcal{F}_{Z_1^n, Z_1^{n\prime}}} (P_n' - P_n)f \geq t/2\right]\,.$$

Let $\sigma_1, \ldots, \sigma_n$ be n independent random variables such that $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$ (they are called Rademacher variables). We notice that the quantities $(P_n' - P_n)f$ and $\frac{1}{n}\sum_{i=1}^n \sigma_i(f(Z_i') - f(Z_i))$ have the same distribution since changing one σ_i corresponds to exchanging Z_i and Z_i'. Hence we have

$$I \leq \mathbb{E}\left[\mathbb{P}_\sigma\left[\sup_{f \in \mathcal{F}_{Z_1^n, Z_1^{n\prime}}} \frac{1}{n}\sum_{i=1}^n \sigma_i(f(Z_i') - f(Z_i)) \geq t/2\right]\right]\,,$$

and the union bound leads to

$$I \leq \mathbb{E}\left[N\left(\mathcal{F}, Z_1^n, Z_1^{n\prime}\right) \max_f \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i(f(Z_i') - f(Z_i)) \geq t/2\right]\right]\,.$$

Since $\sigma_i(f(Z_i') - f(Z_i)) \in [-1, 1]$, Hoeffding's inequality finally gives

$$I \leq \mathbb{E}\left[N\left(\mathcal{F}, Z, Z'\right)\right] e^{-nt^2/8}\,.$$

The rest of the proof is as before. □

5 Capacity Measures

We have seen so far three measures of *capacity* or size of classes of function: the VC dimension and growth function both distribution independent, and the VC entropy which depends on the distribution. Apart from the VC dimension, they are usually hard or impossible to compute. There are however other measures which not only may give sharper estimates, but also have properties that make their computation possible from the data only.

5.1 Covering Numbers

We start by endowing the function class $\mathcal{F}$ with the following (random) metric

$$d_n(f, f') = \frac{1}{n}|\{f(Z_i) \neq f'(Z_i) : i = 1, \ldots, n\}|\,.$$

This is the normalized Hamming distance of the 'projections' on the sample. Given such a metric, we say that a set $f_1, \ldots, f_N$ *covers* $\mathcal{F}$ at radius ε if

$$\mathcal{F} \subset \cup_{i=1}^{N} B(f_i, \varepsilon) \,.$$

We then define the covering numbers of $\mathcal{F}$ as follows.

Definition 4 (Covering number). *The covering number of $\mathcal{F}$ at radius ε, with respect to d_n, denoted by $N(\mathcal{F}, \varepsilon, n)$ is the minimum size of a cover of radius ε.*

Notice that it does not matter if we apply this definition to the original class $\mathcal{G}$ or the loss class $\mathcal{F}$, since $N(\mathcal{F}, \varepsilon, n) = N(\mathcal{G}, \varepsilon, n)$.

The covering numbers characterize the size of a function class as measured by the metric d_n. The rate of growth of the logarithm of $N(\mathcal{G}, \varepsilon, n)$ usually called the metric entropy, is related to the classical concept of vector dimension. Indeed, if $\mathcal{G}$ is a compact set in a d-dimensional Euclidean space, $N(\mathcal{G}, \varepsilon, n) \approx \varepsilon^{-d}$.

When the covering numbers are finite, it is possible to approximate the class $\mathcal{G}$ by a finite set of functions (which cover $\mathcal{G}$). Which again allows to use the finite union bound, provided we can relate the behavior of all functions in $\mathcal{G}$ to that of functions in the cover. A typical result, which we provide without proof, is the following.

Theorem 4. *For any $t > 0$,*

$$\mathbb{P}\left[\exists g \in \mathcal{G} :\, R(g) > R_n(g) + t\right] \leq 8\mathbb{E}\left[N(\mathcal{G}, t, n)\right] e^{-nt^2/128} \,.$$

Covering numbers can also be defined for classes of real-valued functions.

We now relate the covering numbers to the VC dimension. Notice that, because the functions in $\mathcal{G}$ can only take two values, for all $\varepsilon > 0$, $N(\mathcal{G}, \varepsilon, n) \leq |\mathcal{G}_{Z_1^n}| = N(\mathcal{G}, Z_1^n)$. Hence the VC entropy corresponds to log covering numbers at minimal scale, which implies $N(\mathcal{G}, \varepsilon, n) \leq h \log \frac{en}{h}$, but one can have a considerably better result.

Lemma 3 (Haussler). *Let $\mathcal{G}$ be a class of VC dimension h. Then, for all $\varepsilon > 0$, all n, and any sample,*

$$N(\mathcal{G}, \varepsilon, n) \leq Ch(4e)^h \varepsilon^{-h} \,.$$

The interest of this result is that the upper bound does not depend on the sample size n.

The covering number bound is a generalization of the VC entropy bound where the scale is adapted to the error. It turns out that this result can be improved by considering all scales (see Section 5.2).

5.2 Rademacher Averages

Recall that we used in the proof of Theorem 3 Rademacher random variables, i.e. independent $\{-1, 1\}$-valued random variables with probability $1/2$ of taking either value.

For convenience we introduce the following notation (signed empirical measure) $R_n f = \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)$. We will denote by $\mathbb{E}_\sigma$ the expectation taken with respect to the Rademacher variables (i.e. conditionally to the data) while $\mathbb{E}$ will denote the expectation with respect to all the random variables (i.e. the data, the ghost sample and the Rademacher variables).

Definition 5 (Rademacher averages). *For a class $\mathcal{F}$ of functions, the Rademacher average is defined as*

$$\mathcal{R}(\mathcal{F}) = \mathbb{E} \sup_{f\in\mathcal{F}} R_n f \,,$$

and the conditional Rademacher average is defined as

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} R_n f \,.$$

We now state the fundamental result involving Rademacher averages.

Theorem 5. *For all $\delta > 0$, with probability at least $1-\delta$,*

$$\forall f \in \mathcal{F}, Pf \le P_n f + 2\mathcal{R}(\mathcal{F}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}\,,$$

and also, with probability at least $1-\delta$,

$$\forall f \in \mathcal{F}, Pf \le P_n f + 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{2\log\frac{2}{\delta}}{n}}\,.$$

It is remarkable that one can obtain a bound (second part of the theorem) which depends solely on the data.

The proof of the above result requires a powerful tool called a concentration inequality for empirical processes.
Actually, Hoeffding's inequality is a (simple) concentration inequality, in the sense that when n increases, the empirical average is concentrated around the expectation. It is possible to generalize this result to functions that depend on i.i.d. random variables as shown in the theorem below.

Theorem 6 (McDiarmid [17]). *Assume for all $i = 1, \ldots, n$,*

$$\sup_{z_1,\ldots,z_n,z_i'} |F(z_1,\ldots,z_i,\ldots,z_n) - F(z_1,\ldots,z_i',\ldots,z_n)| \le c\,,$$

then for all $\varepsilon > 0$,

$$\mathbb{P}\left[|F - \mathbb{E}\left[F\right]| > \varepsilon\right] \le 2\exp\left(-\frac{2\varepsilon^2}{nc^2}\right)\,.$$

The meaning of this result is thus that, as soon as one has a function of n independent random variables, which is such that its variation is bounded when one variable is modified, the function will satisfy a Hoeffding-like inequality.

Proof of Theorem 5. To prove Theorem 5, we will have to follow the following three steps:

1. Use *concentration* to relate $\sup_{f \in \mathcal{F}} Pf - P_n f$ to its expectation,
2. use *symmetrization* to relate the expectation to the Rademacher average,
3. use *concentration* again to relate the Rademacher average to the conditional one.

We first show that McDiarmid's inequality can be applied to $\sup_{f \in \mathcal{F}} Pf - P_n f$. We denote temporarily by P_n^i the empirical measure obtained by modifying one element (e.g. Z_i is replaced by Z_i') of the sample. It is easy to check that the following holds

$$|\sup_{f \in \mathcal{F}}(Pf - P_n f) - \sup_{f \in \mathcal{F}}(Pf - P_n^i f)| \le \sup_{f \in \mathcal{F}} |P_n^i f - P_n f| \,.$$

Since $f \in \{0, 1\}$ we obtain

$$|P_n^i f - P_n f| = \frac{1}{n}|f(Z_i') - f(Z_i)| \le \frac{1}{n} \,,$$

and thus McDiarmid's inequality can be applied with $c = 1/n$. This concludes the first step of the proof.

We next prove the (first part of the) following symmetrization lemma.

Lemma 4. *For any class* $\mathcal{F}$*,*

$$\mathbb{E} \sup_{f \in \mathcal{F}} Pf - P_n f \le 2\mathbb{E} \sup_{f \in \mathcal{F}} R_n f \,,$$

and

$$\mathbb{E} \sup_{f \in \mathcal{F}} |Pf - P_n f| \ge \frac{1}{2} \mathbb{E} \sup_{f \in \mathcal{F}} \mathcal{R}_n f - \frac{1}{2\sqrt{n}} \,.$$

Proof. We only prove the first part. We introduce a ghost sample and its corresponding measure P_n'. We successively use the fact that $\mathbb{E} P_n' f = Pf$ and the supremum is a convex function (hence we can apply Jensen's inequality, see Appendix A):

$$\begin{aligned}
&\mathbb{E} \sup_{f \in \mathcal{F}} Pf - P_n f \\
&= \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{E}\left[P_n' f\right] - P_n f \\
&\le \mathbb{E} \sup_{f \in \mathcal{F}} P_n' f - P_n f \\
&= \mathbb{E}_\sigma \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i (f(Z_i') - f(Z_i))\right] \\
&\le \mathbb{E}_\sigma \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(Z_i')\right] + \mathbb{E}_\sigma \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n -\sigma_i f(Z_i))\right] \\
&= 2\mathbb{E} \sup_{f \in \mathcal{F}} R_n f \,.
\end{aligned}$$

where the third step uses the fact that $f(Z_i) - f(Z_i')$ and $\sigma_i(f(Z_i) - f(Z_i'))$ have the same distribution and the last step uses the fact that the $\sigma_i f(Z_i)$ and $-\sigma_i f(Z_i')$ have the same distribution. □

The above already establishes the first part of Theorem 5. For the second part, we need to use concentration again. For this we apply McDiarmid's inequality to the following functional

$$F(Z_1, \dots, Z_n) = \mathcal{R}_n(\mathcal{F}) .$$

It is easy to check that F satisfies McDiarmid's assumptions with $c = \frac{1}{n}$. As a result, $\mathbb{E}F = \mathcal{R}(\mathcal{F})$ can be sharply estimated by $F = \mathcal{R}_n(\mathcal{F})$.

Loss Class and Initial Class. In order to make use of Theorem 5 we have to relate the Rademacher average of the loss class to those of the initial class. This can be done with the following derivation where one uses the fact that σ_i and $\sigma_i Y_i$ have the same distribution.

$$\begin{aligned}
\mathcal{R}(\mathcal{F}) &= \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbb{1}_{g(X_i)\neq Y_i}\right] \\
&= \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1}{2}(1 - Y_i g(X_i))\right] \\
&= \frac{1}{2}\mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i Y_i g(X_i)\right] = \frac{1}{2}\mathcal{R}(\mathcal{G}) .
\end{aligned}$$

Notice that the same is valid for conditional Rademacher averages, so that we obtain that with probability at least $1 - \delta$,

$$\forall g \in \mathcal{G}, R(g) \leq R_n(g) + \mathcal{R}_n(\mathcal{G}) + \sqrt{\frac{2\log\frac{2}{\delta}}{n}} .$$

Computing the Rademacher Averages. We now assess the difficulty of actually computing the Rademacher averages. We write the following.

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(X_i)\right] \\
&= \frac{1}{2} + \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} -\frac{1 - \sigma_i g(X_i)}{2}\right] \\
&= \frac{1}{2} - \mathbb{E}\left[\inf_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \frac{1 - \sigma_i g(X_i)}{2}\right] \\
&= \frac{1}{2} - \mathbb{E}\left[\inf_{g\in\mathcal{G}} R_n(g, \sigma)\right] .
\end{aligned}$$

This indicates that, given a sample and a choice of the random variables $\sigma_1, \dots, \sigma_n$, computing $\mathcal{R}_n(\mathcal{G})$ is not harder than computing the empirical risk minimizer in $\mathcal{G}$. Indeed, the procedure would be to generate the σ_i randomly and minimize the empirical error in $\mathcal{G}$ with respect to the labels σ_i.

An advantage of rewriting $\mathcal{R}_n(\mathcal{G})$ as above is that it gives an intuition of what it actually measures: it measures how much the class $\mathcal{G}$ can fit random noise. If the class $\mathcal{G}$ is very large, there will always be a function which can perfectly fit the σ_i and then $\mathcal{R}_n(\mathcal{G}) = 1/2$, so that there is no hope of uniform convergence to zero of the difference between true and empirical risks.

For a finite set with $|\mathcal{G}| = N$, one can show that

$$\mathcal{R}_n(\mathcal{G}) \le 2\sqrt{\log N/n}\,,$$

where we again see the logarithmic factor $\log N$. A consequence of this is that, by considering the projection on the sample of a class $\mathcal{G}$ with VC dimension h, and using Lemma 1, we have

$$\mathcal{R}(\mathcal{G}) \le 2\sqrt{\frac{h\log\frac{en}{h}}{n}}\,.$$

This result along with Theorem 5 allows to recover the Vapnik Chervonenkis bound with a concentration-based proof.

Although the benefit of using concentration may not be entirely clear at that point, let us just mention that one can actually improve the dependence on n of the above bound. This is based on the so-called *chaining* technique. The idea is to use covering numbers at all scales in order to capture the geometry of the class in a better way than the VC entropy does.

One has the following result, called Dudley's entropy bound

$$\mathcal{R}_n(\mathcal{F}) \le \frac{C}{\sqrt{n}}\int_0^\infty \sqrt{\log N(\mathcal{F}, t, n)}\,dt\,.$$

As a consequence, along with Haussler's upper bound, we can get the following result

$$\mathcal{R}_n(\mathcal{F}) \le C\sqrt{\frac{h}{n}}\,.$$

We can thus, with this approach, remove the unnecessary $\log n$ factor of the VC bound.

6 Advanced Topics

In this section, we point out several ways in which the results presented so far can be improved. The main source of improvement actually comes, as mentioned earlier, from the fact that Hoeffding and McDiarmid inequalities do not make use of the variance of the functions.

6.1 Binomial Tails

We recall that the functions we consider are binary valued. So, if we consider a fixed function f, the distribution of $P_n f$ is actually a binomial law of parameters Pf and n (since we are summing n i.i.d. random variables $f(Z_i)$ which can either be 0 or 1 and are equal to 1 with probability $\mathbb{E}f(Z_i) = Pf$). Denoting $p = Pf$, we can have an exact expression for the deviations of $P_n f$ from Pf:

$$\mathbb{P}\left[Pf - P_n f \geq t\right] = \sum_{k=0}^{\lfloor n(p-t) \rfloor} \binom{n}{k} p^k (1-p)^{n-k} .$$

Since this expression is not easy to manipulate, we have used an upper bound provided by Hoeffding's inequality. However, there exist other (sharper) upper bounds. The following quantities are an upper bound on $\mathbb{P}\left[Pf - P_n f \geq t\right]$,

$$\begin{array}{rl}
\left(\frac{1-p}{1-p-t}\right)^{n(1-p-t)} \left(\frac{p}{p+t}\right)^{n(p+t)} & \text{(exponential)} \\
e^{-\frac{np}{1-p}((1-t/p)\log(1-t/p)+t/p)} & \text{(Bennett)} \\
e^{-\frac{nt^2}{2p(1-p)+2t/3}} & \text{(Bernstein)} \\
e^{-2nt^2} & \text{(Hoeffding)}
\end{array}$$

Examining the above bounds (and using inversion), we can say that roughly speaking, the small deviations of $Pf - P_n f$ have a Gaussian behavior of the form $\exp(-nt^2/2p(1-p))$ (i.e. Gaussian with variance $p(1-p)$) while the large deviations have a Poisson behavior of the form $\exp(-3nt/2)$.
So the tails are heavier than Gaussian, and Hoeffding's inequality consists in upper bounding the tails with a Gaussian with maximum variance, hence the term $\exp(-2nt^2)$.

Each function $f \in \mathcal{F}$ has a different variance $Pf(1-Pf) \leq Pf$. Moreover, for each $f \in \mathcal{F}$, by Bernstein's inequality, with probability at least $1-\delta$,

$$Pf \leq P_n f + \sqrt{\frac{2Pf \log\frac{1}{\delta}}{n}} + \frac{2\log\frac{1}{\delta}}{3n} .$$

The Gaussian part (second term in the right hand side) dominates (for Pf not too small, or n large enough), and it depends on Pf. We thus want to combine Bernstein's inequality with the union bound and the symmetrization.

6.2 Normalization

The idea is to consider the ratio

$$\frac{Pf - P_n f}{\sqrt{Pf}} .$$

Here ($f \in \{0,1\}$), $\mathsf{Var} f \leq Pf^2 = Pf$

The reason for considering this ration is that after normalization, fluctuations are more 'uniform' in the class $\mathcal{F}$. Hence the supremum in

$$\sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}}$$

not necessarily attained at functions with large variance as it was the case previously.

Moreover, we know that our goal is to find functions with small error Pf (hence small variance). The normalized supremum takes this into account.

We now state a result similar to Theorem 2 for the normalized supremum.

Theorem 7 (Vapnik-Chervonenkis, [18]). *For $\delta > 0$ with probability at least $1-\delta$,*

$$\forall f \in \mathcal{F},\ \frac{Pf - P_n f}{\sqrt{Pf}} \leq 2\sqrt{\frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}}\,,$$

and also with probability at least $1-\delta$,

$$\forall f \in \mathcal{F},\ \frac{P_n f - Pf}{\sqrt{P_n f}} \leq 2\sqrt{\frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}}\,.$$

Proof. We only give a sketch of the proof. The first step is a variation of the symmetrization lemma

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}} \geq t\right] \leq 2\mathbb{P}\left[\sup_{f\in\mathcal{F}} \frac{P'_n f - P_n f}{\sqrt{(P_n f + P'_n f)/2}} \geq t\right].$$

The second step consists in randomization (with Rademacher variables)

$$\cdots = 2\mathbb{E}\left[\mathbb{P}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i(f(Z'_i) - f(Z_i))}{\sqrt{(P_n f + P'_n f)/2}} \geq t\right]\right].$$

Finally, one uses a tail bound of Bernstein type. □

Let us explore the consequences of this result.
From the fact that for non-negative numbers A, B, C,

$$A \leq B + C\sqrt{A} \Rightarrow A \leq B + C^2 + \sqrt{B}C\,,$$

we easily get for example

$$\forall f \in \mathcal{F},\ Pf \leq P_n f + 2\sqrt{P_n f \frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}} + 4\frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}\,.$$

In the ideal situation where there is no noise (i.e. $Y = t(X)$ almost surely), and $t \in \mathcal{G}$, denoting by g_n the empirical risk minimizer, we have $R^* = 0$ and also $R_n(g_n) = 0$. In particular, when $\mathcal{G}$ is a class of VC dimension h, we obtain

$$R(g_n) = O\left(\frac{h \log n}{n}\right) .$$

So, in a way, Theorem 7 allows to interpolate between the best case where the rate of convergence is $O(h \log n/n)$ and the worst case where the rate is $O(\sqrt{h \log n/n})$ (it does not allow to remove the $\log n$ factor in this case).

It is also possible to derive from Theorem 7 relative error bounds for the minimizer of the empirical error. With probability at least $1 - \delta$,

$$\begin{aligned} R(g_n) \le R(g^*) + 2\sqrt{R(g^*)\frac{\log S_{\mathcal{G}}(2n) + \log \frac{4}{\delta}}{n}} \\ +4\frac{\log S_{\mathcal{G}}(2n) + \log \frac{4}{\delta}}{n} . \end{aligned}$$

We notice here that when $R(g^*) = 0$ (i.e. $t \in \mathcal{G}$ and $R^* = 0$), the rate is again of order $1/n$ while, as soon as $R(g^*) > 0$, the rate is of order $1/\sqrt{n}$. Therefore, it is not possible to obtain a rate with a power of n in between $-1/2$ and -1.

The main reason is that the factor of the square root term $R(g^*)$ is not the right quantity to use here since it does not vary with n. We will see later that one can have instead $R(g_n) - R(g^*)$ as a factor, which is usually converging to zero with n increasing. Unfortunately, Theorem 7 cannot be applied to functions of the type $f - f^*$ (which would be needed to have the mentioned factor), so we will need a refined approach.

6.3 Noise Conditions

The refinement we seek to obtain requires certain specific assumptions about the noise function $s(x)$. The ideal case being when $s(x) = 0$ everywhere (which corresponds to $R^* = 0$ and $Y = t(X)$). We now introduce quantities that measure how well-behaved the noise function is.

The situation is favorable when the regression function $\eta(x)$ is not too close to 0, or at least not too often close to 1/2. Indeed, $\eta(x) = 0$ means that the noise is maximum at x ($s(x) = 1/2$) and that the label is completely undetermined (any prediction would yield an error with probability 1/2).

Definitions. There are two types of conditions.

Definition 6 (Massart's Noise Condition). *For some $c > 0$, assume*

$$|\eta(X)| > \frac{1}{c} \textit{ almost surely} .$$

This condition implies that there is no region where the decision is completely random, or the noise is bounded away from 1/2.

Definition 7 (Tsybakov's Noise Condition). *Let $\alpha \in [0,1]$, assume that one the following equivalent conditions is satisfied*

$$\begin{aligned}
&(i)\ \exists c > 0,\ \forall g \in \{-1,1\}^{\mathcal{X}}, \\
&\qquad \mathbb{P}\left[g(X)\eta(X) \le 0\right] \le c(R(g) - R^*)^\alpha \\
&(ii)\ \exists c > 0,\ \forall A \subset \mathcal{X},\ \int_A dP(x) \le c\Big(\int_A |\eta(x)| dP(x)\Big)^\alpha \\
&(iii)\ \exists B > 0,\ \forall t \ge 0,\ \mathbb{P}\left[|\eta(X)| \le t\right] \le B t^{\frac{\alpha}{1-\alpha}}
\end{aligned}$$

Condition (iii) is probably the easiest to interpret: it means that $\eta(x)$ is close to the critical value 0 with low probability.

We indicate how to prove that conditions $(i), (ii)$ and (iii) are indeed equivalent:

$(i) \Leftrightarrow (ii)$ It is easy to check that $R(g) - R^* = \mathbb{E}\left[|\eta(X)| \mathbb{1}_{g\eta \le 0}\right]$. For each function g, there exists a set A such that $\mathbb{1}_A = \mathbb{1}_{g\eta \le 0}$

$(ii) \Rightarrow (iii)$ Let $A = \{x : |\eta(x)| \le t\}$

$$\begin{aligned}
\mathbb{P}\left[|\eta| \le t\right] = \int_A dP(x) &\le c\Big(\int_A |\eta(x)| dP(x)\Big)^\alpha \\
&\le c t^\alpha \Big(\int_A dP(x)\Big)^\alpha \\
\Rightarrow \mathbb{P}\left[|\eta| \le t\right] &\le c^{\frac{1}{1-\alpha}} t^{\frac{\alpha}{1-\alpha}}
\end{aligned}$$

$(iii) \Rightarrow (i)$ We write

$$\begin{aligned}
&R(g) - R^* = \mathbb{E}\left[|\eta(X)|\, g\eta \le 0\right] \\
&\ge t\mathbb{E}\left[\mathbb{1}_{g\eta \le 0} \mathbb{1}_{|\eta| t}\right] \\
&= t\mathbb{P}\left[|\eta|\, t\right] - t\mathbb{E}\left[\mathbb{1}_{g\eta > 0} \mathbb{1}_{|\eta| t}\right] \\
&\ge t(1 - B t^{\frac{\alpha}{1-\alpha}}) - t\mathbb{P}\left[g\eta > 0\right] = t(\mathbb{P}\left[g\eta \le 0\right] - B t^{\frac{\alpha}{1-\alpha}}) .
\end{aligned}$$

Taking $t = \left(\frac{(1-\alpha)\mathbb{P}[g\eta \le 0]}{B}\right)^{(1-\alpha)/\alpha}$ finally gives

$$\mathbb{P}\left[g\eta \le 0\right] \le \frac{B^{1-\alpha}}{(1-\alpha)^(1-\alpha)\alpha^\alpha} (R(g) - R^*)^\alpha .$$

We notice that the parameter α has to be in $[0,1]$. Indeed, one has the opposite inequality

$$R(g) - R^* = \mathbb{E}\left[|\eta(X)| \mathbb{1}_{g\eta \le 0}\right] \le \mathbb{E}\left[\mathbb{1}_{g\eta \le 0}\right] = \mathbb{P}\left[g(X)\eta(X) \le 0\right] ,$$

which is incompatible with condition (i) if $\alpha > 1$.

We also notice that when $\alpha = 0$, Tsybakov's condition is void, and when $\alpha = 1$, it is equivalent to Massart's condition.

Consequences. The conditions we impose on the noise yield a crucial relationship between the variance and the expectation of functions in the so-called relative loss class defined as

$$\tilde{\mathcal{F}} = \{(x,y) \mapsto f(x,y) - \mathbb{1}_{t(x)\neq y} : f \in \mathcal{F}\} \,.$$

This relationship will allow to exploit Bernstein type inequalities applied to this latter class.

Under Massart's condition, one has (written in terms of the initial class) for $g \in \mathcal{G}$,

$$\mathbb{E}\left[(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{t(X)\neq Y})^2\right] \leq c(R(g) - R^*) \,,$$

or, equivalently, for $f \in \tilde{\mathcal{F}}$, $\mathsf{Var} f \leq Pf^2 \leq cPf$. Under Tsybakov's condition this becomes for $g \in \mathcal{G}$,

$$\mathbb{E}\left[(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{t(X)\neq Y})^2\right] \leq c(R(g) - R^*)^\alpha \,,$$

and for $f \in \tilde{\mathcal{F}}$, $\mathsf{Var} f \leq Pf^2 \leq c(Pf)^\alpha$.

In the finite case, with $|\mathcal{G}| = N$, one can easily apply Bernstein's inequality to $\tilde{\mathcal{F}}$ and the finite union bound to get that with probability at least $1-\delta$, for all $g \in \mathcal{G}$,

$$R(g) - R^* \leq R_n(g) - R_n(t) + \sqrt{\frac{8c(R(g)-R^*)^\alpha \log \frac{N}{\delta}}{n}} + \frac{4\log\frac{N}{\delta}}{3n} \,.$$

As a consequence, when $t \in \mathcal{G}$, and g_n is the minimizer of the empirical error (hence $R_n(g) \leq R_n(t)$), one has

$$R(g_n) - R^* \leq C \left(\frac{\log \frac{N}{\delta}}{n}\right)^{\frac{1}{2-\alpha}} \,,$$

which always better than $n^{-1/2}$ for $\alpha > 0$ and is valid even if $R^* > 0$.

6.4 Local Rademacher Averages

In this section we generalize the above result by introducing a localized version of the Rademacher averages. Going from the finite to the general case is more involved than what has been seen before. We first give the appropriate definitions, then state the result and give a proof sketch.

Definitions. Local Rademacher averages refer to Rademacher averages of subsets of the function class determined by a condition on the variance of the function.

Definition 8 (Local Rademacher Average). *The local Rademacher average at radius $r \geq 0$ for the class $\mathcal{F}$ is defined as*

$$\mathcal{R}(\mathcal{F}, r) = \mathbb{E} \sup_{f \in \mathcal{F}: Pf^2 \leq r} R_n f \,.$$

The reason for this definition is that, as we have seen before, the crucial ingredient to obtain better rates of convergence is to use the variance of the functions. Localizing the Rademacher average allows to focus on the part of the function class where the fast rate phenomenon occurs, that are functions with small variance.

Next we introduce the concept of a sub-root function, a real-valued function with certain monotony properties.

Definition 9 (Sub-Root Function). *A function $\psi : \mathbb{R} \to \mathbb{R}$ is sub-root if*

(i) *ψ is non-decreasing,*
(ii) *ψ is non negative,*
(iii) *$\psi(r)/\sqrt{r}$ is non-increasing.*

An immediate consequence of this definition is the following result.

Lemma 5. *A sub-root function*

(i) *is continuous,*
(ii) *has a unique (non-zero)* fixed point *r^* satisfying $\psi(r^*) = r^*$.*

Figure 6 shows a typical sub-root function and its fixed point.

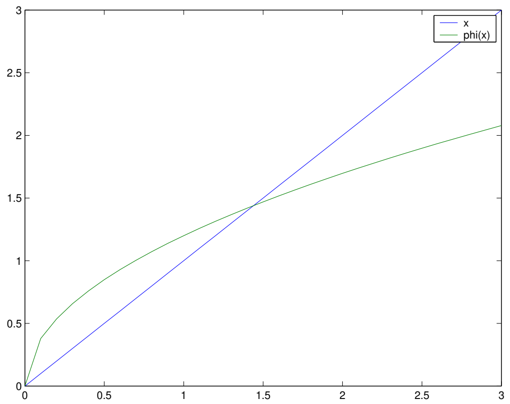

Fig. 6. An example of a sub-root function and its fixed point.

Before seeing the rationale for introducing the sub-root concept, we need yet another definition, that of a 'star-hull' (somewhat similar to a convex hull).

Definition 10 (Star-Hull). *Let $\mathcal{F}$ be a set of functions. Its star-hull is defined as*

$$\star\mathcal{F} = \{\alpha f : f \in \mathcal{F},\, \alpha \in [0, 1]\}\,.$$

Now, we state a lemma that indicates that by taking the star-hull of a class of functions, we are guaranteed that the local Rademacher average behaves like a sub-root function, and thus has a unique fixed point. This fixed point will turn out to be the key quantity in the relative error bounds.

Lemma 6. *For any class of functions $\mathcal{F}$,*

$$\mathcal{R}_n(\star\mathcal{F}, r) \text{ is sub-root}.$$

One legitimate question is whether taking the star-hull does not enlarge the class too much. One way to see what the effect is on the size of the class is to compare the metric entropy (log covering numbers) of $\mathcal{F}$ and of $\star\mathcal{F}$. It is possible to see that the entropy increases only by a logarithmic factor, which is essentially negligible.

Result. We now state the main result involving local Rademacher averages and their fixed point.

Theorem 8. *Let $\mathcal{F}$ be a class of bounded functions (e.g. $f \in [-1,1]$) and r^* be the fixed point of $\mathcal{R}(\star\mathcal{F}, r)$. There exists a constant $C > 0$ such that with probability at least $1-\delta$,*

$$\forall f \in \mathcal{F},\ Pf - P_n f \leq C\left(\sqrt{r^* \mathsf{Var} f} + \frac{\log\frac{1}{\delta} + \log\log n}{n}\right).$$

If in addition the functions in $\mathcal{F}$ satisfy $\mathsf{Var} f \leq c(Pf)^\beta$, then one obtains that with probability at least $1-\delta$,

$$\forall f \in \mathcal{F},\ Pf \leq C\left(P_n f + (r^*)^{\frac{1}{2-\beta}} + \frac{\log\frac{1}{\delta} + \log\log n}{n}\right).$$

Proof. We only give the main steps of the proof.

1. The starting point is Talagrand's inequality for empirical processes, a generalization of McDiarmid's inequality of Bernstein type (i.e. which includes the variance). This inequality tells that with high probability,

$$\sup_{f\in\mathcal{F}} Pf - P_n f \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} Pf - P_n f\right] + c\sqrt{\sup_{f\in\mathcal{F}} \mathsf{Var} f/n} + c'/n,$$

   for some constants c, c'.
2. The second step consists in 'peeling' the class, that is splitting the class into subclasses according to the variance of the functions

$$\mathcal{F}_k = \{f : \mathsf{Var} f \in [x^k, x^{k+1})\},$$

3. We can then apply Talagrand's inequality to each of the sub-classes separately to get with high probability
$$\sup_{f\in\mathcal{F}_k} Pf - P_nf \le \mathbb{E}\left[\sup_{f\in\mathcal{F}_k} Pf - P_nf\right] + c\sqrt{x\mathsf{Var}f/n} + c'/n\,,$$
4. Then the symmetrization lemma allows to introduce local Rademacher averages. We get that with high probability
$$\forall f\in\mathcal{F},\ Pf - P_nf \le 2\mathcal{R}(\mathcal{F}, x\mathsf{Var}f) + c\sqrt{x\mathsf{Var}f/n} + c'/n\,.$$
5. We then have to 'solve' this inequality. Things are simple if $\mathcal{R}$ behaves like a square root function since we can upper bound the local Rademacher average by the value of its fixed point. With high probability,
$$Pf - P_nf \le 2\sqrt{r^*\mathsf{Var}f} + c\sqrt{x\mathsf{Var}f/n} + c'/n\,.$$
6. Finally, we use the relationship between variance and expectation
$$\mathsf{Var}f \le c(Pf)^\alpha\,,$$
and solve the inequality in Pf to get the result.

□

We will not got into the details of how to apply the above result, but we give some remarks about its use.

An important example is the case where the class $\mathcal{F}$ is of finite VC dimension h. In that case, one has
$$\mathcal{R}(\mathcal{F}, r) \le C\sqrt{\frac{rh\log n}{n}}\,,$$
so that $r^* \le C\frac{h\log n}{n}$. As a consequence, we obtain, under Tsybakov condition, a rate of convergence of Pf_n to Pf^* is $O(1/n^{1/(2-\alpha)})$. It is important to note that in this case, the rate of convergence of P_nf to Pf in $O(1/\sqrt{n})$. So we obtain a fast rate by looking at the relative error. These fast rates can be obtained provided $t\in\mathcal{G}$ (but it is not needed that $R^*=0$). This requirement can be removed if one uses structural risk minimization or regularization.

Another related result is that, as in the global case, one can obtain a bound with data-dependent (i.e. conditional) local Rademacher averages
$$\mathcal{R}_n(\mathcal{F}, r) = \mathbb{E}_\sigma \sup_{f\in\mathcal{F}:Pf^2\le r} R_nf\,.$$
The result is the same as before (with different constants) under the same conditions as in Theorem 8. With probability at least $1-\delta$,
$$Pf \le C\left(P_nf + (r_n^*)^{\frac{1}{2-\alpha}} + \frac{\log\frac{1}{\delta} + \log\log n}{n}\right)$$

where r_n^* is the fixed point of a sub-root upper bound of $\mathcal{R}_n(\mathcal{F}, r)$.

Hence, we can get improved rates when the noise is well-behaved and these rates interpolate between $n^{-1/2}$ and n^{-1}. However, it is not in general possible to estimate the parameters (c and α) entering in the noise conditions, but we will not discuss this issue further here. Another point is that although the capacity measure that we use seems 'local', it does depend on all the functions in the class, but each of them is implicitly appropriately rescaled. Indeed, in $\mathcal{R}(\star\mathcal{F}, r)$, each function $f \in \mathcal{F}$ with $Pf^2 \geq r$ is considered at scale r/Pf^2.

Bibliographical remarks. Hoeffding's inequality appears in [19]. For a proof of the contraction principle we refer to Ledoux and Talagrand [20].

Vapnik-Chervonenkis-Sauer-Shelah's lemma was proved independently by Sauer [21], Shelah [22], and Vapnik and Chervonenkis [18]. For related combinatorial results we refer to Alesker [23], Alon, Ben-David, Cesa-Bianchi, and Haussler [24], Cesa-Bianchi and Haussler [25], Frankl [26], Haussler [27], Szarek and Talagrand [28].

Uniform deviations of averages from their expectations is one of the central problems of empirical process theory. Here we merely refer to some of the comprehensive coverages, such as Dudley [29], Giné [30], Vapnik [1], van der Vaart and Wellner [31]. The use of empirical processes in classification was pioneered by Vapnik and Chervonenkis [18, 15] and re-discovered 20 years later by Blumer, Ehrenfeucht, Haussler, and Warmuth [32], Ehrenfeucht, Haussler, Kearns, and Valiant [33]. For surveys see Anthony and Bartlett [2], Devroye, Györfi, and Lugosi [4], Kearns and Vazirani [7], Natarajan [12], Vapnik [14, 1].

The question of how $\sup_{f\in\mathcal{F}}(P(f) - P_n(f))$ behaves has been known as the Glivenko-Cantelli problem and much has been said about it. A few key references include Alon, Ben-David, Cesa-Bianchi, and Haussler [24], Dudley [34, 35, 36], Talagrand [37, 38], Vapnik and Chervonenkis [18, 39].

The VC dimension has been widely studied and many of its properties are known. We refer to Anthony and Bartlett [2], Assouad [40], Cover [41], Dudley [42, 29], Goldberg and Jerrum [43], Karpinski and A. Macintyre [44], Khovanskii [45], Koiran and Sontag [46], Macintyre and Sontag [47], Steele [48], and Wenocur and Dudley [49].

The bounded differences inequality was formulated explicitly first by McDiarmid [17] who proved it by martingale methods (see the surveys [17], [50]), but closely related concentration results have been obtained in various ways including information-theoretic methods (see Alhswede, Gács, and Körner [51], Marton [52], [53],[54], Dembo [55], Massart [56] and Rio [57]), Talagrand's induction method [58],[59],[60] (see also Luczak and McDiarmid [61], McDiarmid [62], Panchenko [63, 64, 65]) and the so-called "entropy method", based on logarithmic Sobolev inequalities, developed by Ledoux [66],[67], see also Bobkov and Ledoux [68], Massart [69], Rio [57], Boucheron, Lugosi, and Massart [70], [71], Boucheron, Bousquet, Lugosi, and Massart [72], and Bousquet [73].

Symmetrization lemmas can be found in Giné and Zinn [74] and Vapnik and Chervonenkis [18, 15].

The use of Rademacher averages in classification was first promoted by Koltchinskii [75] and Bartlett, Boucheron, and Lugosi [76], see also Koltchinskii and Panchenko [77, 78], Bartlett and Mendelson [79], Bartlett, Bousquet, and Mendelson [80], Bousquet, Koltchinskii, and Panchenko [81], Kégl, Linder, and Lugosi [82].

A Probability Tools

This section recalls some basic facts from probability theory that are used throughout this tutorial (sometimes without explicitly mentioning it).

We denote by A and B some events (i.e. elements of a σ-algebra), and by X some real-valued random variable.

A.1 Basic Facts

- Union:
$$\mathbb{P}\left[A \text{ or } B\right] \leq \mathbb{P}\left[A\right] + \mathbb{P}\left[B\right] .$$
- Inclusion: If $A \Rightarrow B$, then $\mathbb{P}\left[A\right] \leq \mathbb{P}\left[B\right]$.
- Inversion: If $\mathbb{P}\left[X > t\right] \leq F(t)$ then with probability at least $1-\delta$,
$$X \leq F^{-1}(\delta) .$$
- Expectation: If $X \geq 0$,
$$\mathbb{E}\left[X\right] = \int_0^\infty \mathbb{P}\left[X \geq t\right] dt .$$

A.2 Basic Inequalities

All the inequalities below are valid as soon as the right-hand side exists.

- Jensen: for f convex,
$$f(\mathbb{E}\left[X\right]) \leq \mathbb{E}\left[f(X)\right] .$$
- Markov: If $X \geq 0$ then for all $t > 0$,
$$\mathbb{P}\left[X \geq t\right] \leq \frac{\mathbb{E}\left[X\right]}{t} .$$
- Chebyshev: for $t > 0$,
$$\mathbb{P}\left[\left|X - \mathbb{E}\left[X\right]\right| \geq t\right] \leq \frac{\mathsf{Var} X}{t^2} .$$
- Chernoff: for all $t \in \mathbb{R}$,
$$\mathbb{P}\left[X \geq t\right] \leq \inf_{\lambda \geq 0} \mathbb{E}\left[e^{\lambda(X-t)}\right] .$$

B No Free Lunch

We can now give a formal definition of consistency and state the core results about the impossibility of universally good algorithms.

Definition 11 (Consistency). *An algorithm is consistent if for any probability measure P,*

$$\lim_{n\to\infty} R(g_n) = R^* \text{ almost surely.}$$

It is important to understand the reasons that make possible the existence of consistent algorithms. In the case where the input space $\mathcal{X}$ is countable, things are somehow easy since even if there is no relationship at all between inputs and outputs, by repeatedly sampling data independently from P, one will get to see an increasing number of different inputs which will eventually converge to all the inputs. So, in the countable case, an algorithm which would simply learn 'by heart' (i.e. makes a majority vote when the instance has been seen before, and produces an arbitrary prediction otherwise) would be consistent.

In the case where $\mathcal{X}$ is not countable (e.g. $\mathcal{X} = \mathbb{R}$), things are more subtle. Indeed, in that case, there is a seemingly innocent assumption that becomes crucial: to be able to define a probability measure P on $\mathcal{X}$, one needs a σ-algebra on that space, which is typically the Borel σ-algebra. So the hidden assumption is that P is a Borel measure. This means that the topology of $\mathbb{R}$ plays a role here, and thus, the target function t will be Borel measurable. In a sense this guarantees that it is possible to approximate t from its value (or approximate value) at a finite number of points. The algorithms that will achieve consistency are thus those who use the topology in the sense of 'generalizing' the observed values to neighborhoods (e.g. local classifiers). In a way, the measurability of t is one of the crudest notions of smoothness of functions.

We now cite two important results. The first one tells that for a fixed sample size, one can construct arbitrarily bad problems for a given algorithm.

Theorem 9 (No Free Lunch, see e.g. [4]). *For any algorithm, any n and any $\varepsilon > 0$, there exists a distribution P such that $R^* = 0$ and*

$$\mathbb{P}\left[R(g_n) \geq \frac{1}{2} - \varepsilon\right] = 1\,.$$

The second result is more subtle and indicates that given an algorithm, one can construct a problem for which this algorithm will converge as slowly as one wishes.

Theorem 10 (No Free Lunch at All, see e.g. [4]). *For any algorithm, and any sequence (a_n) that converges to 0, there exists a probability distribution P such that $R^* = 0$ and*

$$R(g_n) \geq a_n\,.$$

In the above theorem, the 'bad' probability measure is constructed on a countable set (where the outputs are not related at all to the inputs so that no generalization is possible), and is such that the rate at which one gets to see new inputs is as slow as the convergence of a_n.

Finally we mention other notions of consistency.

Definition 12 (VC consistency of ERM). *The ERM algorithm is consistent if for any probability measure P,*

$$R(g_n) \to R(g^*) \text{ in probability,}$$

and

$$R_n(g_n) \to R(g^*) \text{ in probability.}$$

Definition 13 (VC non-trivial consistency of ERM). *The ERM algorithm is non-trivially consistent for the set $\mathcal{G}$ and the probability distribution P if for any $c \in \mathbb{R}$,*

$$\inf_{f \in \mathcal{F}: Pf > c} P_n(f) \to \inf_{f \in \mathcal{F}: Pf > c} P(f) \text{ in probability.}$$

References

1. Vapnik, V.: Statistical Learning Theory. John Wiley, New York (1998)
2. Anthony, M., Bartlett, P.L.: Neural Network Learning: Theoretical Foundations. Cambridge University Press, Cambridge (1999)
3. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth International, Belmont, CA (1984)
4. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York (1996)
5. Duda, R., Hart, P.: Pattern Classification and Scene Analysis. John Wiley, New York (1973)
6. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, New York (1972)
7. Kearns, M., Vazirani, U.: An Introduction to Computational Learning Theory. MIT Press, Cambridge, Massachusetts (1994)
8. Kulkarni, S., Lugosi, G., Venkatesh, S.: Learning pattern classification—a survey. IEEE Transactions on Information Theory **44** (1998) 2178–2206 Information Theory: 1948–1998. Commemorative special issue.
9. Lugosi, G.: Pattern classification and learning theory. In Györfi, L., ed.: Principles of Nonparametric Learning, Springer, Viena (2002) 5–62
10. McLachlan, G.: Discriminant Analysis and Statistical Pattern Recognition. John Wiley, New York (1992)
11. Mendelson, S.: A few notes on statistical learning theory. In Mendelson, S., Smola, A., eds.: Advanced Lectures in Machine Learning. LNCS 2600, Springer (2003) 1–40
12. Natarajan, B.: Machine Learning: A Theoretical Approach. Morgan Kaufmann, San Mateo, CA (1991)
13. Vapnik, V.: Estimation of Dependencies Based on Empirical Data. Springer-Verlag, New York (1982)
14. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag, New York (1995)
15. Vapnik, V., Chervonenkis, A.: Theory of Pattern Recognition. Nauka, Moscow (1974) (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.

16. von Luxburg, U., Bousquet, O., Schölkopf, B.: A compression approach to support vector model selection. The Journal of Machine Learning Research **5** (2004) 293–323
17. McDiarmid, C.: On the method of bounded differences. In: Surveys in Combinatorics 1989, Cambridge University Press, Cambridge (1989) 148–188
18. Vapnik, V., Chervonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications **16** (1971) 264–280
19. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association **58** (1963) 13–30
20. Ledoux, M., Talagrand, M.: Probability in Banach Space. Springer-Verlag, New York (1991)
21. Sauer, N.: On the density of families of sets. Journal of Combinatorial Theory Series A **13** (1972) 145–147
22. Shelah, S.: A combinatorial problem: Stability and order for models and theories in infinity languages. Pacific Journal of Mathematics **41** (1972) 247–261
23. Alesker, S.: A remark on the Szarek-Talagrand theorem. Combinatorics, Probability, and Computing **6** (1997) 139–144
24. Alon, N., Ben-David, S., Cesa-Bianchi, N., Haussler, D.: Scale-sensitive dimensions, uniform convergence, and learnability. Journal of the ACM **44** (1997) 615–631
25. Cesa-Bianchi, N., Haussler, D.: A graph-theoretic generalization of the Sauer-Shelah lemma. Discrete Applied Mathematics **86** (1998) 27–35
26. Frankl, P.: On the trace of finite sets. Journal of Combinatorial Theory, Series A **34** (1983) 41–45
27. Haussler, D.: Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. Journal of Combinatorial Theory, Series A **69** (1995) 217–232
28. Szarek, S., Talagrand, M.: On the convexified Sauer-Shelah theorem. Journal of Combinatorial Theory, Series B **69** (1997) 183–192
29. Dudley, R.: Uniform Central Limit Theorems. Cambridge University Press, Cambridge (1999)
30. Giné, E.: Empirical processes and applications: an overview. Bernoulli **2** (1996) 1–28
31. van der Waart, A., Wellner, J.: Weak convergence and empirical processes. Springer-Verlag, New York (1996)
32. Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.: Learnability and the Vapnik-Chervonenkis dimension. Journal of the ACM **36** (1989) 929–965
33. Ehrenfeucht, A., Haussler, D., Kearns, M., Valiant, L.: A general lower bound on the number of examples needed for learning. Information and Computation **82** (1989) 247–261
34. Dudley, R.: Central limit theorems for empirical measures. Annals of Probability **6** (1978) 899–929
35. Dudley, R.: Empirical processes. In: Ecole de Probabilité de St. Flour 1982, Lecture Notes in Mathematics #1097, Springer-Verlag, New York (1984)
36. Dudley, R.: Universal Donsker classes and metric entropy. Annals of Probability **15** (1987) 1306–1326
37. Talagrand, M.: The Glivenko-Cantelli problem. Annals of Probability **15** (1987) 837–870
38. Talagrand, M.: Sharper bounds for Gaussian and empirical processes. Annals of Probability **22** (1994) 28–76

39. Vapnik, V., Chervonenkis, A.: Necessary and sufficient conditions for the uniform convergence of means to their expectations. Theory of Probability and its Applications **26** (1981) 821–832
40. Assouad, P.: Densité et dimension. Annales de l'Institut Fourier **33** (1983) 233–282
41. Cover, T.: Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Transactions on Electronic Computers **14** (1965) 326–334
42. Dudley, R.: Balls in R^k do not cut all subsets of $k + 2$ points. Advances in Mathematics **31 (3)** (1979) 306–308
43. Goldberg, P., Jerrum, M.: Bounding the Vapnik-Chervonenkis dimension of concept classes parametrized by real numbers. Machine Learning **18** (1995) 131–148
44. Karpinski, M., Macintyre, A.: Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks. Journal of Computer and System Science **54** (1997)
45. Khovanskii, A.G.: Fewnomials. Translations of Mathematical Monographs, vol. 88, American Mathematical Society (1991)
46. Koiran, P., Sontag, E.: Neural networks with quadratic VC dimension. Journal of Computer and System Science **54** (1997)
47. Macintyre, A., Sontag, E.: Finiteness results for sigmoidal "neural" networks. In: Proceedings of the 25th Annual ACM Symposium on the Theory of Computing, Association of Computing Machinery, New York (1993) 325–334
48. Steele, J.: Existence of submatrices with all possible columns. Journal of Combinatorial Theory, Series A **28** (1978) 84–88
49. Wenocur, R., Dudley, R.: Some special Vapnik-Chervonenkis classes. Discrete Mathematics **33** (1981) 313–318
50. McDiarmid, C.: Concentration. In Habib, M., McDiarmid, C., Ramirez-Alfonsin, J., Reed, B., eds.: Probabilistic Methods for Algorithmic Discrete Mathematics, Springer, New York (1998) 195–248
51. Ahlswede, R., Gács, P., Körner, J.: Bounds on conditional probabilities with applications in multi-user communication. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **34** (1976) 157–177 (correction in 39:353–354,1977).
52. Marton, K.: A simple proof of the blowing-up lemma. IEEE Transactions on Information Theory **32** (1986) 445–446
53. Marton, K.: Bounding $\bar{d}$-distance by informational divergence: a way to prove measure concentration. Annals of Probability **24** (1996) 857–866
54. Marton, K.: A measure concentration inequality for contracting Markov chains. Geometric and Functional Analysis **6** (1996) 556–571 Erratum: 7:609–613, 1997.
55. Dembo, A.: Information inequalities and concentration of measure. Annals of Probability **25** (1997) 927–939
56. Massart, P.: Optimal constants for Hoeffding type inequalities. Technical report, Mathematiques, Université de Paris-Sud, Report 98.86 (1998)
57. Rio, E.: Inégalités de concentration pour les processus empiriques de classes de parties. Probability Theory and Related Fields **119** (2001) 163–175
58. Talagrand, M.: A new look at independence. Annals of Probability **24** (1996) 1–34 (Special Invited Paper).
59. Talagrand, M.: Concentration of measure and isoperimetric inequalities in product spaces. Publications Mathématiques de l'I.H.E.S. **81** (1995) 73–205
60. Talagrand, M.: New concentration inequalities in product spaces. Inventiones Mathematicae **126** (1996) 505–563
61. Luczak, M.J., McDiarmid, C.: Concentration for locally acting permutations. Discrete Mathematics (2003) to appear

62. McDiarmid, C.: Concentration for independent permutations. Combinatorics, Probability, and Computing **2** (2002) 163–178
63. Panchenko, D.: A note on Talagrand's concentration inequality. Electronic Communications in Probability **6** (2001)
64. Panchenko, D.: Some extensions of an inequality of Vapnik and Chervonenkis. Electronic Communications in Probability **7** (2002)
65. Panchenko, D.: Symmetrization approach to concentration inequalities for empirical processes. Annals of Probability **to appear** (2003)
66. Ledoux, M.: On Talagrand's deviation inequalities for product measures. ESAIM: Probability and Statistics **1** (1997) 63–87 `http://www.emath.fr/ps/`.
67. Ledoux, M.: Isoperimetry and Gaussian analysis. In Bernard, P., ed.: Lectures on Probability Theory and Statistics, Ecole d'Eté de Probabilités de St-Flour XXIV-1994 (1996) 165–294
68. Bobkov, S., Ledoux, M.: Poincaré's inequalities and Talagrands's concentration phenomenon for the exponential distribution. Probability Theory and Related Fields **107** (1997) 383–400
69. Massart, P.: About the constants in Talagrand's concentration inequalities for empirical processes. Annals of Probability **28** (2000) 863–884
70. Boucheron, S., Lugosi, G., Massart, P.: A sharp concentration inequality with applications. Random Structures and Algorithms **16** (2000) 277–292
71. Boucheron, S., Lugosi, G., Massart, P.: Concentration inequalities using the entropy method. The Annals of Probability **31** (2003) 1583–1614
72. Boucheron, S., Bousquet, O., Lugosi, G., Massart, P.: Moment inequalities for functions of independent random variables. The Annals of Probability (2004) to appear.
73. Bousquet, O.: A Bennett concentration inequality and its application to suprema of empirical processes. C. R. Acad. Sci. Paris **334** (2002) 495–500
74. Giné, E., Zinn, J.: Some limit theorems for empirical processes. Annals of Probability **12** (1984) 929–989
75. Koltchinskii, V.: Rademacher penalties and structural risk minimization. IEEE Transactions on Information Theory **47** (2001) 1902–1914
76. Bartlett, P., Boucheron, S., Lugosi, G.: Model selection and error estimation. Machine Learning **48** (2001) 85–113
77. Koltchinskii, V., Panchenko, D.: Empirical margin distributions and bounding the generalization error of combined classifiers. Annals of Statistics **30** (2002)
78. Koltchinskii, V., Panchenko, D.: Rademacher processes and bounding the risk of function learning. In Giné, E., Mason, D., Wellner, J., eds.: High Dimensional Probability II. (2000) 443–459
79. Bartlett, P., Mendelson, S.: Rademacher and Gaussian complexities: risk bounds and structural results. Journal of Machine Learning Research **3** (2002) 463–482
80. Bartlett, P., Bousquet, O., Mendelson, S.: Localized Rademacher complexities. In: Proceedings of the 15th annual conference on Computational Learning Theory. (2002) 44–48
81. Bousquet, O., Koltchinskii, V., Panchenko, D.: Some local measures of complexity of convex hulls and generalization bounds. In: Proceedings of the 15th Annual Conference on Computational Learning Theory, Springer (2002) 59–73
82. Antos, A., Kégl, B., Linder, T., Lugosi, G.: Data-dependent margin-based generalization bounds for classification. Journal of Machine Learning Research **3** (2002) 73–98