Journal of Machine Learning Research 5 (2004) 219–238 Submitted 12/02; Published 3/04

On the Importance of Small Coordinate Projections

Shahar Mendelson SHAHAR.MENDELSON@ANU.EDU.AU
Petra Philips PETRA.PHILIPS@ANU.EDU.AU
RSISE, The Australian National University
Canberra, ACT 0200, Australia

Editor: Peter Bartlett

Abstract

It has been recently shown that sharp generalization bounds can be obtained when the function class from which the algorithm chooses its hypotheses is "small" in the sense that the Rademacher averages of this function class are small. We show that a new more general principle guarantees good generalization bounds. The new principle requires that random coordinate projections of the function class evaluated on random samples are "small" with high probability and that the random class of functions allows symmetrization. As an example, we prove that this geometric property of the function class is exactly the reason why the two lately proposed frameworks, the *luckiness* (Shawe-Taylor et al., 1998) and the *algorithmic luckiness* (Herbrich and Williamson, 2002), can be used to establish generalization bounds.

Keywords: Statistical learning theory, generalization bounds, data-dependent complexities, coordinate projections

1. Introduction

Generalization bounds are used to show that, with high probability, functions produced by a learning algorithm have a small error, and as such, can be used to approximate the unknown target. For many years, such bounds were obtained by deviation estimates of empirical means from the actual mean, *uniformly* over the whole class of functions from which the algorithm produces its hypothesis. Thus, classes which satisfy the uniform law of large numbers (so-called uniform Glivenko-Cantelli or GC classes) have played a central role in Machine Learning literature. More recently, other methods of deriving generalization bounds were developed, in which the "size" of the function class from which the algorithm chooses a hypothesis is not specified *a priori*. Examples of such methods are the *luckiness* (Shawe-Taylor et al., 1998) and the *algorithmic luckiness* (Herbrich and Williamson, 2002) frameworks, but although in both cases one can obtain generalization bounds, they seem to be based on completely different arguments.

In this article, we show that the bounds obtained in all these frameworks follow from the same general principle. This principle requires that coordinate projections of a function subclass on random samples are "small" with high probability.

We consider the following setting for the learning problem: let Ω be a measurable input space, $t : \Omega \longrightarrow \mathbb{R}$ an unknown real-valued target function, $H = \{h \,|\, h : \Omega \longrightarrow \mathbb{R}\}$ a class of hypothesis functions, and μ an unknown probability distribution on Ω. Let $(X_1, \ldots, X_n) \in \Omega^n$ be a finite training sample, where each X_i is generated randomly, independently, according to μ. Based on the values of the target function on this sample, $(t(X_1), \ldots, t(X_n))$, the goal of a learning algorithm is to choose a function $h^* \in H$ which is a good estimator of the target function t. A quantitative measure of how well a function $h \in H$ approximates t is given by a loss function $l : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$. Typical examples of loss functions are the 0-1 loss for classification defined by $l(r,s) = 0$ if $r = s$ and $l(r,s) = 1$ if $r \neq s$ or the square-loss for regression tasks $l(r,s) = (r-s)^2$. In what follows we will assume a bounded loss function and therefore, without loss of generality, $l : \mathbb{R} \times \mathbb{R} \longrightarrow [-1,1]$. For every $h \in H$ we define the associated loss function $l_h : \Omega \longrightarrow [-1,1]$, $l_h(x) := l(h(x), t(x))$ and denote by $F = \{l_h : \Omega \longrightarrow [-1,1] \,|\, h \in H\}$ the loss class associated with the learning problem. If h^* is the best estimate for t in H, we call $F' = \{l_h - l_{h^*} \,|\, h \in H\}$ the excess loss class.

The best estimate for t is defined to be the $h^* \in H$ for which the expected loss (also called risk) over all possible observations is as small as possible, that is, $\int l_{h*}(x) d\mu(x) = \mathbb{E}_\mu l_{h*} \approx \inf_{h \in H} \mathbb{E}_\mu l_h$. Empirical risk minimization algorithms are based on the philosophy that it is possible to approximate this expectation with the empirical mean, and choose instead a hypothesis $\hat{h} \in H$ for which $\frac{1}{n}\sum_{i=1}^n l_{\hat{h}}(x_i) \approx \inf_{h \in H} \frac{1}{n}\sum_{i=1}^n l_h(x_i)$. Therefore, the relationship between expected and empirical loss is of crucial importance for the performance of these learning algorithms.

In the classical approach to obtain generalization bounds, called GC-type bounds, one investigates the probability that, for *any* hypothesis in the class, the deviation of the empirical mean from the actual mean of its associated loss function is larger than a given threshold, that is,

$$Pr\Big\{\sup_{f \in F} \big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^n f(X_i)\big| \geq t\Big\}, \tag{1.1}$$

where μ is a probability measure on Ω, $X_1, \ldots, X_n$ are independent random variables distributed according to μ, F is the loss class associated with the learning problem, and $\mathbb{E}_\mu f$ denotes $\mathbb{E} f(X_i)$.

Classes of functions F which, independently of the underlying measure μ, satisfy the law of large numbers, that is, the probability (1.1) tends to 0 as n goes to infinity uniformly in μ, are called *uniform Glivenko-Cantelli classes*. For these classes learning is possible. Historically, uniform Glivenko-Cantelli classes were characterized by a finite combinatorial dimension (e.g., a finite VC dimension in the 0,1-case) (see Vapnik and Chervonenkis, 1971; Alon et al., 1997). The ability to bound the tails of the random variable in (1.1) is therefore due to the fact that the class F is "small" in some sense.

In Mendelson (2002a,b) it has been shown that parameters which also characterize the uniform Glivenko-Cantelli property are the *uniform Rademacher averages* of the class F, defined as

$$R_n(F) := \sup_{\{x_1,\ldots,x_n\} \subset \Omega^n} \mathbb{E}_\varepsilon \sup_{f \in F} \big|\sum_{i=1}^n \varepsilon_i f(x_i)\big|,$$

where $(\varepsilon_i)_{i=1}^n$ are independent, symmetric, $\{-1,1\}$-valued random variables, also known as Rademacher random variables. The necessary and sufficient condition for a class F to be a uniform Glivenko-Cantelli class is that $R_n(F) = o(n)$. In this case, tail estimates for (1.1) which are independent of the underlying measure μ can be as fast as the order of e^{-cnt^2}. Therefore, the Rademacher averages $R_n(F)$ seem to be a reasonable notion of "size" for a function class F in the context of learning via the uniform law of large numbers.

It was shown in Mendelson (2002b) that—if F satisfies mild structural assumptions—it is possible to derive sharper bounds on the learning problem by bounding the probability

$$Pr\Big\{\exists f \in F,\ \frac{1}{n}\sum_{i=1}^n f(X_i) \leq t,\ \mathbb{E}_\mu f \geq 2t\Big\} \tag{1.2}$$

instead of (1.1). The difference in the two cases is that in (1.1) one controls the deviation of the empirical means from the actual one for *all* the functions in the class, whereas in (1.2) the control is only for functions with "small" empirical mean, that is, potential minimizers for the actual mean. The tail estimates for (1.2) can be as good as e^{-cnt} instead of e^{-cnt^2} and are governed by the Rademacher averages $R_n(F^t)$ of the class $F^t := \{f \in F : \mathbb{E}_\mu f^2 \leq t\}$.

In other words, in both these cases measure independent estimates depend on the fact that for *all* coordinate projections, the projected sets $\{(f(x_1),...,f(x_n)) : f \in F\}$ and respectively $\{(f(x_1),...,f(x_n)) : f \in F^t\}$ are small in the sense that they have small Rademacher averages. Clearly, this is a property of the class F, and if one wishes to obtain useful bounds, one has to assume *a priori* that F is small.

In this article we show that the fact that classes have "small" coordinate projections with high probability (for a fixed probability measure) is the reason that the tails of (1.1) and (1.2) are well behaved. More surprising is the fact that this is also the reason why several other (seemingly very different) approaches yield generalization bounds.

The method of analysis we use is a combination of a symmetrization with respect to a random subclass and sharp concentration results. The need to investigate random subclasses is simple, as the starting point is that, ultimately, one wants to control the generalization ability *only* for the hypothesis functions which are reachable by the specific learning algorithm when presented with the actual training sample. Therefore, it suffices to obtain estimates on

$$Pr\Big\{\sup_{f \in F'} \Big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\Big| \geq t\Big\}, \tag{1.3}$$

where $F' \subset F$ and $\hat{f} \in F'$, where $\hat{f}$ is the loss function associated with the hypothesis produced by the algorithm from the sample $(X_1,\ldots,X_n)$. Although one can hope that F' has a smaller "size" than F, it is not possible to use the classical uniform generalization bounds because F' depends on the random training sample and could change with the sample.

The outline of this paper is as follows: in Section 2 we will first present a symmetrization procedure which is performed with respect to a random subclass of functions. The proof is a modification of the original proof of the standard symmetrization argument. The probability in Equation (1.3) can be therefore related to the probability of having large Rademacher sums for the (random) coordinate projections of this random subclass. Sharp generalization bounds can be obtained when the "size" of the set of coordinate projections of the random subclass is "small" in the sense that with high probability the Rademacher averages associated with a random projection of a random subclass are small (Corollary 2.5). We conclude that the general principle which ensures learnability and fast error rates consists of the combination of three main ingredients: a symmetrization procedure, a sharp concentration inequality, and a small "size" for the set of random coordinate projections.

In Section 3 we show how apparently different approaches fall within this general framework. In Section 3.1 we present the GC case (1.1) as an easy corollary of the symmetrization procedure. It is considerably more difficult to obtain the sharp rates for (1.2), a fact which is demonstrated in Section 3.2. In both these examples, the class whose coordinate projections have to be controlled is not random, but a deterministic subset of F.

There are several examples in which one really requires random subclasses of F. The examples we shall present are of the luckiness and algorithmic luckiness frameworks. The reason we chose these examples and not others is because our methods give considerably shorter proofs that seem to clarify the reason why luckiness and algorithmic luckiness work to the extent that they do.

In the luckiness and algorithmic luckiness frameworks, it is possible to avoid the detour via the worst-case quantity in (1.1) and to derive bounds using additional prior knowledge on the learning algorithm or on the training sample. The bounds stated in these approaches are, as opposed to the existing classical ones, data- or algorithm-dependent.

In the luckiness framework (Shawe-Taylor et al., 1998), prior knowledge about the connection between the actual sample and the functions in F is quantified through a luckiness function. A "fortunate" property of this luckiness function (ω-smallness) ensures good tail estimates for (1.3). One example for a luckiness function is the size of the margin for linear classifiers.

The algorithmic luckiness framework (Herbrich and Williamson, 2002) generalizes the luckiness framework. Prior knowledge about the link between the functions learned by the algorithm and the actual sample are formulated through an algorithmic luckiness function whose property of "ω-smallness" enables one to bound the generalization error.

Although there are no examples in which these methods yield results clearly better than other approaches, one merit of these two frameworks is that they set up a possible theoretical basis that could enable one to directly estimate data-dependent and algorithm-dependent generalization bounds. Unfortunately, the "ω-smallness" property is somewhat technically complicated and seems unnatural. The notion of complexity for the function class employed in these frameworks is covering numbers, and it was unexplored how these frameworks link to approaches using Rademacher averages as a notion of size.

We show that the ω-smallness condition is just a way of ensuring that a random coordinate projection of the random set is "small" and this suffices to recover the original generalization bounds. Hence, both luckiness and algorithmic luckiness fall within the general framework.

Note also that other examples of generalization bounds for data-dependent hypothesis classes that could be obtained using our methods are presented in Gat (1999) and Cannon et al. (2002). In fact, the proofs in Gat (1999) and Cannon et al. (2002) are based on a similar symmetrization argument. The notion of "size" which is employed is simply that of cardinality, and thus these results easily fall within the framework presented here (for details see Mendelson and Philips, 2003).

We end this introduction with some notation which will be used throughout the article. In the following, F is a class of real-valued functions defined on a measurable space Ω which take values in $[-1,1]$ and μ is a probability measure on Ω. Ω^n denotes the product space $\Omega \times \cdots \times \Omega$. Let $X_1,...,X_n$ be independent random variables distributed according to μ and let $(Y_1,...,Y_n)$ be an independent copy of $(X_1,...,X_n)$. μ_n denotes the random empirical probability measure supported on $\{X_1,...,X_n\}$, that is, $\mu_n := n^{-1}\sum_{i=1}^n \delta_{X_i}$. Pr_μ and $\mathbb{E}_\mu$ are the probability and the expectation with respect to μ and Pr_X and $\mathbb{E}_X$ denote the probability and the expectation with respect to the random vector $X = (X_1,\dots,X_n)$ (and therefore with respect to μ^n). $\mathbb{E}_\mu f$ is the expectation and $\mathrm{var}(f)$ is the variance of the random variable $f(X_i)$. In general, for any random variable Z, Pr_Z and $\mathbb{E}_Z$ are the probability and the expectation with respect to the distribution of Z.

Let F/X be the random set $\{(f(X_1),...,f(X_n)) : f \in F\}$, that is, the coordinate projection of the set F onto the random set of coordinates X. $VC(F)$ is the VC-dimension of F if F is a boolean class of functions.

Set ℓ_p^n to be $\mathbb{R}^n$ with the norm $\|x\|_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ and put B_p^n to be the unit ball of ℓ_p^n. ℓ_∞^n is $\mathbb{R}^n$ endowed with the norm $\|x\|_\infty := \sup_{1\le i\le n}|x_i|$, let $L_\infty(\Omega)$ be the set of bounded functions on Ω with respect to the norm $\|f\|_\infty := \sup_{\omega\in\Omega}|f(\omega)|$, and denote its unit ball by $B(L_\infty(\Omega))$. For a probability measure μ on a measurable space Ω and $1 \le p < \infty$, let $L_p(\mu)$ be the space of measurable functions on Ω with a finite norm $\|f\|_{L_p(\mu)} := (\int |f|^p d\mu)^{1/p}$.

Let (Y,d) be a metric space. If $F \subset Y$ then for every $\varepsilon > 0$, $N(\varepsilon, F, d)$ is the minimal number of open balls (with respect to the metric d) needed to cover F. A corresponding set $\{y_1, \ldots, y_m\} \subset Y$ of minimal cardinality such that for every $f \in F$ there is some y_i with $d(f, y_i) < \varepsilon$ is called an ε-cover of F. For $1 \le p < \infty$, denote by $N\big(\varepsilon, F, L_p(\mu_n)\big)$ the covering number of F at scale ε with respect to the $L_p(\mu_n)$ norm. Similarly, one can define the packing number at scale ε, which is the maximal cardinality of a set $\{y_1, \ldots, y_k\} \subset F$ such that for every $i \neq j$, $d(y_i, y_j) \ge \varepsilon$. Denote the ε-packing numbers by $M(\varepsilon, F, d)$ and note that for every $\varepsilon > 0$, $N(\varepsilon, F, d) \le M(\varepsilon, F, d) \le N(\varepsilon/2, F, d)$. If S is a set, we denote its complement by S^c.

Finally, throughout this article all absolute constants are denoted by c, C or K. Their values may change from line to line, or even within the same line.

2. Symmetrization

For every integer n, let F_n and F_n^{sym} be set-valued functions which assign to each $\sigma_n \in \Omega^n$ a subset of F. We assume that F_n^{sym} is invariant to permutations, that is, for every $\sigma_n \in \Omega^n$ and every permutation $\pi(\sigma_n)$ of σ_n, $F_n^{\mathrm{sym}}(\sigma_n) = F_n^{\mathrm{sym}}(\pi(\sigma_n))$, in which case we say that F_n^{sym} is symmetric.

The question we wish to address in this section is how to estimate the probability

$$Pr_X\Big\{ \sup_{f \in F_n(\sigma_n)} \big| \mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^n f(X_i) \big| \ge t \Big\}, \tag{2.1}$$

where $\sigma_n = (X_1, \ldots, X_n)$ is a random sample. Note that the worst-case probability (1.1) is a special case of (2.1), where F_n is the constant function mapping every sample to the whole function class F, $F_n(\sigma_n) = F$. Another extreme case occurs when $F_n(\sigma_n) = \{\hat f\}$, where $\hat f$ is the loss function associated with the hypothesis produced by a learning algorithm from the sample σ_n.

We will show that by employing an additional assumption on the functions F_n and F_n^{sym} which relates the random subsets $F_n(\sigma_n)$ to *symmetric* random subsets dependent on a double-sample, it is possible to upper bound (2.1) in terms of Rademacher sums associated with sets of coordinate projections.

Assumption 2.1 *There exists a constant* $\delta > 0$ *such that for every* $t > 0$*,*

$$\begin{aligned} &Pr_{X\times Y}\Big\{ \exists f \in F_n(\sigma_n),\ \Big| \frac{1}{n}\sum_{i=1}^n \big(f(X_i) - f(Y_i)\big) \Big| \ge t \Big\} \le \\ &Pr_{X\times Y}\Big\{ \exists f \in F_{2n}^{sym}(\sigma_n, \tau_n),\ \Big| \frac{1}{n}\sum_{i=1}^n \big(f(X_i) - f(Y_i)\big) \Big| \ge t \Big\} + \delta, \end{aligned} \tag{2.2}$$

where $\sigma_n = (X_1, \ldots, X_n)$ *and* $\tau_n = (Y_1, \ldots, Y_n)$.

This assumption quantifies that by replacing the original random subset of hypotheses with another symmetric random subset dependent on the double-sample—and which is therefore invariant under permutations of this double-sample—the probability of having large deviations of empirical means evaluated on the sample and ghost sample increases by at most δ and therefore not "too much".

Indeed, in all the applications we present δ can be made as small as we require. One extreme case occurs when for every double-sample (σ_n, τ_n),

$$F_n(\sigma_n) \subseteq F_{2n}^{\mathrm{sym}}(\sigma_n, \tau_n),$$

in which case Assumption 2.1 holds trivially with a constant $\delta = 0$. For example, if both set-valued maps are the constant function $F_n(\sigma_n) = F_n^{\text{sym}}(\sigma_n) = F$, then $\delta = 0$.

Given F_n, one can always define a mapping F_{2n}^{sym} to satisfy Assumption 2.1 as the symmetric extension of F_n: for every double-sample (σ_n, τ_n), $F_{2n}^{\text{sym}}(\sigma_n, \tau_n)$ is defined to be the union of all subsets corresponding to the first half of permutations of the double-sample (σ_n, τ_n),

$$F_{2n}^{\text{sym}}(\sigma_n, \tau_n) := \bigcup_{\pi \in S_{2n}} F_n\big(\pi(\sigma_n, \tau_n)|_{i=1}^n\big), \tag{2.3}$$

where S_{2n} is the set of permutations on $(1, \ldots, 2n)$, and $\pi(\sigma_n, \tau_n)|_{i=1}^n$ is the first half of the permuted double-sample. However, Assumption 2.1 allows us to replace the original subset $F_n(\sigma_n)$ even with a potentially "smaller" symmetric subset $F_{2n}^{\text{sym}}(\sigma_n, \tau_n)$ as long as the change in probabilities can be controlled.

The importance of Assumption 2.1 lies in the fact that it enables one to bound the probability (2.1) by proving a similar symmetrization argument to that employed in the original proof of the uniform Glivenko-Cantelli property.

The following symmetrization theorem is the main result of this section.

Theorem 2.1 *If Assumption 2.1 holds then for every $t > 0$,*

$$\Big(1 - \frac{4}{nt^2} \sup_{f \in F} \text{var}(f)\Big) \cdot Pr_X \Big\{ \exists f \in F_n(\sigma_n),\ \Big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^n f(X_i)\Big| \geq t \Big\}$$
$$\leq 2 Pr_{X \times Y} Pr_\varepsilon \Big\{ \exists f \in F_{2n}^{sym}(\sigma_n, \tau_n),\ \Big|\sum_{i=1}^n \varepsilon_i f(X_i)\Big| \geq \frac{nt}{4} \Big\} + \delta, \tag{2.4}$$

where $\sigma_n = (X_1, ..., X_n)$ and $\tau_n = (Y_1, ..., Y_n)$.

As the examples we present in the next section show, most of the standard methods used in Learning Theory fall within the general framework of Theorem 2.1. The advantage of Theorem 2.1 is that it reduces the analysis to a geometric problem, namely, estimating the Rademacher sums associated with the coordinate projection onto $\sigma_n = (X_1, ..., X_n)$ of the random class $F_{2n}^{\text{sym}}(\sigma_n, \tau_n)$,

$$F_{2n}^{\text{sym}}(\sigma_n, \tau_n)/\sigma_n := \Big\{ \big(f(X_1), ..., f(X_n)\big) :\ f \in F_{2n}^{\text{sym}}(\sigma_n, \tau_n) \Big\}.$$

The proof is done in two steps which are the same as in the standard symmetrization procedure: a symmetrization by a ghost sample which relates the deviation of the mean from the empirical mean to the deviation of the empirical means evaluated on two different samples; and a symmetrization by signs which relates the latter deviation to the probability of having "large" Rademacher sums $\sup_{f \in F_{2n}^{\text{sym}}(\sigma_n, \tau_n)} |\sum_{i=1}^n \varepsilon_i f(X_i)|$. We present the proof for the sake of completeness.

Lemma 2.2 (Symmetrization by a Ghost Sample) *For every $t > 0$,*

$$\Big(1 - \frac{4n}{t^2} \sup_{f \in F} \text{var}(f)\Big) Pr_X \Big\{ \exists f \in F_n(\sigma_n),\ \Big|\sum_{i=1}^n \big(f(X_i) - \mathbb{E}_\mu f\big)\Big| \geq t \Big\}$$
$$\leq Pr_{X \times Y} \Big\{ \exists f \in F_n(\sigma_n),\ \Big|\sum_{i=1}^n \big(f(X_i) - f(Y_i)\big)\Big| \geq \frac{t}{2} \Big\}.$$

Proof. Define the random processes $Z_i(f) = f(X_i) - \mathbb{E}_\mu f$ and $W_i(f) = f(Y_i) - \mathbb{E}_\mu f$, and fix $t > 0$. Let $\sigma_n = (X_1, ..., X_n)$, put $\beta = \inf_{f \in F} Pr\{|\sum_{i=1}^n Z_i(f)| \leq t/2\}$ and set $A = \{\sigma_n : \exists f \in F_n(\sigma_n), |\sum_{i=1}^n Z_i(f)| \geq t\}$. For every element in A there is some $f \in F_n(\sigma_n)$ and a realization of Z_i such that $|\sum_{i=1}^n Z_i(f)| \geq t$. Fix this realization and f and observe that by the triangle inequality, if $|\sum_{i=1}^n W_i(f)| \leq t/2$ then $\left|\sum_{i=1}^n \left(Z_i(f) - W_i(f)\right)\right| \geq t/2$. Since $(W_i)_{i=1}^n$ is an independent copy of $(Z_i)_{i=1}^n$,

$$\begin{aligned}\beta &\leq Pr_Y\left\{\left|\sum_{i=1}^n W_i(f)\right| \leq \frac{t}{2}\right\} \leq Pr_Y\left\{\left|\sum_{i=1}^n W_i(f) - \sum_{i=1}^n Z_i(f)\right| \geq \frac{t}{2}\right\} \\ &\leq Pr_Y\left\{\exists f \in F_n(\sigma_n),\ \left|\sum_{i=1}^n W_i(f) - \sum_{i=1}^n Z_i(f)\right| \geq \frac{t}{2}\right\}.\end{aligned}$$

Since the two extreme sides of this inequality are independent of the specific selection of f, this inequality holds on the set A. Integrating with respect to X on A it follows that

$$\begin{aligned}&\beta Pr_X\left\{\exists f \in F_n(\sigma_n),\ \left|\sum_{i=1}^n Z_i(f)\right| \geq t\right\} \\ \leq &Pr_X Pr_Y\left\{\exists f \in F_n(\sigma_n),\ \left|\sum_{i=1}^n \left(W_i(f) - Z_i(f)\right)\right| \geq \frac{t}{2}\right\}.\end{aligned}$$

Finally, to estimate β, note that by Chebyshev's inequality

$$Pr\left\{\left|\sum_{i=1}^n Z_i(f)\right| \geq \frac{t}{2}\right\} \leq \frac{4n}{t^2}\mathrm{var}(f)$$

for every $f \in F$, and thus, $\beta \geq 1 - (4n/t^2)\sup_{f \in F}\mathrm{var}(f)$. ∎

Proposition 2.3 (Symmetrization by Random Signs) *Let F_{2n}^{sym} be a symmetric map. Then, for any probability measure μ and every $t > 0$,*

$$\begin{aligned}&Pr_{X\times Y}\left\{\exists f \in F_{2n}^{sym}(\sigma_n, \tau_n),\ \left|\frac{1}{n}\sum_{i=1}^n \left(f(X_i) - f(Y_i)\right)\right| \geq t\right\} \\ &\qquad \leq 2Pr_{X\times Y}Pr_\varepsilon\left\{\exists f \in F_{2n}^{sym}(\sigma_n, \tau_n),\ \left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| \geq \frac{nt}{2}\right\},\end{aligned}$$

where $\sigma_n = (X_1, ..., X_n)$, $\tau_n = (Y_1, ..., Y_n)$, and $(\varepsilon_i)_{i=1}^n$ are independent Rademacher variables.

Proof. By the symmetry of F_{2n}^{sym} it follows that for every $\{\varepsilon_1, ..., \varepsilon_n\} \in \{-1, 1\}^n$,

$$\begin{aligned}&Pr_{X\times Y}\left\{\exists f \in F_{2n}^{\mathrm{sym}}(\sigma_n, \tau_n),\ \left|\frac{1}{n}\sum_{i=1}^n \left(f(X_i) - f(Y_i)\right)\right| \geq t\right\} \\ &\qquad = Pr_{X\times Y}\left\{\exists f \in F_{2n}^{\mathrm{sym}}(\sigma_n, \tau_n),\ \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(f(X_i) - f(Y_i)\right)\right| \geq t\right\}.\end{aligned}$$

Taking the expectation with respect to the random signs (that is, with respect to the Rademacher random variables), the proof follows from the triangle inequality and the fact that $(X_1, ..., X_n)$ has the same distribution as $(Y_1, ..., Y_n)$. ∎

Proof of Theorem 2.1. The claim follows immediately by combining Lemma 2.2, Assumption 2.1 and Proposition 2.3. ■

We can now relate the Rademacher sums from Theorem 2.1 to the Rademacher averages of $F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)/\sigma_n$ by employing concentration inequalities for the random variable

$$Z = \sup_{f\in F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)} \Big|\sum_{i=1}^n \varepsilon_i f(X_i)\Big| = \sup_{v\in F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)/\sigma_n} \Big|\sum_{i=1}^n \varepsilon_i v_i\Big|$$

around its conditional mean $\mathbb{E}_\varepsilon(Z|X_1,...,X_n,Y_1,...,Y_n)$.

For example, we state one particular concentration result which follows directly from martingale methods (see, e.g., McDiarmid, 1989) for functions with bounded differences:

Theorem 2.4 (Concentration) *For every set $V\subset B_\infty^n$ and every $t>0$,*

$$Pr_\varepsilon\Big\{\sup_{v\in V}\Big|\sum_{i=1}^n\varepsilon_i v_i\Big| - \mathbb{E}_\varepsilon\sup_{v\in V}\Big|\sum_{i=1}^n\varepsilon_i v_i\Big| > t\Big\} \le e^{-\frac{t^2}{2n}}. \tag{2.5}$$

Proof. Define $h(\varepsilon_1,\dots,\varepsilon_n) := \sup_{v\in V}\Big|\sum_{i=1}^n\varepsilon_i v_i\Big|$. By the triangle inequality, for every $1\le i\le n$,

$$\sup_{\{\varepsilon_1,\dots,\varepsilon_n,\tilde{\varepsilon}_i\}} |h(\varepsilon_1,\dots,\varepsilon_n) - h(\varepsilon_1,\dots,\varepsilon_{i-1},\tilde{\varepsilon}_i,\varepsilon_{i+1},\dots,\varepsilon_n)| \le 2,$$

and the claim follows directly from McDiarmid's inequality for h. ■

In particular, setting V to be the (random) coordinate projection of $F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)$ onto σ_n,

$$V := F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)/\sigma_n = \Big\{\big(f(X_1),...,f(X_n)\big) :\ f\in F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)\Big\}$$

and A_t to be the set of double-samples (σ_n,τ_n) with small Rademacher averages for the projections onto σ_n,

$$A_t := \Big\{(\sigma_n,\tau_n)\ :\ \mathbb{E}_\varepsilon\sup_{v\in V}\Big|\sum_{i=1}^n\varepsilon_i v_i\Big| \le nt/8\Big\},$$

it follows by the union bound and Equation (2.5) that

$$\begin{aligned} &Pr_{X\times Y}Pr_\varepsilon\Big\{\exists f\in F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n),\ \Big|\sum_{i=1}^n\varepsilon_i f(X_i)\Big| > \frac{nt}{4}\Big\} \\ &\qquad \le Pr_{X\times Y}\{A_t^c\} + e^{-\frac{nt^2}{128}}. \end{aligned}$$

Corollary 2.5 *If Assumption 2.1 holds, then for every $t>0$*

$$\begin{aligned} &\Big(1-\frac{4}{nt^2}\sup_{f\in F}\mathrm{var}(f)\Big)\cdot Pr_X\Big\{\exists f\in F_n(\sigma_n),\ \Big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^n f(X_i)\Big| \ge t\Big\} \\ &\qquad \le 2\Big(Pr_{X\times Y}\{A_t^c\} + e^{-\frac{nt^2}{128}}\Big) + \delta, \end{aligned}$$

where $V := F_{2n}^{sym}(\sigma_n,\tau_n)/\sigma_n$ and $A_t := \big\{(\sigma_n,\tau_n)\ :\ \mathbb{E}_\varepsilon\sup_{v\in V}\big|\sum_{i=1}^n\varepsilon_i v_i\big|\le nt/8\big\}$.

This corollary illustrates how Assumption 2.1 is sufficient to guarantee a generalization bound with tails of order e^{-cnt^2} for a learning algorithm drawing its hypotheses from the random set $F_n(\sigma_n)$, as soon as the Rademacher averages of the projection of $F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)$ onto σ_n are "small" with high probability.

3. Examples

In the previous section we have proved that, under the Assumption 2.1, a small "size" for the set of random coordinate projections together with a sharp concentration inequality allow us to derive tail estimates for random subsets F_n. In particular, in order to obtain tail estimates of the order of e^{-cnt^2}, by Corollary 2.5, it is sufficient to find symmetric random subsets $F_{2n}^{\mathrm{sym}}(\sigma_n, \tau_n)$ which satisfy Assumption 2.1 and for which the probability

$$Pr_{X \times Y}\left\{\mathbb{E}_\varepsilon\Big(\sup_{f \in F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)} \big|\sum_{i=1}^n \varepsilon_i f(X_i)\big| \Big| X_1,\ldots,X_n,Y_1,\ldots,Y_n\Big) > \frac{nt}{8}\right\},$$

is small. Now we are ready to show how apparently different approaches fall within the framework of Theorem 2.1.

Indeed, we will show that the tail estimate on (1.1) and the generalization bounds in the luckiness and the algorithmic luckiness frameworks can be derived directly from Corollary 2.5. We will illustrate this by specifying the corresponding maps F_n and F_{2n}^{sym}, and showing that for every fixed double-sample (σ_n, τ_n),

$$\mathbb{E}_\varepsilon \sup_{f \in F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)} \big|\sum_{i=1}^n \varepsilon_i f(x_i)\big|,$$

are sufficiently small (of the order $o(n)$).

As we present below, to recover the better estimates for (1.2) as in Mendelson (2002b) is more delicate because it requires a sharper concentration result than (2.5). We will show that these estimates as well follow from Theorem 2.1, by proving a different concentration inequality which will enable us to obtain the desired rates of the order of e^{-cnt}.

3.1 Glivenko-Cantelli Classes

In this section we will demonstrate how one can recover the optimal deviation estimates for uniform Glivenko-Cantelli classes directly from Corollary 2.5.

F is called a *uniform Glivenko-Cantelli class (GC class)* if for every $t > 0$

$$\lim_{n\to\infty} \sup_\mu Pr\left\{\sup_{f\in F} \big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^n f(X_i)\big| \geq t\right\} = 0.$$

If F is a uniform GC class, by selecting the constant functions

$$F_n(\sigma_n) = F, \;\; F_n^{\mathrm{sym}}(\sigma_n) = F,$$

Assumption 2.1 is trivially satisfied with $\delta = 0$ and

$$F_{2n}^{\mathrm{sym}}(\sigma_n, \tau_n)/\sigma_n = \Big\{\big(f(X_1),\ldots,f(X_n)\big) : \; f \in F\Big\}$$

for every double-sample (σ_n, τ_n). The fact that these coordinate projections are "small" follows from the characterization of uniform GC classes, an observation we shall return to later.

Theorem 3.1 *(Mendelson, 2002a) A class of uniformly bounded functions is a uniform GC class if and only if*

$$\lim_{n\to\infty} \sup_{\{x_1,\ldots,x_n\}\subset\Omega^n} \frac{1}{n}\mathbb{E}_\varepsilon \sup_{f\in F} \big|\sum_{i=1}^n \varepsilon_i f(x_i)\big| = 0.$$

Recall that

$$R_n(F) = \sup_{\{x_1,\ldots,x_n\}\subset\Omega^n} \mathbb{E}_\varepsilon \sup_{f\in F} \Big|\sum_{i=1}^n \varepsilon_i f(x_i)\Big|$$

and note that Theorem 3.1 ensures that the bound obtained from Theorem 2.1 is nonempty.

In particular, for every $t > 0$ let n_0 be such that for every $n \geq n_0$, $R_n(F) \leq nt/8$. Since $F \subset B\big(L_\infty(\Omega)\big)$ then $\sup_{f\in F} \mathrm{var}(f) \leq 1$, and thus $1-4\sup_{f\in F}\mathrm{var}(f)/nt^2 \geq 1/2$, provided that $n \geq 8/t^2$. Thus, by Corollary 2.5 and selecting $N = \max\{8/t^2, n_0\}$ it follows that for every integer $n > N$ and for any probability measure μ,

$$Pr\Big\{\sup_{f\in F}\Big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^n f(X_i)\Big| \geq t\Big\} \leq 8e^{-\frac{nt^2}{128}}.$$

In cases where one has *a priori* estimates on the size of the class (e.g., the shattering dimension or the uniform entropy), one can recover the optimal GC deviation results. For example, if $VC(F) = d$, then $R_n(F) \leq C\sqrt{dn}$ where C is an absolute constant, implying that one can take $n_0 = Cd/t^2$, and thus, for every $n \geq Ct^{-2}\max\{d, \log(1/\delta)\}$,

$$Pr\Big\{\sup_{f\in F}\Big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^n f(X_i)\Big| \geq t\Big\} \leq \delta.$$

Similar estimates can be recovered for classes with a polynomial shattering dimension by applying the bounds on $R_n(F)$ from Mendelson (2002a).

3.2 Learning Sample Complexity and Error Bounds

The learning sample complexity is governed by the probability that the empirical risk minimization algorithm fails, that is, it is the probability that an empirical minimizer of the loss functional (or more generally, an "almost empirical minimizer") will have a relatively large expectation. Formally, our aim is to estimate

$$Pr\Big\{\exists f \in F,\ \frac{1}{n}\sum_{i=1}^n f(X_i) \leq t,\ \mathbb{E}_\mu f \geq 2t\Big\}, \tag{3.1}$$

where F is the excess squared-loss class.

The required tail estimates follow from two principles: The first is a mild structural assumption on F, namely that F is star-shaped around 0 (i.e., for every $f \in F$ and $0 \leq t \leq 1$, $tf \in F$); the second is that there is some $B > 0$ such that for every $f \in F$, $\mathbb{E}_\mu f^2 \leq B\mathbb{E}_\mu f$. Note that there are many examples of loss classes for which this second assumption could be verified. For example, for nonnegative bounded loss functions, the associated loss function classes satisfy this property. For convex classes of functions bounded by 1, the associated excess squared-loss class satisfies this property as well, a result that was first shown in Lee, Bartlett, and Williamson (1998) and improved and extended in Bartlett, Jordan, and McAuliffe (2003) and Mendelson (2002b).

Under these two assumptions, one can show (Mendelson, 2002b) that for every $t > 0$,

$$\begin{aligned} &Pr\Big\{\exists f \in F,\ \frac{1}{n}\sum_{i=1}^n f(X_i) \leq t,\ \mathbb{E}_\mu f \geq 2t\Big\} \\ &\leq 2Pr\Big\{\sup_{f\in F,\ \mathbb{E}_\mu f^2 \leq Bt}\Big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^n f(X_i)\Big| \geq t\Big\}. \end{aligned} \tag{3.2}$$

For the sake of simplicity we present our results for $B = 1$, which is the case if F consists of nonnegative functions. The general case follows an identical path.

It is possible to obtain sharper deviation estimates for (3.2)—of the order of e^{-cnt} instead of e^{-cnt^2} like in the uniform GC case—as long as the largest variance of a class member is of the same order of magnitude as the required deviation. This follows directly from Talagrand's inequality (see Mendelson, 2002b; Bartlett, Bousquet, and Mendelson, 2004), and was the basis for the estimates on (3.1) in Mendelson (2002b). Unfortunately, the method of proof used in Mendelson (2002b) cannot be used directly in a way which fits our general principle. We thus present a different proof which uses the fact that "most" coordinate projections of $\{f \in F : \mathbb{E}_\mu f^2 \leq t\}$ are contained in a "small" Euclidean ball. Although the proof is slightly more complicated than the original one, the significance of having "small" coordinate projections is better exhibited.

Theorem 3.2 below is the main result of this section.

Theorem 3.2 *There are absolute constants K, c and c_1 for which the following holds. Let $F \subset B\big(L_\infty(\Omega)\big)$ be star-shaped around 0 such that for every $f \in F$, $\mathbb{E}_\mu f^2 \leq \mathbb{E}_\mu f$. If $t \geq c_1/n$ satisfies that*

$$\mathbb{E} \sup_{f \in F,\ \mathbb{E}_\mu f^2 \leq t} \Big| \sum_{i=1}^n \varepsilon_i f(X_i) \Big| \leq \frac{nt}{16}, \tag{3.3}$$

then

$$Pr\Big\{ \exists f \in F,\ \frac{1}{n}\sum_{i=1}^n f(X_i) \leq t,\ \mathbb{E}_\mu f \geq 2t \Big\} \leq Ke^{-cnt}.$$

For example, if t_0 is the minimal t which satisfies (3.3), F is a loss class in a proper learning scenario, and f^* denotes the loss function associated with the empirical minimizer, then with probability larger than $1 - Ke^{-cnt_0}$,

$$\mathbb{E}_\mu f^* \leq 2t_0.$$

Note that it is possible to estimate t_0, either via *a priori* assumptions on the function class, such as assumptions on the shattering dimension or the uniform entropy as in Mendelson (2002b), or from the sampled data as in Bartlett, Bousquet, and Mendelson (2004). For example, one can show (Mendelson, 2002b; Bartlett, Bousquet, and Mendelson, 2004) that if F is the star-shaped hull of a Boolean class G and 0, and if $VC(G) = d$, then $t_0 = O\big(\frac{d}{n}\log\big(\frac{en}{d}\big)\big)$. Hence, there are absolute constants c and C such that with probability larger that $1 - c(d/n)^d$,

$$\mathbb{E}_\mu f^* \leq C\frac{d}{n}\log\Big(\frac{en}{d}\Big).$$

The rest of this section will be devoted to the proof of Theorem 3.2.

First, let us denote the subset of functions in F with variance bounded by t by

$$F^t := \{f \in F,\ \mathbb{E}_\mu f^2 \leq t\}.$$

From (3.2) it follows that by setting $F_n = F_{2n}^{\text{sym}} := F^t$, and applying Theorem 2.1 (Assumption 2.1 holds trivially with $\delta = 0$), the probability we want to estimate is bounded by the probability $Pr_{X\times Y} Pr_\varepsilon \Big\{ \exists f \in F^t,\ \Big|\sum_{i=1}^n \varepsilon_i f(X_i)\Big| \geq \frac{nt}{4} \Big\}$.

We will show that the condition (3.3) is just a way of ensuring that, with high probability, the coordinate projections of F^t onto a random sample are small. This, together with a sharp concentration result will yield the desired result.

Let

$$Z_t(X_1,...,X_n) := \mathbb{E}_\varepsilon \sup_{f\in F^t} \big|\sum_{i=1}^n \varepsilon_i f(X_i)\big|.$$

The first step in the proof is based on an inequality due to Boucheron, Lugosi, and Massart which will allow us to bound the probability that Z_t deviates from its expectation.

Theorem 3.3 *(Boucheron, Lugosi, and Massart, 2003) Let $V_1,...,V_n$ be independent, identically distributed random variables which take values in a Banach space B, and assume that $\|V_i\| \le 1$ almost surely. Set*

$$Z := \mathbb{E}\big(\|\sum_{i=1}^n \varepsilon_i V_i\| \ \big|\ V_1,...,V_n\big).$$

Then, for any $t > 0$,

$$Pr\big(Z \ge \mathbb{E}Z + t\big) \le e^{-\frac{t^2}{2\mathbb{E}Z+2t/3}}.$$

To apply this theorem to $Z = Z_t$, let $B = \ell_\infty(F)$, which is the set of all bounded functions $z : F \to \mathbb{R}$ such that $\|z\|_{\ell_\infty(F)} = \sup_{f\in F}|z(f)|$. Let $V_i := X_i$ and define $X_i(f) := f(X_i)$. Hence, $\|X_i\|_B \le 1$ and $\|\sum_{i=1}^n \varepsilon_i V_i\|_B = \sup_{f\in F}|\sum_{i=1}^n \varepsilon_i f(X_i)|$.

If t is such that $\mathbb{E}Z_t \le nt/16$ then by Theorem 3.3,

$$Pr_X\Big\{\mathbb{E}_\varepsilon \sup_{f\in F^t}\Big|\sum_{i=1}^n \varepsilon_i f(X_i)\Big| > \frac{nt}{8}\Big\} \le e^{-cnt}, \tag{3.4}$$

where c is an absolute constant. Hence, with $F_{2n}^{\text{sym}}(\sigma_n,\tau_n) = F^t$ and

$$A_t = \{(\sigma_n,\tau_n) :\ \mathbb{E}_\varepsilon \sup_{f\in F^t}|\sum_{i=1}^n \varepsilon_i f(X_i)| \le nt/8\},$$

where $\sigma_n = (X_1,...,X_n)$ and $\tau_n = (Y_1,...,Y_n)$ as before, it follows that

$$Pr_{X\times Y}\{A_t^c\} \le e^{-cnt}.$$

For a fixed $(\sigma_n,\tau_n) \in A$, we require a sharp concentration result for

$$W_t(\varepsilon_1,...,\varepsilon_n) := \sup_{f\in F^t}|\sum_{i=1}^n \varepsilon_i f(X_i)|,$$

since (2.5) leads only to a tail estimate of e^{-cnt^2}. To that end, we use Talagrand's convex-distance inequality (Talagrand, 1995) (let us mention that our estimates also follow from an earlier result due to Johnson and Schechtman 1991). We will formulate the concentration result only in the context we require (see Ledoux, 2001, pg. 76).

Theorem 3.4 *Let $T \subset \ell_2^n$ and set $\sigma := \sup_{t\in T}\|t\|_{\ell_2^n}$. Define the random variable $G := \sup_{t\in T}|\sum_{i=1}^n \varepsilon_i t_i|$, and denote its median by M_G. Then for every $r > 0$,*

$$Pr\big\{|G - M_G| > r\big\} \le 4e^{-r^2/4\sigma^2},$$

and $|\mathbb{E}G - M_G| \le 4\pi\sigma$.

In our case, T is the image of F under the coordinate projection onto the random sample $\sigma_n = (X_1,...,X_n)$ where $(\sigma_n,\tau_n) \in A$. In order to bound σ we shall estimate the probability that a coordinate projection has a small diameter in ℓ_2^n.

Theorem 3.5 *Let F be a class of functions which map Ω into $[-1,1]$. For every $x > 0$ and r which satisfies that*

$$\mathbb{E} \sup_{f\in F,\ \mathbb{E}_\mu f^2 \le r} \Big|\sum_{i=1}^n \varepsilon_i f(X_i)\Big| \le \frac{nr}{20} - \frac{11x}{20},$$

then with probability at least $1-2e^{-x}$,

$$\Big\{f \in F:\ \mathbb{E}_\mu f^2 \le r\Big\} \subset \Big\{f \in F:\ \sum_{i=1}^n f^2(X_i) \le 2rn\Big\}.$$

Proof. The proof follows directly from the contraction inequality (see, e.g., Theorem 2.8 in Bartlett, Bousquet, and Mendelson, 2004) for $\phi(x) = x^2$ combined with Corollary 2.7 in Bartlett, Bousquet, and Mendelson (2004). ■

By our selection of t, it is easy to see that there is an absolute constant c, such that if $x = cnt$, then with probability larger that $1-2e^{-cnt}$, the radius of the projected set $F^t/\sigma_n \subset \ell_2^n$ is smaller than $\sqrt{2nt}$. In particular, we have

Corollary 3.6 *There are absolute constants c and c_1 for which the following holds. For every $t \ge c_1/n$ such that $\mathbb{E}Z_t \le nt/16$, there is a set A_t' of samples (σ_n,τ_n) which has probability larger than $1-3e^{-cnt}$, on which the set $V = \big\{\big(f(X_1),...,f(X_n)\big):\ f \in F^t\big\}$ is such that $\mathbb{E}_\varepsilon \sup_{v\in V} |\sum_{i=1}^n \varepsilon_i v_i| \le nt/8$ and $\sup_{v\in V} \|v\|_{\ell_2^n} \le \sqrt{2nt}$.*

Combining this corollary with Theorem 3.4, for every such set V,

$$Pr_\varepsilon\Big\{\sup_{v\in V}\Big|\sum_{i=1}^n \varepsilon_i v_i\Big| \ge \frac{nt}{4}\Big\} \le Pr_\varepsilon\Big\{\sup_{v\in V}\Big|\sum_{i=1}^n \varepsilon_i v_i\Big| \ge \mathbb{E}_\varepsilon \sup_{v\in V}\Big|\sum_{i=1}^n \varepsilon_i v_i\Big| + \frac{nt}{8}\Big\} \le 4e^{-cnt}$$

for an absolute constant c. Hence, there are absolute constants c and K such that

$$Pr_{X\times Y} Pr_\varepsilon\Big\{\exists f \in F^t,\ \Big|\sum_{i=1}^n \varepsilon_i f(X_i)\Big| \ge \frac{nt}{4}\Big\} \le Ke^{-cnt}. \tag{3.5}$$

Proof of Theorem 3.2. For every $t > 0$ let $F_n = F_{2n}^{\mathrm{sym}} := F^t$, and thus Assumption 2.1 holds with $\delta = 0$. Since for every $f \in F^t$, $\mathbb{E}_\mu f^2 \le t$, then

$$\big(1 - \frac{4}{nt^2}\sup_{f\in F^t} \mathrm{var}(f)\big) \ge \frac{1}{2}$$

provided that $t \ge 8/n$. Now the assertion follows from (3.2), Theorem 2.1, and (3.5). ■

3.3 Luckiness

In the luckiness framework introduced in Shawe-Taylor et al. (1998), bounds on the generalization error of functions are formulated *a posteriori*, after having seen a sample σ_n. The bounds are given in terms of an upper bound on some empirical, computable quantity dependent on the sample.

In the following, let n be a fixed sample size, d is a given fixed integer, and set $\delta \in (0,1]$. Three concepts are used in the luckiness framework. The first is the luckiness function $L : F \times \cup_k \Omega^k \longrightarrow \mathbb{R}$ which is invariant under permutations of the sample, that is, it depends only on the set $\{x_1,...,x_k\}$.

Using the luckiness function one can construct sample dependent subsets of F, called *lucky sets* in the following manner; for every sample ζ and $f \in F$, the lucky set consists of all the functions luckier on this sample than the given function, that is,

$$H(f,\zeta) := \{g \in F : L(g,\zeta) \geq L(f,\zeta)\}.$$

Observe that the luckiness function imposes a structure of increasing subsets of F, because $H(g,\zeta) \subseteq H(f,\zeta)$ if and only if $L(g,\zeta) \geq L(f,\zeta)$, a fact which will allow us to define $F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)$.

Lemma 3.7 *For every integer d and sample ζ there is a unique set $H_d(\zeta)$ with the following properties:*

1. $M\left(\frac{1}{n}, H_d(\zeta), L_1(\mu_n)\right) \leq 2^d$*, where μ_n is the empirical measure supported on ζ.*
2. *If $f \in F$ satisfies that $M\left(\frac{1}{n}, H(f,\zeta), L_1(\mu_n)\right) \leq 2^d$ then $f \in H_d(\zeta)$.*

Proof. Let $A := \{f \in F \,:\, M\left(\frac{1}{n}, H(f,\zeta), L_1(\mu_n)\right) \leq 2^d\}$ and set $H_d(\zeta) := \bigcup_{f \in A} H(f,\zeta)$. To see that $H_d(\zeta)$ has the required properties, note that if $K \subset H_d(\zeta)$ is a finite $1/n$-separated set with respect to $L_1(\mu_n)$, then there is some $f \in A$ such that $K \subset H(f,\zeta)$, implying that $|K| \leq 2^d$. The second property and the uniqueness are easily verified. ■

For every double-sample $\zeta = (\sigma_n,\tau_n)$ we set

$$F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n) := H_d(\sigma_n,\tau_n), \tag{3.6}$$

and observe that this random class is permutation invariant, implying that F_{2n}^{sym} is symmetric.

The second ingredient in the luckiness framework, the ω-function, $\omega : \mathbb{R} \times \mathbb{N} \times (0,1] \to \mathbb{N}$, is used to define $F_n(\sigma_n)$. Given a luckiness function L and an ω-function, then for a fixed integer d and $\delta \in (0,1]$, define

$$F_n(\sigma_n) := \{f \in F :\ \omega\big(L(f,\sigma_n), n, \delta\big) \leq 2^d\}. \tag{3.7}$$

The third ingredient is the ω-smallness condition, which is a joint property of the luckiness and ω functions. It states that for every $n \in \mathbb{N}$, every $\delta \in (0,1]$ and every probability measure μ

$$Pr_{X\times Y}\left\{\exists f \in F : M\left(\tfrac{1}{n}, H(f,(\sigma_n,\tau_n)), L_1(\mu_{2n})\right) > \omega\big(L(f,\sigma_n), n, \delta\big)\right\} < \delta. \tag{3.8}$$

Examples for luckiness functions are the empirical VC-dimension of a binary function class with respect to a sample—in this case all lucky sets are equal to the whole set F—and the margin of linear classifiers. Their corresponding ω functions can be found in Shawe-Taylor et al. (1998). Although the luckiness framework gives a unified proof for existing generalization bounds, finding a pair of luckiness and ω-functions seems to be difficult, because of the quite technical and counterintuitive ω-smallness condition.

The following result shows that the ω-smallness of L ensures that Assumption 2.1 holds, and that, with high probability, $F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)/\sigma_n$ is sufficiently small. Therefore, it is just a way of requiring that $F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)$ has, with high probability, small random coordinate projections.

Lemma 3.8 *For fixed integers n and d, and* $\delta\in(0,1]$*, let* F_n *and* F_{2n}^{sym} *be defined as in* (3.7) *and* (3.6)*. If a luckiness function L and an* ω*-function satisfy the* ω*-smallness condition* (3.8)*, then for every* $t>0$*,*

$$Pr_{X\times Y}\Big\{\exists f\in F_n(\sigma_n),\ \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i)-f(Y_i)\big)\Big|\geq t\Big\}$$
$$\leq Pr_{X\times Y}\Big\{\exists f\in F_{2n}^{sym}(\sigma_n,\tau_n),\ \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i)-f(Y_i)\big)\Big|\geq t\Big\}+\delta.$$

Proof. For a fixed double sample $\zeta=(\sigma_n,\tau_n)$ let μ_{2n} be the empirical measure supported on ζ. Put

$$A_\zeta:=\Big\{f\in F\ :M\big(\tfrac{1}{2n},H(f,(\sigma_n,\tau_n)),L_1(\mu_{2n})\big)\leq\omega\big(L(f,\sigma_n),n,\delta\big)\Big\}$$

and

$$B_\zeta:=\Big\{f\in F:M\big(\tfrac{1}{2n},H(f,(\sigma_n,\tau_n)),L_1(\mu_{2n})\big)\leq 2^d\Big\}.$$

Note that $F_n(\sigma_n)\cap A_\zeta\subseteq B_\zeta\subseteq F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)$. By the ω-smallness condition,

$$Pr_{X\times Y}\{\exists f\in(A_\zeta)^c\}\leq\delta,$$

and by the union bound for disjoint sets,

$$\begin{aligned}Pr_{X\times Y}&\Big\{\exists f\in F_n(\sigma_n),\ \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i)-f(Y_i)\big)\Big|\geq t\Big\}\\&=Pr_{X\times Y}\Big\{\exists f\in F_n(\sigma_n)\cap A_\zeta,\ \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i)-f(Y_i)\big)\Big|\geq t\Big\}\\&\quad+Pr_{X\times Y}\Big\{\exists f\in F_n(\sigma_n)\cap(A_\zeta)^c,\ \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i)-f(Y_i)\big)\Big|\geq t\Big\},\end{aligned}\tag{3.9}$$

and our claim follows. ∎

Now, we are ready to formulate the generalization bound for the luckiness framework.

Theorem 3.9 *Let L and* ω *be functions satisfying the* ω*-smallness condition* (3.8)*. Then, for every probability measure* μ*, every* $d\in\mathbb{N}$ *and every* $\delta\in(0,1]$*, there is a set of probability larger than* $1-12\delta$ *such that if* $\omega\big(L(f,\sigma_n),n,\delta\big)\leq 2^d$*, then*

$$\Big|\mathbb{E}_\mu f-\frac{1}{n}\sum_{i=1}^{n}f(X_i)\Big|\leq C\sqrt{\frac{d}{n}\log\frac{1}{\delta}},$$

where C is an absolute constant.

Proof. Let F_n and F_{2n}^{sym} be defined as above, and observe that

$$M\big(\tfrac{1}{n}, F_{2n}^{\text{sym}}(\sigma_n,\tau_n), L_1(\mu_{2n})\big) \le 2^d \tag{3.10}$$

for every (σ_n,τ_n). By Corollary 2.5 we have to estimate

$$Pr_{X\times Y}\Big\{\mathbb{E}_\varepsilon \sup_{f\in F_{2n}^{\text{sym}}(\sigma_n,\tau_n)} \Big|\sum_{i=1}^n \varepsilon_i f(X_i)\Big| > \frac{nt}{8}\Big\},$$

where $\sigma_n = (X_1,\dots,X_n)$ and $\tau_n = (Y_1,\dots,Y_n)$. Let

$$V := \Big\{\big(f(X_1),\dots,f(X_n)\big) : f \in F_{2n}^{\text{sym}}(\sigma_n,\tau_n)\Big\} \subset \ell_2^n,$$

put μ_{2n} to be the empirical measure supported on $\zeta = (\sigma_n,\tau_n)$ and set ν_n to be the empirical measure supported on σ_n. Note that for every f,g, $\mathbb{E}_{\mu_{2n}}|f-g| \ge \mathbb{E}_{\nu_n}|f-g|/2$. Thus, every $1/n$-cover of $F_{2n}^{\text{sym}}(\sigma_n,\tau_n)$ in $L_1(\mu_{2n})$ is a $2/n$-cover of the same set in $L_1(\nu_n)$. In particular, if A is a maximal $1/n$-packing of $F_{2n}^{\text{sym}}(\sigma_n,\tau_n)$ in $L_1(\mu_{2n})$, it is a $2/n$ cover of that set in $L_1(\nu_n)$. It is easy to verify that $B\big(L_1(\nu_n)\big) = nB_1^n$, and in particular, $V \subset A + \frac{2}{n}\cdot nB_1^n = A + 2B_1^n$, where $A+B = \{a+b \,: a\in A,\, b\in B\}$. By the triangle inequality,

$$\begin{aligned}\mathbb{E}_\varepsilon \sup_{f\in F_{2n}^{\text{sym}}(\sigma_n,\tau_n)} \Big|\sum_{i=1}^n \varepsilon_i f(x_i)\Big| &= \mathbb{E}_\varepsilon \sup_{v\in V}\Big|\sum_{i=1}^n \varepsilon_i v_i\Big| = \mathbb{E}_\varepsilon \sup_{a\in A,\, b\in B_1^n}\Big|\sum_{i=1}^n \varepsilon_i(a_i+2b_i)\Big| \\ &\le \mathbb{E}_\varepsilon \sup_{a\in A}\Big|\sum_{i=1}^n \varepsilon_i a_i\Big| + 2\mathbb{E}_\varepsilon \sup_{b\in B_1^n}\Big|\sum_{i=1}^n \varepsilon_i b_i\Big|.\end{aligned}$$

The first term can be bounded by a corollary of Slepian's inequality (Pisier, 1989), which states that there is an absolute constant C such that for every $A \subset \ell_2^n$,

$$\mathbb{E}_g \sup_{a\in A}\Big|\sum_{i=1}^n g_i a_i\Big| \le C\sqrt{\log|A|}\sup_{u,v\in A}\|u-v\|_2,$$

where $(g_i)_{i=1}^n$ are independent standard gaussian random variables.

Since our class consists of functions bounded by 1, then $V \subset B_\infty^n \subset \sqrt{n}B_2^n$ and since the Rademacher averages are upper bounded (up to an absolute constant) by the gaussian ones (Milman and Schechtman, 2001), then

$$\mathbb{E}_\varepsilon \sup_{a\in A}\Big|\sum_{i=1}^n \varepsilon_i a_i\Big| \le C\mathbb{E}_g \sup_{a\in A}\Big|\sum_{i=1}^n g_i a_i\Big| \le C\sqrt{\log|A|}\sqrt{n} \le C\sqrt{nd},$$

where the final inequality holds because $|A| \le 2^d$ by (3.10).

In order to estimate the second term, one can apply the triangle inequality to show that

$$\mathbb{E}_\varepsilon \sup_{b\in B_1^n}\Big|\sum_{i=1}^n \varepsilon_i b_i\Big| \le 1.$$

In conclusion

$$\mathbb{E}_\varepsilon \sup_{f\in F_{2n}^{\text{sym}}(\sigma_n,\tau_n)}\Big|\sum_{i=1}^n \varepsilon_i f(x_i)\Big| \le C\sqrt{nd}.$$

To complete the proof, apply Corollary 2.5 for $t = C\sqrt{\frac{d}{n}\log(1/\delta)}$. ■

3.4 Algorithmic Luckiness

In the algorithmic luckiness framework (Herbrich and Williamson, 2002), the generalization error bound is also formulated *a posteriori*, after having seen a sample. It differs from the luckiness framework because it gives bounds on the generalization error of the function learned by the learning algorithm from the sample at hand. Again, the bound is given in terms of a computable quantity dependent on the sample and on the algorithm.

In a similar fashion to the luckiness framework, an algorithmic luckiness function and an ω-function are introduced in order to define the functions F_n and F_{2n}^{sym}. The functions L and ω satisfy a joint smallness condition which ensures that Assumption 2.1 holds, and that the size of the projection $F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)/\sigma_n$ is sufficiently small.

As we did before, fix a sample size n, an integer d and some $\delta \in (0,1]$. Denote by $\mathcal{A}$ a fixed learning algorithm, by $\mathcal{A}(\zeta)$ the loss function associated with the hypothesis produced by the algorithm from the sample ζ, and set $\mathcal{A}(F) = \{f = \mathcal{A}(\zeta) : \ \zeta \in \Omega^n\}$.

The algorithmic luckiness function is a function $L : \mathcal{A}(F) \longrightarrow \mathbb{R}$. For a sample ζ of size $2n$, the *lucky set* $G(\zeta)$ is defined as the subset of losses of functions learned by the algorithm on the first half of the sample, when permuting the whole sample, as long as the function the algorithm produced on the first half of the permuted sample is "luckier" than on the original one. Formally, let S_{2n} be the set of permutations on $\{1,...,2n\}$, and for every $\zeta = (\zeta_1,...,\zeta_{2n})$, set $\zeta|_{i=1}^n = (\zeta_1,...,\zeta_n)$. Define the lucky set as

$$G(\zeta) := \Big\{\mathcal{A}\big(\pi(\zeta)|_{i=1}^n\big) : \ L\big(\mathcal{A}(\pi(\zeta)|_{i=1}^n)\big) \geq L\big(\mathcal{A}(\zeta|_{i=1}^n)\big),\ \pi \in S_{2n}\Big\}.$$

If $G_{\mathcal{A}}(\zeta)$ is the subset of losses corresponding to functions learned by $\mathcal{A}$ on the first half of all the permutations of the double-sample ζ, then $G(\zeta) \subset G_{\mathcal{A}}(\zeta)$, and clearly, $|G_{\mathcal{A}}(\zeta)| \leq (2n)! < \infty$. Therefore, we can order the functions in decreasing order according to their luckiness. Define the ordered set

$$G_{\mathcal{A}}(\zeta) := \Big[\underbrace{f_1, f_2, f_3, \ldots, f_{k-1}, f_k}_{G(\zeta)}, f_{k+1}, \ldots, f_m\Big],$$

and for the sake of simplicity, assume that for every $i < j$, $L(f_i) > L(f_j)$. Only a small modification is required in the general case, where some functions might have the same luckiness.

Set $f_k = \mathcal{A}(\zeta|_{i=1}^n)$ and let $G_{\mathcal{A}}^{\ell}(\zeta)$ be the subset consisting of the first ℓ functions in $G_{\mathcal{A}}(\zeta)$, that is, $G_{\mathcal{A}}^{\ell}(\zeta) = \{f_1, f_2, f_3, \ldots, f_\ell\}$.

For the given integer d and the double-sample (σ_n,τ_n) put k^* to be the largest integer such that

$$M\big(\tfrac{1}{n}, G_{\mathcal{A}}^{k^*}((\sigma_n,\tau_n)), L_1(\mu_{2n})\big) \leq 2^d \text{ and } M\big(\tfrac{1}{n}, G_{\mathcal{A}}^{k^*+1}((\sigma_n,\tau_n)), L_1(\mu_{2n})\big) > 2^d.$$

Then, by setting

$$F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n) := G_{\mathcal{A}}^{k^*}((\sigma_n,\tau_n)) \tag{3.11}$$

it follows that F_{2n}^{sym} is symmetric, since the learning algorithm is permutation invariant.

The ω-function, $\omega : \mathbb{R} \times \mathbb{N} \times (0,1] \longrightarrow \mathbb{N}$ is used to define $F_n(\sigma_n)$. Indeed, define

$$F_n(\sigma_n) := \begin{cases} \{\mathcal{A}(\sigma_n)\} & \text{if } \omega\big(L(\mathcal{A}(\sigma_n)), n, \delta\big) \leq 2^d \\ \emptyset & \text{otherwise,} \end{cases} \tag{3.12}$$

and note that $|F_n(\sigma_n)| \leq 1$.

Finally, the ω-smallness condition states that for every integer n, every $\delta \in (0,1]$, and every probability measure μ,

$$P_{X\times Y}\Big\{M\big(\tfrac{1}{n}, G((\sigma_n,\tau_n)), L_1(\mu_{2n})\big) \geq \omega\big(L(\mathcal{A}(\sigma_n)), n, \delta\big)\Big\} < \delta, \tag{3.13}$$

and as we show, it assures that Assumption 2.1 holds.

Lemma 3.10 *Let $\mathcal{A}$ be a learning algorithm, fix an integer d and some $\delta \in (0,1]$, and let F_n and F_{2n}^{sym} be as in* (3.12) *and* (3.11). *If a luckiness function and ω-function satisfy the ω-smallness condition* (3.13)*, then for every $t > 0$*

$$Pr_{X\times Y}\Big\{\exists f \in F_n(\sigma_n) : \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(Y_i)\big)\Big| \geq t\Big\}$$
$$\leq Pr_{X\times Y}\Big\{\exists f \in F_{2n}^{sym}(\sigma_n,\tau_n) : \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(Y_i)\big)\Big| \geq t\Big\} + \delta.$$

Proof. For every double sample $\zeta = (\sigma_n,\tau_n)$, let μ_{2n} be the empirical measure supported on (σ_n,τ_n) and define two random sets in the following manner. Let $A_\zeta := \{\mathcal{A}(\sigma_n)\}$ if $M\big(\frac{1}{n}, G((\sigma_n,\tau_n)), L_1(\mu_{2n})\big) < \omega\big(L(\mathcal{A}(\sigma_n)), n, \delta\big)$ and the empty set otherwise, and put $B_\zeta := \{\mathcal{A}(\sigma_n)\}$ if $M\big(\frac{1}{n}, G((\sigma_n,\tau_n)), L_1(\mu_{2n})\big) \leq 2^d$ and the empty set otherwise. Note that for every ζ, $F_n(\sigma_n)\cap A_\zeta \subset B_\zeta \subset F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)$. Moreover, if $F_n(\sigma_n)\cap (A_\zeta)^c \neq \emptyset$, then $F_n(\sigma_n) = \{\mathcal{A}(\sigma_n)\}$ and $A_\zeta = \emptyset$. Thus, by the ω-smallness condition,

$$Pr_{X\times Y}\Big\{F_n(\sigma_n)\cap (A_\zeta)^c \neq \emptyset\Big\} \leq Pr_{X\times Y}\Big\{A_\zeta = \emptyset\Big\} < \delta.$$

Finally, for every $t > 0$,

$$\begin{aligned}
&P_{X\times Y}\Big\{\exists f \in F_n(\sigma_n),\ \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(Y_i)\big)\Big| \geq t\Big\}\\
&= Pr_{X\times Y}\Big\{\exists f \in F_n(\sigma_n)\cap A_\zeta,\ \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(Y_i)\big)\Big| \geq t\Big\}\\
&+ Pr_{X\times Y}\Big\{\exists f \in F_n(\sigma_n)\cap \big(A_\zeta\big)^c,\ \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(Y_i)\big)\Big| \geq t\Big\}\\
&\leq Pr_{X\times Y}\Big\{\exists f \in F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n),\ \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(Y_i)\big)\Big| \geq t\Big\} + \delta,
\end{aligned}$$

as claimed. ■

The definition of $F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)$ assures that the covering numbers of $F_{2n}^{\mathrm{sym}}(\sigma_n,\tau_n)$ are small, and by Corollary 2.5 we obtain a result analogous to Theorem 3.9, which recovers the main result of Herbrich and Williamson (2002).

Theorem 3.11 *Let $\mathcal{A}$ be a learning algorithm which takes values in $B\big(L_\infty(\Omega)\big)$, and let L and ω be functions satisfying the ω-smallness condition* (3.13). *Then, for every probability measure*

μ, every $d \in \mathbb{N}$ and every $\delta \in (0,1]$, there is a set of probability at least $1-12\delta$ such that if $\omega\big(L(\mathcal{A}(\sigma_n)),n,\delta\big) \leq 2^d$, then

$$\Big|\mathbb{E}_\mu\big(\mathcal{A}(\sigma_n)\big) - \mathbb{E}_{\mu_n}\big(\mathcal{A}(\sigma_n)\big)\Big| \leq C\sqrt{\frac{d}{n}\log\frac{1}{\delta}},$$

where C is an absolute constant.

Acknowledgements

We would like to thank Ran Bachrach, Olivier Bousquet, Gideon Schechtman, and Bob Williamson for their valuable suggestions and comments.

References

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, vol. 44(4), 615-631, 1997.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 2004, to appear.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, UC Berkeley, 2003.

S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, vol 31(3), 1583–1614, 2003.

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes-Rendus de l'Académie Scientifique de Paris*, Ser. I, 334, 495–500, 2002.

A. H. Cannon, J. M. Ettinger, D. R. Hush, and J. C. Scovel. Machine learning with data dependent hypothesis classes. *Journal of Machine Learning Research*, 2, 335–358, 2002.

Y. Gat. A bound concerning the generalization ability of a certain class of learning algorithms. Technical Report 548, UC Berkeley, March 1999.

R. Herbrich and R. C. Williamson. Algorithmic luckiness. *Journal of Machine Learning Research*, 3, 175–212, 2002.

W. B. Johnson and G. Schechtman. A remark on Talagrand's deviation inequality for Rademacher functions. In *Functional Analysis (Austin, TX, 1987/1989)*, Lecture Notes in Mathematics 1470, 72–77, Springer, 1991.

M. Ledoux. *The concentration of measure phenomenon*. Mathematical Surveys and Monographs, Vol 89, AMS, 2001.

W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5), 1974–1980, 1998.

C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, London Mathematical Society Lecture Note Series 141, 148–188, Cambridge University Press, 1989.

S. Mendelson. Rademacher averages and phase transitions in Glivenko-Cantelli classes. *IEEE Transactions on Information Theory*, 48(1), 251–263, 2002a.

S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7), 1977–1991, 2002b.

S. Mendelson. A few notes on Statistical Learning Theory. In *Proceedings of the Machine Learning Summer School, Canberra 2002, S. Mendelson and A. J. Smola (Eds.)*, Lecture Notes in Computer Sciences LNCS 2600, 1–40, Springer, 2003.

S. Mendelson and P. Philips. Random subclass bounds. In *Proceedings of the 16th Annual Conference on Computational Learning Theory*, 329–343, Springer, 2003.

S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Inventiones Mathematicae*, 152, 37–55, 2003.

V. D. Milman and G. Schechtman. *Asymptotic Theory of Finite Dimensional Normed Spaces*. Lecture Notes in Mathematics 1200, Springer, 2001.

G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge University Press, 1989.

J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5), 1926–1940, 1998.

M. Talagrand. Majorizing measures: The generic chaining. *The Annals of Probability*, 24, 1049–1103, 1996.

M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.* 81, 73–205, 1995.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2), 264–280, 1971.