RADEMACHER PENALTIES AND STRUCTURAL RISK MINIMIZATION

Vladimir Koltchinskii

We suggest a penalty function to be used in various problems of structural risk minimization. This penalty is data dependent and is based on the sup-norm of the so called Rademacher process indexed by the underlying class of functions (sets). The standard complexity penalties, used in learning problems and based on the VC-dimensions of the classes, are conservative upper bounds (in a probabilistic sense, uniformly over the set of all underlying distributions) for the penalty we suggest. Thus, for a particular distribution of training examples one can expect better performance of learning algorithms with the data-driven Rademacher penalties. We obtain oracle inequalities for the theoretical risk of estimators, obtained by structural minimization of the empirical risk with Rademacher penalties. The inequalities imply some form of optimality of the empirical risk minimizers. We also suggest an iterative approach to structural risk minimization with Rademacher penalties, in which the hierarchy of classes is not given in advance, but is determined in the data-driven iterative process of risk minimization. We prove probabilistic oracle inequalities for the theoretical risk of the estimators based on this approach as well.

Index Terms: Structural Risk Minimization, Iterative Structural Risk Minimization, Rademacher Penalty, Oracle Inequalities, Empirical Process, Classification

1. Dimension based penalties and Rademacher penalties in risk minimization. Let Y be a $\{0,1\}$-valued random variable (label) to be predicted based on an observation of another random variable X taking values in a measurable space $(S,\mathcal{A})$. A decision rule is a measurable set $C\in\mathcal{A}$, or, equivalently, the measurable function $g=I_C$, where

$$I_C(x):=\begin{cases}1 & \text{if } x\in C\\ 0 & \text{otherwise.}\end{cases}$$

The risk of the decision rule C is defined by $L(C):=\mathbb{P}(\{Y\neq I_C(X)\})$. It is well known that the optimal decision rule (the one that minimizes the risk on $\mathcal{A}$) is given by

$$C_{\text{opt}}:=\Big\{x:\mathbb{P}\{Y=0|X=x\}\le\mathbb{P}\{Y=1|X=x\}\Big\}.$$

To determine the set C_{opt} one has to know the joint distribution of (X,Y). Most often, this distribution is unknown and determining the decision rule is to be based on the sample $((X_1,Y_1),\dots,(X_n,Y_n))$ of independent copies of (X,Y) (the training data). Given a class $\mathcal{C}$ of decision rules, the estimate of the "optimal" decision rule is determined by minimization of the empirical risk $\hat{C}_n:=\text{argmin}\{L_n(C):C\in\mathcal{C}\}$, where $L_n(C)$ is the average classification error of the decision rule C on the training data:

$$L_n(C):=n^{-1}\sum_{j=1}^{n}I_{\{Y_j\neq I_C(X_j)\}}.$$

This is the well known method of empirical risk minimization frequently used in the problems of concept learning (pattern recognition, statistical classification) at least since the landmark works of Vapnik and Chervonenkis (1971, 1974) (see also Vapnik (1982, 1995, 1998), Devroye, Györfi and Lugosi (1996), Vidyasagar (1997)). It plays also an important role in computational learning theory (Valiant (1984), Blumer, Ehrenfeucht, Haussler and Warmuth (1989)).

The choice of the class $\mathcal{C}$ of decision rules poses a hard problem. Most often, the available prior information about the unknown distribution of (X,Y) is not enough to determine a reasonable class $\mathcal{C}$ that contains C_{opt}. In an attempt to make the minimal risk $\min_{C\in\mathcal{C}}L(C)$ smaller, one can try to choose very large class $\mathcal{C}$. This results in poor approximation of the risk L by the empirical risk L_n on the class $\mathcal{C}$. In such cases,

The research was partially supported by NSA grant MDA904-99-1-0031 and by a Humboldt Research Fellowship

V. Koltchinskii is with the Department of Mathematics and Statistics, The University of New Mexico, Albuquerque NM 87131-1141, USA; e-mail: vlad@math.unm.edu

the solution $\hat{C}_n$ of the empirical risk minimization problem does not have to be close to $C_{\rm opt}$ and the risk of this solution does not have to be small. This leads to the necessity to take into account the "complexity" of the class $\mathcal{C}$. The standard way to measure the complexity is based on the notion of VC-dimension of the class. Given a finite set $F \subset S$, denote $\Delta^{\mathcal{C}}(F) := {\rm card}(\{F \cap C : C \in \mathcal{C}\})$ and

$$m(\mathcal{C}, n) := \sup\{\Delta^{\mathcal{C}}(F) : F \subset S,\ {\rm card}(F) = n\},\ n \geq 1.$$

Then the VC-dimension of the class $\mathcal{C}$ is defined as $V(\mathcal{C}) := \sup\{n \geq 1 : m(\mathcal{C}, n) = 2^n\}$.

Consider now a nondecreasing sequence $\{\mathcal{C}_N\}_{N\geq 1}$ of classes of decision rules (a sieve). Vapnik's method of *structural risk minimization* is based on minimizing the so called penalized empirical risk:

$$\hat{C} := {\rm argmin}\{L_n(C) : C \in \hat{\mathcal{C}}_{\hat{N}}\},\ \hat{N} := {\rm argmin}\{N \geq 1 : \min_{C \in \mathcal{C}_N} L_n(C) + {\rm pen}(n; N)\},$$

where ${\rm pen}(n; N)$ is the complexity penalty of the class $\mathcal{C}_N$. A standard choice of the complexity penalty is as follows:

$$\text{(1.1)} \qquad {\rm pen}(n; N) := \sqrt{\frac{\log(4e^8 m(\mathcal{C}_N, n^2)) + N}{2n}},$$

which is, roughly, const $\sqrt{(V(\mathcal{C}_N)\log n + N)/n}$ (see, e.g., Lugosi and Zeger (1996)). This particular choice is based on the following bound (due to Devroye) for the deviations of the empirical risk from the theoretical one uniformly over a class $\mathcal{C}$ of the decision rules:

$$\text{(1.2)} \qquad \mathbb{P}\{\sup_{C \in \mathcal{C}} \bigl|L_n(C) - L(C)\bigr| \geq \varepsilon\} \leq 4e^8 m(\mathcal{C}, n^2) e^{-2n\varepsilon^2}.$$

Lugosi and Zeger (1996) established the following bounds for the estimator $\hat{C}$:

$$\text{(1.3)} \qquad \mathbb{P}\bigl\{L(\hat{C}) - \inf_{C \in \mathcal{C}_N} L(C) \geq \varepsilon\bigr\} \leq e^{-n\varepsilon^2/2} + 4e^8 m(\mathcal{C}_N; n^2) e^{-n\varepsilon^2/8},$$

which holds for all $\varepsilon > 4{\rm pen}(n; N)$, and

$$\text{(1.4)} \qquad \mathbb{E}L(\hat{C}) - L_0 \leq \inf_{N \geq 1}\Bigl[\inf_{C \in \mathcal{C}_N} L(C) - L_0 + \sqrt{\frac{16 V(\mathcal{C}_N)\log n + 8(N + 11)}{n}}\Bigr],$$

where $L_0 := \inf_{N\geq 1}\inf_{C \in \mathcal{C}_N} L(C)$.

Given a class $\mathcal{C}$ of decision rules and a number $L_0 \in (0, 1/2)$, let $\mathcal{P}(\mathcal{C}; L_0)$ be the set of all distributions of (X, Y) such that $L(C) \geq L_0$ for all $C \in \mathcal{C}$. Suppose that $V(\mathcal{C}) \geq 2$. Devroye, Györfi and Lugosi (1996) gave a minimax lower bound for the risk of arbitrary empirical decision rule, based on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ (see their Theorem 14.5). Namely, for any such a decision rule $\tilde{C}$, there exists a distribution of training examples from the set $\mathcal{P}(\mathcal{C}; L_0)$ such that

$$\text{(1.5)} \qquad \mathbb{E}L(\tilde{C}) - L_0 \geq e^{-8}\sqrt{\frac{L_0(V(\mathcal{C}) - 1)}{24n}}$$

for all $n \geq (2L_0)^{-1}((1 - 2L_0)^{-2} \vee 9)(V(\mathcal{C}) - 1)$.

Let $\{\mathcal{C}_N\}$ be an increasing sequence of VC-classes such that $\{V(\mathcal{C}_N)\}$ is strictly increasing and for some constant $D > 0$ $V(\mathcal{C}_{N+1}) \leq DV(\mathcal{C}_N),\ N \geq 1$. Let $\{\delta_N\}$ be a sequence such that $\delta_N \downarrow 0$. Let $\mathcal{P} := \mathcal{P}(\{\mathcal{C}_N\}; \{\delta_N\}; L_0)$ be the class of all distributions of (X, Y) such that $0 \leq \inf_{C \in \mathcal{C}_N} L(C) - L_0 \leq \delta_N,\ N \geq 1$. It follows from (1.4) and (1.5) that with some constants $A, B > 0$

$$\text{(1.6)} \qquad \sup_{\mathcal{P}} \mathbb{E}L(\hat{C}) - L_0 \leq A \inf_{N \geq 1}\Bigl[\delta_N + \sqrt{\frac{V(\mathcal{C}_N)\log n}{n}}\Bigr]$$

and

$$\inf_{\tilde{C}} \sup_{\mathcal{P}} \mathbb{E}L(\tilde{C}) - L_0 \geq B \inf_{N\geq 1}\Bigl[\delta_N + \sqrt{\frac{V(\mathcal{C}_N)}{n}}\Bigr]. \tag{1.7}$$

Thus the estimator $\hat{C}$, obtained using the structural risk minimization approach, is optimal in the minimax sense up to a logarithmic factor and up to constants.

A natural measure of complexity of the class $\mathcal{C}$ of decision rules in the problems of empirical risk minimization is the accuracy of empirical approximation on the class $\mathcal{C}$, defined by $\|L_n - L\|_{\mathcal{C}} := \sup_{C\in\mathcal{C}}\bigl|L_n(C) - L(C)\bigr|$, or as the expectation of this quantity. The bound (1.2) is uniform with respect to all the distributions of (X, Y) and therefore it does not have to be optimal for a particular distribution. Also, the constants in this bound are not best possible and the VC-dimension of the class $\mathcal{C}$ of decision rules is often unknown and has to be replaced by its upper bound (this is the case, for instance, for some classes of neural networks). This hierarchy of non-optimal upper bounds leads to the fact that the penalty function pen(n, N), defined by (1.1), is often much larger than the "ideal" penalty $\mathbb{E}\|L_n - L\|_{\mathcal{C}_N}$. The "ideal" penalty, however, can not be used in practice since the distribution of (X, Y) is unknown. Therefore, rather conservative upper bounds, described above, are to be used instead.

In the recent literature on nonparametric estimation, an approach quite similar to the structural risk minimization is often referred to as *the method of sieves.* Birgé and Massart (1996), Barron, Birgé and Massart (1999) have studied rather thoroughly the penalty functions to be used in the problems of adaptive estimation on sieves. They used powerful Talagrand's concentration and deviation inequalities for empirical processes (Talagrand (1996a,b), Ledoux (1996), Massart (1999)) to obtain the so called *oracle inequalities* for the theoretical risk of their estimators. The method of oracle inequalities has become a rather popular way to prove optimality properties of nonparametric statistical estimators (see Johnstone (1998)). The Birgé-Massart penalties are also based on the dimensions of the classes of functions (metric entropy dimensions or VC-type dimensions). Their approach works rather well in some examples of sieves that frequently occur in the problems of nonparametric regression and density estimation (for example, for nested families of Sobolev ellipsoids). In such cases, the Birgé-Massart penalties provide rather sharp upper bounds for the accuracy of empirical approximation. This is not always the case, however, in the problems of concept learning. In these problems, the dimension based penalties often overestimate the value of $\mathbb{E}\|L_n - L\|_{\mathcal{C}}$, which imposes unnecessary restrictions on the complexity of the classes of decision rules and results in prohibitively large sample sizes required to guarantee a reasonable accuracy of learning.

In this paper, we suggest a data based penalty, defined by $\rho(n; N) := R_n(\mathcal{C}_N)$, where

$$R_n(\mathcal{C}) := \sup_{C\in\mathcal{C}}\Bigl|n^{-1}\sum_{j=1}^{n} r_j I_{\{Y_j \neq I_C(X_j)\}}\Bigr| \tag{1.8}$$

$\{r_n\}_{n\geq 1}$ being a Rademacher sequence (i.e. a sequence of independent random variables taking values $+1$ and -1 with probability $1/2$ each), independent of $\{(X_n, Y_n)\}$. We call such a penalty *the Rademacher penalty.* The quantities similar to $R_n(\mathcal{C})$ have been frequently used in the so called symmetrization inequalities for empirical processes (see Lemma 2.4 below). The method of Rademacher symmetrization, known in many areas of Analysis and Probability, was brought to the empirical processes theory by Koltchinskii (1981), Pollard (1982), and, especially, Giné and Zinn (1984). It allowed them to simplify substantially the proofs of the original Vapnik and Chervonenkis (1971, 1974) results and to develop the techniques of uniform bounds for empirical processes to the level they could be used to prove uniform versions of the central limit theorem (see Dudley (1999) and van der Vaart and Wellner (1996) for thorough account of these developments). Despite the theoretical importance of the Rademacher symmetrization, its use as a tool of statistical inference has been rather limited. Using $R_n(\mathcal{C})$ as a (computable) measure of the accuracy of empirical approximation on the class $\mathcal{C}$ is actually a special case of the so called weighted bootstrap (see van der Vaart and Wellner (1996)). Recently, Koltchinskii, Abdallah, Ariola, Dorato and Panchenko (1999) used similar quantities in statistical learning problems that occur in control theory.

It is easy to check that computing the Rademacher penalty is equivalent to the solution of empirical risk minimization problem for "randomly relabeled" sample. Indeed, we have

$$R_n(\mathcal{C}) = \sup_{C\in\mathcal{C}}\Bigl[n^{-1}\sum_{j=1}^{n} r_j I_{\{Y_j\neq I_C(X_j)\}}\Bigr] \bigvee \Bigl(-\inf_{C\in\mathcal{C}}\Bigl[n^{-1}\sum_{j=1}^{n} r_j I_{\{Y_j\neq I_C(X_j)\}}\Bigr]\Bigr),$$

so, it is enough to compute separately the supremum and the infimum above. Let us consider, for instance, the supremum. We have

$$\sum_{j=1}^{n} r_j I_{\{Y_j\neq I_C(X_j)\}} = \sum_{r_j=+1,Y_j=1}(1-I_C(X_j)) + \sum_{r_j=+1,Y_j=0} I_C(X_j) -$$

$$-\sum_{r_j=-1,Y_j=1}(1-I_C(X_j)) - \sum_{r_j=-1,Y_j=0} I_C(X_j) = \sum_{j=1,\dots n:Y_j=1} r_j + \sum_{j=1}^{n}\sigma_j I_C(X_j),$$

where $\sigma_j := -(2Y_j-1)r_j$. Thus, maximizing $\sum_{j=1}^{n} r_j I_{\{Y_j\neq I_C(X_j)\}}$ over $C\in\mathcal{C}$ is equivalent to maximizing $\sum_{j=1}^{n}\sigma_j I_C(X_j)$. Next, we have

$$\sum_{j=1}^{n}\sigma_j I_C(X_j) = \sum_{\sigma_j=+1} I_C(X_j) - \sum_{\sigma_j=-1} I_C(X_j) =$$

$$= -\sum_{\sigma_j=-1} I_C(X_j) - \sum_{\sigma_j=+1}(1-I_C(X_j)) + \mathrm{card}\{j=1,\dots,n:\sigma_j=+1\}.$$

Hence, the problem can be reduced to minimizing

$$\sum_{\sigma_j=-1} I_C(X_j) + \sum_{\sigma_j=+1}(1-I_C(X_j)) = \sum_{j=1}^{n} I_{\{\tilde{Y}_j\neq I_C(X_j)\}},$$

where $\tilde{Y}_j = 0$ iff $\sigma_j=-1$, and $\tilde{Y}_j=1$ otherwise. The above argument also shows that $R_n(\mathcal{C})$ can be viewed as a measure of "separation power" of the class $\mathcal{C}$ of decision rules. Indeed, if the value of $R_n(\mathcal{C})$ is large, the class of decision rules $\mathcal{C}$ would separate the "positive" examples from the "negative" ones with a small error even if the labels were assigned at random. This indicates that the class $\mathcal{C}$ is too large (a reasonable class of decision rules should separate the positive examples from the negative ones in the case of correct labels, but should not do this when the labels are randomly misplaced).

In the next section, we describe more general version of structural minimization of empirical risk with Rademacher penalties. This version also applies to the problems of function learning and regression. We prove probabilistic oracle inequalities (of the same type as (1.3), (1.4)) that give upper bounds for the (theoretical) risk of the functions that approximately minimize the penalized empirical risk. The inequalities show some form of optimality of the procedure of structural risk minimization with Rademacher penalties. In a special case of the sieve formed by VC-classes of sets (concepts), the decision rule, obtained by the method of structural risk minimization with Rademacher penalties, has the optimal value of risk (up to a multiplicative constant).

One of the problems with the implementation of the method of Rademacher penalization is the necessity to compute the penalties, which, as we have shown above, is equivalent to solving precisely the problem of minimization of the empirical risk for randomly relabeled data. In many cases only approximate solution of this problem is available and the accuracy of approximation is not known precisely. We consider in the last two sections a possible way to get around these difficulties. Namely, we develop in these sections a method of *iterative structural risk minimization* with Rademacher penalties. Instead of using the hierarchy of function classes given in advance, this method allows one to determine finite data dependent pools of functions in the data-driven process of empirical risk minimization. The Rademacher penalties are now computed by maximizing the Rademacher process over these finite pools of functions. This resembles the recent work of some other authors on developing more flexible data-driven versions of risk minimization (such as "simple empirical covering" of Buescher and Kumar (1996), "structural risk minimization over data-dependent hierarchies" of Shawe-Taylor, Bartlett, Williamson and Anthony (1996), "self-bounding learning" of Freund (1998)). It is also worth mentioning that popular boosting algorithms are, in fact, methods of iterative structural minimization of risk. They are known to produce classifiers of rather high complexity and with small classification error, seemingly overcoming the standard difficulties related to overfitting the data. This could be due to the fact that the right measure of complexity for such iterative risk minimization algorithms should be data dependent and based on the functions actually involved in the iteration process rather than on VC-dimensions of the huge classes of classifiers of which the actual iteration pool is only a small part. We prove probabilistic oracle inequalities, showing some form of optimality of iterative structural risk minimization.

2. Oracle inequalities for structural risk minimization with Rademacher penalties. Let $(S, \mathcal{A})$ be a measurable space and let $\{X_n\}_{n\geq 1}$ be a sequence of i.i.d. observations in this space with common distribution P. We assume that this sequence is defined on a probability space $(\Omega, \Sigma, \mathbb{P})$. Denote $\mathcal{P}(S) := \mathcal{P}(S, \mathcal{A})$ the set of all probability measures on $(S, \mathcal{A})$. Let P_n be the empirical measure based on the sample $(X_1, \dots, X_n)$:

$$P_n := n^{-1} \sum_{j=1}^{n} \delta_{X_j}, \text{ where } \delta_x(A) := \begin{cases} 1 & x \in A \\ 0 & \text{otherwise.} \end{cases}$$

Given a probability measure μ on $(S; \mathcal{A})$ (e.g. P or P_n) and a μ-integrable function f, we define $\mu(f) := \int_S f d\mu$, and in what follows we frequently identify μ with the mapping $f \mapsto \mu(f)$. Given a class $\mathcal{F}$ of measurable functions from $(S, \mathcal{A})$ into $[0, 1]$, we denote

$$\Delta_n(\mathcal{F}) := \|P_n - P\|_{\mathcal{F}} \text{ and } R_n(\mathcal{F}) := \|n^{-1} \sum_{j=1}^{n} r_j \delta_{X_j} \|_{\mathcal{F}}.$$

Here $\| \cdot \|_{\mathcal{F}}$ stands for the norm of the space $\ell^\infty(\mathcal{F})$ of all uniformly bounded real valued functions on $\mathcal{F}$: $\|Y\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |Y(f)|,\ Y : \mathcal{F} \mapsto \mathbb{R}$.

To avoid dealing with complicated measurability issues that frequently occur in the theory of empirical processes, we assume in what follows that the classes of functions we are working with are countable. However, all the results of the paper are true if this assumption is replaced by standard assumptions of empirical measurability of the classes, the probability measure $\mathbb{P}$ is replaced by outer probability, expectation $\mathbb{E}$ is replaced by outer expectation, etc. (see Dudley (1999) or van der Vaart and Wellner (1996) for the discussion of these issues).

Consider a family $\{\mathcal{F}_m : m \in \mathcal{M}\}$ of classes of measurable functions from $(S, \mathcal{A})$ into $[0, 1]$ (a sieve). The set $\mathcal{M}$ is supposed to be countable. We assume in what follows that for different classes in the sieve one can use different sample sizes. We denote these sample sizes $\{n_m :\ m \in \mathcal{M}\}$. Let $\{t_m :\ m \in \mathcal{M}\}$ be a set of positive real numbers. We define an "ideal" penalty function by

$$\mathcal{I}(m) := \mathcal{I}(m; \{\mathcal{F}_m, n_m, t_m : m \in \mathcal{M}\}) := 5\mathbb{E}\Delta_{n_m}(\mathcal{F}_m) + \frac{6t_m + 2}{\sqrt{n_m}} \tag{2.1}$$

and an empirical Rademacher penalty function by

$$\mathcal{E}(m) := \mathcal{E}(m; \{\mathcal{F}_m, n_m, t_m : m \in \mathcal{M}\}) := 2R_{n_m}(\mathcal{F}_m) + \frac{3t_m}{\sqrt{n_m}}. \tag{2.2}$$

Given $\delta > 0$, we define a random variable $\hat{m} \in \mathcal{M}$ and an estimate $\hat{f} := \hat{f}_\delta \in \mathcal{F}_{\hat{m}}$ ($\hat{m}$ and $\hat{f}$ depend on the data $\{X_j : j = 1, \dots, n_m\}_{m\in\mathcal{M}}$), such that

$$\inf_{m\in\mathcal{M}} \Big[\inf_{f\in\mathcal{F}_m} P_{n_m}(f) + \mathcal{E}(m) \Big] + \delta \geq P_{n_{\hat{m}}}(\hat{f}) + \mathcal{E}(\hat{m}). \tag{2.3}$$

In the setting of Section 1, the space S is to be replaced by $S \times \{0,1\}$. The sieve in this case is the family $\{\mathcal{F}_N : N \geq 1\}$, where

$$\mathcal{F}_N := \Big\{ f_C : C \in \mathcal{C}_N \Big\},\ N \geq 1,\ f_C(x,y) := I_{\{y \neq I_C(x)\}},\ x \in S, y \in \{0,1\}.$$

2.1. Theorem. *The following inequalities hold:*

$$\sup_{P \in \mathcal{P}(S)} \mathbb{P}\Big\{ P(\hat{f}) \geq \inf_{m \in \mathcal{M}} \big[\inf_{f \in \mathcal{F}_m} P_{n_m}(f) + \mathcal{E}(m) \big] + \delta \Big\} \leq \sum_{m \in \mathcal{M}} \exp\{-\frac{2}{3} t_m^2\} \tag{2.4}$$

and

$$\sup_{P \in \mathcal{P}(S)} \mathbb{P}\Big\{ P(\hat{f}) \geq \inf_{m \in \mathcal{M}} \big[\inf_{f \in \mathcal{F}_m} P(f) + \mathcal{I}(m) \big] + \delta \Big\} \leq 2 \sum_{m \in \mathcal{M}} \exp\{-\frac{2}{3} t_m^2\}. \tag{2.5}$$

The proof uses a well known exponential inequality for martingale difference sequences (see, e.g., Ledoux and Talagrand (1991), Lemma 1.5, or Devroye, Györfi and Lugosi (1996), Theorem 9.1). This inequality is due to Azuma (1967). Yurinski (1974) suggested a martingale representation of the norms of sums of independent random vectors and opened a way to use this type of inequalities in Probability in Banach spaces. They also found a number of applications in the local theory of Banach spaces (Milman and Schechtman (1986)). Koltchinskii (1985, 1986) applied these inequalities to empirical processes and random entropies, Rhee and Talagrand (1987) used them in their study of NP-complete problems, McDiarmid (1989) considered a broad range of applications in graph theory and combinatorics, Devroye, Györfi and Lugosi (1996) used them in the problems of pattern recognition. Talagrand (1996a, 1996b) has developed much more powerful and sophisticated technique of concentration inequalities that have already been applied to the problems of adaptive nonparametric estimation (Barron, Birgé and Massart (1999)).

More specifically, the following corollary of Azuma's inequality will be used (see Devroye, Györfi and Lugosi (1996), Theorem 9.2). Let $(A, \mathcal{A})$ be a measurable space and let g be a measurable function from A^n into $\mathbb{R}$, such that with some constants $c_i > 0,\ i = 1, \dots, n$

$$|g(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - g(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \leq c_i,$$

for all $x_1, \dots, x_{i-1}, x_i, x_i', x_{i+1}, \dots, x_n \in A,\ i = 1, \dots, n$. Let $Y_1, \dots, Y_n$ be independent random variables with values in $(A, \mathcal{A})$. Then for all $\varepsilon > 0$

$$\mathbb{P}\Big\{ g(Y_1, \dots, Y_n) - \mathbb{E} g(Y_1, \dots, Y_n) \geq \varepsilon \Big\} \leq \exp\{-\frac{2\varepsilon^2}{\sum_{j=1}^n c_j^2}\}$$

and

$$\mathbb{P}\Big\{ \mathbb{E} g(Y_1, \dots, Y_n) - g(Y_1, \dots Y_n) \geq \varepsilon \Big\} \leq \exp\{-\frac{2\varepsilon^2}{\sum_{j=1}^n c_j^2}\}.$$

These inequalities immediately imply the following lemmas.

2.2. Lemma. *For all* $\varepsilon > 0$,

$$\mathbb{P}\{\Delta_n(\mathcal{F}) \geq \mathbb{E}\Delta_n(\mathcal{F}) + \varepsilon\} \leq \exp\{-2\varepsilon^2 n\}$$

and

$$\mathbb{P}\{\mathbb{E}\Delta_n(\mathcal{F}) \geq \Delta_n(\mathcal{F}) + \varepsilon\} \leq \exp\{-2\varepsilon^2 n\}.$$

2.3. Lemma. *For all* $\varepsilon > 0$,

$$\mathbb{P}\{\mathbb{E} R_n(\mathcal{F}) \geq R_n(\mathcal{F}) + \varepsilon\} \leq \exp\{-\varepsilon^2 n/2\}$$

and

$$\mathbb{P}\{R_n(\mathcal{F}) \geq \mathbb{E} R_n(\mathcal{F}) + \varepsilon\} \leq \exp\{-\varepsilon^2 n/2\}.$$

2.4. Lemma. *For all* $\varepsilon > 0$,

$$\mathbb{P}\{\Delta_n(\mathcal{F}) - 2R_n(\mathcal{F}) \geq \mathbb{E}[\Delta_n(\mathcal{F}) - 2R_n(\mathcal{F})] + 3\varepsilon\} \leq \exp\{-18\varepsilon^2 n/25\} \leq \exp\{-\frac{2}{3}\varepsilon^2 n\}$$

and

$$\mathbb{P}\{\Delta_n(\mathcal{F}) + 2R_n(\mathcal{F}) \geq \mathbb{E}[\Delta_n(\mathcal{F}) + 2R_n(\mathcal{F})] + 3\varepsilon\} \leq \exp\{-18\varepsilon^2 n/25\} \leq \exp\{-\frac{2}{3}\varepsilon^2 n\}.$$

Note that inequalities similar to the ones of Lemma 2.4 can be also obtained by combining the bounds of Lemma 2.2 and Lemma 2.3, but this leads to the worse values of constants. [This improvements of the constants was suggested to the author by Don Hush and Clint Scovel, see also Hush and Scovel (1999)].

2.5. Lemma. *The following inequalities hold:*

$$\frac{1}{2}\mathbb{E}R_n(\mathcal{F}) - \frac{1}{2\sqrt{n}} \leq \frac{1}{2}\mathbb{E}\|n^{-1}\sum_{j=1}^{n} r_j(\delta_{X_j} - P)\|_{\mathcal{F}} \leq \mathbb{E}\Delta_n(\mathcal{F}) \leq 2\mathbb{E}R_n(\mathcal{F}).$$

Lemma 2.5 gives symmetrization inequalities for empirical processes. The proofs of the last two inequalities can be found, for instance, in van der Vaart and Wellner (1996). The proof of the first inequality is obvious:

$$\mathbb{E}R_n(\mathcal{F}) \leq \mathbb{E}\|n^{-1}\sum_{j=1}^{n} r_j(\delta_{X_j} - P)\|_{\mathcal{F}} + \mathbb{E}|n^{-1}\sum_{j=1}^{n} r_j| \leq$$

$$\leq \mathbb{E}\|n^{-1}\sum_{j=1}^{n} r_j(\delta_{X_j} - P)\|_{\mathcal{F}} + \mathbb{E}^{1/2}|n^{-1}\sum_{j=1}^{n} r_j|^2 = \mathbb{E}\|n^{-1}\sum_{j=1}^{n} r_j(\delta_{X_j} - P)\|_{\mathcal{F}} + \frac{1}{\sqrt{n}}.$$

Proof of Theorem 2.1. The following bound is obvious (since $\hat{f} \in \mathcal{F}_{\hat{m}}$):

$$P(\hat{f}) \leq P_{n_{\hat{m}}}(\hat{f}) + \Delta_{n_{\hat{m}}}(\mathcal{F}_{\hat{m}}), \tag{2.6}$$

and using the first bound of Lemma 2.4 and the inequality $\mathbb{E}[\Delta(\mathcal{F}_m) - 2R_n(\mathcal{F}_m)] \leq 0$ (Lemma 2.5) we can write

$$\mathbb{P}\Big(\bigcup_{m\in\mathcal{M}}\Big\{\Delta_{n_m}(\mathcal{F}_m) \geq 2R_{n_m}(\mathcal{F}_m) + 3t_m n_m^{-1/2}\Big\}\Big) \leq \sum_{m\in\mathcal{M}}\exp\{-\frac{2}{3}t_m^2\}. \tag{2.7}$$

This implies that with probability at least $1 - \sum_{m\in\mathcal{M}}\exp\{-\frac{2}{3}t_m^2\}$, we have (by the definition of $\hat{f}, \hat{m}$)

$$P(\hat{f}) \leq P_{n_{\hat{m}}}(\hat{f}) + 2R_{n_{\hat{m}}}(\mathcal{F}_{\hat{m}}) + 3t_m n_m^{-1/2} =$$

$$= P_{n_{\hat{m}}}(\hat{f}) + \mathcal{E}(\hat{m}) \leq \inf_{m\in\mathcal{M}}\Big[\inf_{f\in\mathcal{F}_m} P_{n_m}(f) + \mathcal{E}(m)] + \delta, \tag{2.8}$$

and the inequality (2.4) follows.

To prove (2.5), note that, by the second bound of Lemma 2.4, we get

$$\mathbb{P}\Big(\bigcup_{m\in\mathcal{M}}\{\Delta_n(\mathcal{F}_m) + 2R_n(\mathcal{F}_m) \geq \mathbb{E}[\Delta_n(\mathcal{F}_m) + 2R_n(\mathcal{F}_m)] + 3t_m n_m^{-1/2}\}\Big) \leq \sum_{m\in\mathcal{M}}\exp\{-\frac{2}{3}t_m^2\}.$$

We also have (using the bounds of Lemma 2.5)

$$\mathbb{E}R_{n_m}(\mathcal{F}_m) \leq \mathbb{E}\|n_m^{-1}\sum_{j=1}^{n_m} r_j(\delta_{X_j} - P)\|_{\mathcal{F}_m} + n_m^{-1/2} \leq 2\mathbb{E}\Delta_{n_m}(\mathcal{F}_m) + n_m^{-1/2}. \tag{2.9}$$

Therefore

$$\mathbb{P}\Big(\bigcup_{m\in\mathcal{M}}\{\Delta_n(\mathcal{F}_m)+2R_n(\mathcal{F}_m)\geq 5\mathbb{E}\Delta_n(\mathcal{F}_m)+3t_m n_m^{-1/2}+2n_m^{-1/2}\}\Big)\leq$$

$$\leq\sum_{m\in\mathcal{M}}\exp\{-\frac{2}{3}t_m^2\}. \tag{2.10}$$

Using (2.8)–(2.10), we conclude that with probability at least $1-2\sum_{m\in\mathcal{M}}\exp\{-\frac{2}{3}t_m^2\}$, we have

$$P(\hat{f})\leq\inf_{m\in\mathcal{M}}\Big[\inf_{f\in\mathcal{F}_m}P_{n_m}(f)+\mathcal{E}(m)\Big]+\delta\leq$$

$$\leq\inf_{m\in\mathcal{M}}\Big[\inf_{f\in\mathcal{F}_m}P(f)+\Delta_{n_m}(\mathcal{F}_m)+2R_{n_m}(\mathcal{F}_m)+3t_m n_m^{-1/2}\Big]+\delta\leq$$

$$\leq\inf_{m\in\mathcal{M}}\Big[\inf_{f\in\mathcal{F}_m}P(f)+5\mathbb{E}\Delta_{n_m}(\mathcal{F}_m)+2n_m^{-1/2}+6t_m n_m^{-1/2}\Big]+\delta=$$

$$=\inf_{m\in\mathcal{M}}\Big[\inf_{f\in\mathcal{F}_m}P(f)+\mathcal{J}(m)\Big]+\delta,$$

which completes the proof.

□

Let $\Phi(x):=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-u^2/2}du$.

2.6. Corollary. *The following inequality holds for all $P\in\mathcal{P}(S)$:*

$$\mathbb{E}P(\hat{f})\leq\inf_{m\in\mathcal{M}}\big[\inf_{f\in\mathcal{F}_m}P(f)+\mathcal{I}(m)\big]+\delta+6\sqrt{6\pi}\sum_{m\in\mathcal{M}}\frac{1}{\sqrt{n_m}}\big[1-\Phi(\frac{2}{\sqrt{3}}t_m)\big]. \tag{2.11}$$

Proof. Given $\varepsilon>0$, let us replace t_m by $t_m':=t_m+\varepsilon n_m^{1/2}$. Then $\mathcal{I}'(m)=\mathcal{I}(m)+6\varepsilon$ and $\mathcal{E}'(m)=\mathcal{E}(m)+3\varepsilon$. The estimates $\hat{m}$ and $\hat{f}$ remain unchanged. In this case it follows from the inequalities (2.4) and (2.5) that

$$\sup_{P\in\mathcal{P}(S)}\mathbb{P}\Big\{P(\hat{f})\geq\inf_{m\in\mathcal{M}}\big[\inf_{f\in\mathcal{F}_m}P_{n_m}(f)+\mathcal{E}(m)\big]+\delta+3\varepsilon\Big\}\leq\sum_{m\in\mathcal{M}}\exp\{-\frac{2}{3}(t_m+\varepsilon n_m^{1/2})^2\} \tag{2.4′}$$

and

$$\sup_{P\in\mathcal{P}(S)}\mathbb{P}\Big\{P(\hat{f})\geq\inf_{m\in\mathcal{M}}\big[\inf_{f\in\mathcal{F}_m}P(f)+\mathcal{I}(m)\big]+\delta+6\varepsilon\Big\}\leq 2\sum_{m\in\mathcal{M}}\exp\{-\frac{2}{3}(t_m+\varepsilon n_m^{1/2})^2\}. \tag{2.5′}$$

Define

$$\xi:=\Big(P(\hat{f})-\inf_{m\in\mathcal{M}}\big[\inf_{f\in\mathcal{F}_m}P(f)+\mathcal{I}(m)\big]-\delta\Big)/6.$$

It follows from (2.5’) that for all $\varepsilon>0$

$$\mathbb{P}\{\xi^+\geq\varepsilon\}=\mathbb{P}\{\xi\geq\varepsilon\}\leq 2\sum_{m\in\mathcal{M}}\exp\{-\frac{2}{3}(t_m+\varepsilon n_m^{1/2})^2\}.$$

Integrating with respect to ε from 0 to $+\infty$ gives:

$$\mathbb{E}\xi\leq\mathbb{E}\xi^+=\int_0^{+\infty}\mathbb{P}\{\xi^+\geq\varepsilon\}d\varepsilon\leq 2\sum_{m\in\mathcal{M}}\int_0^{+\infty}\exp\{-\frac{2}{3}(t_m+\varepsilon n_m^{1/2})^2\}d\varepsilon=$$

$$= 2\sum_{m\in\mathcal{M}} \frac{1}{\sqrt{n_m}}\int_0^{+\infty} \exp\{-\frac{2}{3}(t_m+v)^2\}dv = \sqrt{6\pi}\sum_{m\in\mathcal{M}}\frac{1}{\sqrt{n_m}}\bigl[1-\Phi(\frac{2}{\sqrt{3}}t_m)\bigr],$$

and (2.11) easily follows.

□

In particular, assume that $n_m \equiv n$ and let $C := 2\sum_{m\in\mathcal{M}}\exp\{-\frac{2}{3}t_m^2\} < +\infty$. Then we have

$$\mathcal{I}(m) := \mathcal{I}(m;n) := 5\mathbb{E}\Delta_n(\mathcal{F}_m) + \frac{6t_m+2}{\sqrt{n}}.$$

Theorem 2.1 and Corollary 2.6 imply that

$$\sup_{P\in\mathcal{P}(S)}\mathbb{P}\Bigl\{P(\hat{f}) \geq \inf_{m\in\mathcal{M}}\bigl[\inf_{f\in\mathcal{F}_m}P(f)+\mathcal{I}(m)\bigr]+\delta+6\varepsilon\Bigr\} \leq C\exp\{-\frac{2}{3}\varepsilon^2 n\}$$

and

$$\mathbb{E}P(\hat{f}) \leq \inf_{m\in\mathcal{M}}\bigl[\inf_{f\in\mathcal{F}_m}P(f)+\mathcal{I}(m)\bigr]+\delta+\frac{(3/2)\sqrt{6\pi}C}{\sqrt{n}}.$$

To be more specific, assume that $\mathcal{M} := \mathbb{N}$ and take $t_m := \gamma(\log m)^{1/2}$ with $\gamma > \sqrt{\frac{3}{2}}$. Then $C := C_\gamma := 2\sum_{m\geq 1} m^{-\frac{2}{3}\gamma^2}$. If, in addition, $\mathbb{E}\Delta_n(\mathcal{F}_m) \leq \frac{D_m(P)}{\sqrt{n}}$ with some $D_m(P) > 0$ (this holds, for instance, if $\mathcal{F}_m$ is a P-Donsker class for all $m \geq 1$), then we have

$$\sup_{P\in\mathcal{P}(S)}\mathbb{P}\Bigl\{P(\hat{f}) \geq \inf_{m\in\mathbb{N}}\bigl[\inf_{f\in\mathcal{F}_m}P(f)+\frac{5D_m(P)+6\gamma\sqrt{\log m}+2}{\sqrt{n}}\bigr]+\delta+6\varepsilon\Bigr\} \leq$$

$$\leq C_\gamma \exp\{-\frac{2}{3}\varepsilon^2 n\} \tag{2.12}$$

and

$$\mathbb{E}P(\hat{f}) \leq \inf_{m\in\mathbb{N}}\bigl[\inf_{f\in\mathcal{F}_m}P(f)+\frac{5D_m(P)+6\gamma\sqrt{\log m}+2}{\sqrt{n}}\bigr]+\delta+\frac{(3/2)\sqrt{6\pi}C_\gamma}{\sqrt{n}}. \tag{2.13}$$

The meaning of these oracle inequalities can be described as follows. Suppose there exists an oracle who knows the distribution P of our data and who can compute any quantity related to this distribution. Then we can ask the oracle to tell us the values of the quantities $D_m(P)$ as well as the quantities $\delta_m(P) := \inf_{f\in\mathcal{F}_m}P(f) - \inf_{m\in\mathbb{N}}\inf_{f\in\mathcal{F}_m}P(f)$, that characterize the approximation error of the minimal risk on the class $\mathcal{F}_m$. Since $\frac{D_m(P)}{\sqrt{n}}$ characterizes the accuracy of empirical approximation on the class $\mathcal{F}_m$, it can be used as a complexity penalty. With such a penalty, a reasonable choice of m is

$$\tilde{m} := \mathrm{argmin}\bigl[\delta_m(P)+\frac{D_m(P)}{\sqrt{n}}\bigr].$$

Thus, one can try to estimate the minimizer of the risk P by minimizing the empirical risk P_n on the class $\mathcal{F}_{\tilde{m}}$. Suppose for simplicity that $\delta = 0$. Then the oracle inequlities above tell us that if the sample size n is large enough, namely, $n > \frac{3}{2}\frac{1}{\varepsilon^2}\log\frac{C_\gamma}{\alpha}$, then for all $P \in \mathcal{P}(S)$ with probability at least $1-\alpha$

$$P(\hat{f}) - \inf_{m\in\mathbb{N}}\inf_{f\in\mathcal{F}_m}P(f) < \inf_{m\in\mathbb{N}}\bigl[\delta_m(P)+\frac{5D_m(P)+6\gamma\sqrt{\log m}+2}{\sqrt{n}}\bigr]+6\varepsilon.$$

Moreover, for all n and all $P \in \mathcal{P}(S)$

$$\mathbb{E}\bigl[P(\hat{f}) - \inf_{m\in\mathbb{N}}\inf_{f\in\mathcal{F}_m}P(f)\bigr] \leq \inf_{m\in\mathbb{N}}\bigl[\delta_m(P)+\frac{5D_m(P)+6\gamma\sqrt{\log m}+2}{\sqrt{n}}\bigr]+\frac{(3/2)\sqrt{6\pi}C_\gamma}{\sqrt{n}}.$$

Thus, using Rademacher penalization allows us to obtain the solution of empirical risk minimization problem that is almost as good (up to constants and a couple of relatively small extra terms) as the one suggested by the oracle.

Being even more specific, one can assume that for each m the class $\mathcal{F}_m := \{I_C : C \in \mathcal{C}_m\}$, where $\mathcal{C}_m$ is a VC-class of sets. In this case, using well known bounds for the expectation of the sup-norm of empirical process and the bounds on uniform entropies of VC-classes (see, e.g., van der Vaart and Wellner (1996), Theorem 2.6.4), one can easily prove that for all $P \in \mathcal{P}(S)$ we can choose $D_m(P) \leq D\sqrt{V(\mathcal{C}_m)}$ with some numerical constant $D > 0$. This allows us to conclude that in the context of the classification problem discussed in section 1 (see (1.6), (1.7)), we have

$$\inf_{\tilde{C}} \sup_{\mathcal{P}} \mathbb{E}L(\tilde{C}) - L_0 \asymp \inf_{N \geq 1} \Big[\delta_N + \sqrt{\frac{V(\mathcal{C}_N)}{n}}\Big]$$

and the best possible (in the minimax sense and up to a constant) asymptotic rate of convergence is attained for the decision rule obtained via structural risk minimization with Rademacher penalties.

It is also worth mentioning that the upper bounds (2.12) and (2.13) depend on the distribution P and, for a particular distribution, they can be much sharper than the "worst case" bounds, depending on the VC-dimensions. On the other hand, these bounds can not be used in practice since the distribution P is unknown. The inequality (2.4) (see also (2.4')) provides a complementary data dependent upper bound on the theoretical risk that can be used instead (and which is sharper than the distribution dependent bound, given by (2.5), as it follows from the proofs).

3. Iterative structural risk minimization with Rademacher penalties. In this section, we consider an abstract iterative procedure of empirical risk minimization. We use Rademacher penalties in this procedure and obtain probabilistic bounds for the theoretical risk of the empirical risk minimizer. Our approach has some similarities with recent work of Freund (1999) on self bounding versions of local search minimization of empirical risk. Instead of using the sieve of function classes given in advance, as it is common in the traditional approach to structural risk minimization and as we did in the previous section, we construct here iteratively two nondecreasing sequences of finite pools of functions: the inner pools $\{\hat{\mathcal{F}}_k^-\}$, that are used to minimize the empirical risk P_n, and the outer pools $\{\hat{\mathcal{F}}_k^+\}$, that are used to compute the Rademacher penalties $R_n(\hat{\mathcal{F}}_k^+)$. In addition to these two data dependent pools, we construct recursively three other nondecreasing sequences of finite pools of functions, $\{\mathcal{F}_k^-\}, \{\mathcal{F}_k\}, \{\mathcal{F}_k^+\}$. These three pools are related to the minimization of the theoretical risk and they depend on the unknown distribution P. The construction of the pools is based on the notion of *extension operator,* that allows one, given a finite path through the space of functions, to extend this path by adding a finite number of new functions, that are used in the process of risk minimization. The extension operator is the main ingredient of our method and its choice would be crucial for designing specific learning algorithms using the method. The pools are constructed in such a way that the inclusions $\mathcal{F}_k^- \subset \hat{\mathcal{F}}_k^- \subset \mathcal{F}_k \subset \hat{\mathcal{F}}_k^+ \subset \mathcal{F}_k^+$ hold for all k with high probability. We obtain an explicit bound for this probability, which enables us to prove the oracle inequalities for iterative structural risk minimizers.

To define things precisely, we need some elementary notions of graph theory. Let $\mathcal{V}$ and $\mathcal{L}$ be sets. $\mathcal{V}$ is supposed to be countable. The elements of $\mathcal{V}$ will be used as vertices of the graphs below and the elements of $\mathcal{L}$ will be used as labels assigned to the vertices. For a graph $\mathcal{G}$, $V(\mathcal{G})$ denotes the set of all vertices of $\mathcal{G}$ and $E(\mathcal{G})$ is the set of all edges. A tree is a connected graph with no cycles. A rooted tree is a tree with a fixed vertex (the root). Given a tree $\mathcal{G}$, an $\mathcal{L}$-labeling of $\mathcal{G}$ is a mapping $L : V(\mathcal{G}) \mapsto \mathcal{L}$. A couple $(\mathcal{G}, L)$ is called an $\mathcal{L}$-labeled tree. Let $v_0 \in V(\mathcal{G})$ be the root of the rooted tree $\mathcal{G}$. We denote $V_0(\mathcal{G}) = \{v_0\}$, $V_1(\mathcal{G})$ the set of all vertices of $\mathcal{G}$ adjacent to $v_0, \dots, V_k(\mathcal{G})$ the set of all vertices of $\mathcal{G}$ connected to V_0 with a path of length k. The number $h(\mathcal{G}) := \max\{k : V_k(\mathcal{G}) \neq \emptyset\}$ will be called the height of the rooted tree $\mathcal{G}$. We denote $T(\mathcal{G})$ the set of all terminal vertices of $\mathcal{G}$ (the vertices of degree 1 that are not equal to v_0). We set $T_k(\mathcal{G}) := T(\mathcal{G}) \cap V_k(\mathcal{G})$. Clearly, $T_{h(\mathcal{G})}(\mathcal{G}) = V_{h(\mathcal{G})}(\mathcal{G})$. We call the vertices in this set alive and the rest of the terminal vertices dead. The set $A(\mathcal{G})$ of all alive vertices is equal to $T_{h(\mathcal{G})}(\mathcal{G})$ and the set $D(\mathcal{G})$ of all dead vertices is equal to $T(\mathcal{G}) \setminus A(\mathcal{G})$. For an $\mathcal{L}$-labeled tree $\mathcal{T} = (\mathcal{G}; L)$, we use the notations $V(\mathcal{T}), V_k(\mathcal{T}), T(\mathcal{T}), T_k(\mathcal{T}), \dots$ that have similar meaning. For a rooted tree $\mathcal{G}$ and $v \in V(\mathcal{G})$, we denote by $\mathcal{G}(v)$ the subtree of $\mathcal{G}$ rooted at the vertex v. For an $\mathcal{L}$-labeled tree $\mathcal{T}$, we use similarly the notation $\mathcal{T}(v)$.

Next we define recursively an ordering on the set of all rooted $\mathcal{L}$-labeled trees as follows. Given $\mathcal{L}$-labeled trees $\mathcal{T}_1, \mathcal{T}_2$, we write $\mathcal{T}_1 \prec \mathcal{T}_2$, iff

(i) the roots of $\mathcal{T}_1$ and $\mathcal{T}_2$ have the same labels;

(ii) $\mathrm{card}(V_1(\mathcal{T}_1)) \leq \mathrm{card}(V_1(\mathcal{T}_2))$ and, moreover, there exists a one-to-one mapping φ from $V_1(\mathcal{T}_1)$ onto $V \subset V_1(\mathcal{T}_2)$ that preserves the labels;

(iii) for any $v \in V_1(\mathcal{T}_1)$, we have $\mathcal{T}_1(v) \prec \mathcal{T}_2(\varphi(v))$.

If $\mathcal{T}_1 \prec \mathcal{T}_2$ and $\mathcal{T}_2 \prec \mathcal{T}_1$, we say that the rooted labeled trees $\mathcal{T}_1$ and $\mathcal{T}_2$ are isomorphic and write $\mathcal{T}_1 \simeq \mathcal{T}_2$.

Next we define *an extension operator* $\mathcal{E}$ on the set of all $\mathcal{L}$-labeled rooted trees. Let $\mathrm{Fin}(\mathcal{L})$ be the class of all finite subsets of $\mathcal{L}$. For all $k \geq 1$, define a mapping $\mathcal{E}_k : \mathcal{L}^k \mapsto \mathrm{Fin}(\mathcal{L})$. Suppose that $\pi = ((v_0, l_0), \dots, (v_k, l_k))$, where $l_j := L(v_j)$, is a path through the labeled tree from the root to a terminal vertex v_k. Let $F := \mathcal{E}_{k+1}(l_0, \dots, l_k)$. Given π, $\mathcal{E}(\pi)$ is obtained by adding $\mathrm{card}(F)$ new vertices (from the set $\mathcal{V}$) to the tree, connecting them with edges to v_k and labeling them with different labels from F. In a special case $F = \emptyset$, the path does not have further extension. We denote $\mathcal{E}(\mathcal{T})$ the tree obtained from $\mathcal{T}$ by extending it in the described way along all the paths from the root to all the alive vertices of $\mathcal{T}$. Clearly, there might be many extentions $\mathcal{E}(\mathcal{T})$, but all of them are isomorphic rooted $\mathcal{L}$-labeled trees, so $\mathcal{E}(\mathcal{T})$ is well defined up to an isomorphism.

We give below some examples of the extension operators that can be used in minimization algorithms. To relate these examples to the risk minimization problems, one should think of the class of functions parametrized by the set $\mathcal{L} : \mathcal{F} := \{f(l, \cdot) : l \in \mathcal{L}\}$.

Example 3.1. Suppose that $\mathcal{L}$ is a countable set and for each $l \in \mathcal{L}$ there exists a finite neighbourhood $N(l) \subset \mathcal{L}$. For $k \geq 0$, define $\mathcal{E}_{k+1}(l_0, \dots, l_k) := N(l_k)$. This defines an extension operator $\mathcal{E}$, which will be called *a local search* operator (because of its connection to the local search minimization algorithm). Starting with a trivial labeled tree $\mathcal{T}_0$ with one labeled vertex (the root) (v_0, l_0), one can define recursively a sequence of labeled trees with root $v_0 : \mathcal{T}_k := \mathcal{E}(\mathcal{T}_{k-1}),\ k = 1, 2, \dots$.

Example 3.2. Assume now that $\mathcal{L} := \mathbb{Z}^d$ and that each point $l \in \mathbb{Z}^d$ is provided with a finite neighbourhood $N(l) \subset \mathbb{Z}^d$. We define $\mathcal{E}_1(l_0) := N(l_0)$ and for $k \geq 1$

$$\mathcal{E}_{k+1}(l_0, \dots, l_k) := \begin{cases} \{2l_k - l_{k-1}, l_k\} & \text{if } l_k \neq l_{k-1} \\ N(l_k) & \text{otherwise.} \end{cases}$$

This defines another extension operator, which will be called *an extension according to the previous pattern.*

Example 3.3. This is a more sophisticated version of the previous example. Suppose that $\mathcal{L} := \mathbb{R}^d$. Let $N \subset \mathbb{R}^d$, $N \ni 0$ be a finite set (a "neighbourhood" of 0). Let $\{\delta_j\}_{j \geq 1}$ be a sequence of positive numbers such that $\delta_j \downarrow 0$. The mappings $\mathcal{E}_k$ will be defined as follows. For $k = 1$, $\mathcal{E}_1(l_0) := l_0 + \delta_1 N$. Suppose that $k \geq 1$. We define $\mathcal{E}_{k+1}(l_0, \dots, l_k)$. If $l_k \neq l_{k-1}$, $\mathcal{E}_{k+1}(l_0, \dots, l_k) := \{2l_k - l_{k-1}, l_k\}$. Otherwise, if $l_k = l_{k-1} \neq l_{k-2}$, we set $\mathcal{E}_{k+1}(l_0, \dots, l_k) := l_k + \delta_1 N$. More generally, if $l_k = l_{k-1} = \dots = l_{k-j} \neq l_{k-j-1}$, we set $\mathcal{E}_{k+1}(l_0, \dots, l_k) := l_k + \delta_j N$. This defines an extension operator that is closely related to some well known minimization algorithms. Given a sequence $(l_0, \dots, l_k)$ of points from $\mathbb{R}^d$, we say that l_j is a double point in this sequence iff $l_j = l_{j-1}$. If the double point occurs in iterative minimization process, it means that the iterations got stuck at this point (for instance, close to a local minimum). The extension operator described above reduces in such a case the size of the iterative step, trying to approach the minimum closer.

Example 3.4. This example is related to some minimization techniques that are suitable for functions of many variables whose graph has the shape of a ravine (the minimization method goes back to Gelfand and Tsetlin). Namely, such an algorithm starts at any point and employs a version of steepest descent method to reach the bottom of the ravine. Then, another point is picked up (far enough from the first one) and the steepest descent method is used again to hit the bottom of the ravine. The two points at the bottom are connected with the segment of straight line and a point in the segment (far enough from the both ends) is picked up. The steepest descent to the bottom starts at this point. Thus, we have now three points at the bottom. The second one and the third one are used to repeat the iterations, and so on.

We define the corresponding extension operator as follows. Suppose that $\mathcal{L} := \mathbb{R}^d$ and that we are given a mapping $\mathbb{R}^d \ni l \mapsto l' \in \mathbb{R}^d$ (the point l' is "far enough" from the point l). Similarly to Example 3.3, we define a finite "neighbourhood" N of 0. As before, $\mathcal{E}_1(l_0) = l_0 + N$. If l_k and l_{k-1} are not double points (in the context of the minimization method described above, the double point occurs when the steepest descent stops and we are at the bottom of the ravine), define $\mathcal{E}_{k+1}(l_0, \dots, l_k) = \{2l_k - l_{k-1}, l_k\}$. If l_k is not a double point, but l_{k-1} is, set $\mathcal{E}_{k+1}(l_0, \dots, l_k) = l_k + N$. Otherwise, if l_k is the first double point, we set $\mathcal{E}_{k+1}(l_0, \dots, l_k) := \{l'_0\}$. Finally, if l_k is a double point and, for $j < k$, l_j is the last double point before l_k, we set $\tilde{l} := \gamma l_j + (1-\gamma) l_k$ ($\gamma \in (0,1)$ being a parameter of the algorithm) and define $\mathcal{E}_{k+1}(l_0, \dots, l_k) := \{\tilde{l}\}$.

Example 3.5. In this example, the extension operator can change the dimension (or the complexity) of the labels. The necessity to do this can occur, for instance, in neural networks learning (when one changes the complexity of the network in the process of learning), and also in such learning procedures as boosting. Assume that $\mathcal{L} := \bigcup_{j=1}^{\infty} \mathcal{L}_j$. Denote $c(l) := \inf\{j \geq 1 : l \in \mathcal{L}_j\}$, $l \in \mathcal{L}$ ($c(l)$ is the "complexity" of l). Suppose also that, for each $j \geq 1$, $\mathcal{E}^{(j)}$ is an extension operator on $\mathcal{L}_j$-labeled rooted trees. Let Γ be a mapping from $\mathcal{L}$ into $\mathcal{L}$ such that $c(\Gamma(l)) = c(l) + 1$, $l \in \mathcal{L}$ (for instance, if $\mathcal{L}_j = \mathbb{R}^j$ and $c(l)$ is equal to the dimension of l, one can define $\Gamma(l) = (l, 0)$). We define an extension operator $\mathcal{E}$ on $\mathcal{L}$-labeled rooted trees as follows. Given a sequence of points $(l_0, \dots, l_k)$ such that $l_k = l_{k-1}$, we set $\mathcal{E}_{k+1}(l_0, \dots, l_k) = \{\Gamma(l_k)\}$. Otherwise, if $l_k \neq l_{k-1}$, let $s \geq 1$ be the smallest integer such that $c(l_k) = c(l_{k-1}) = \ldots = c(l_{k-s+1}) = i \neq c(l_{k-s})$. We define $\mathcal{E}_{k+1}(l_0, \dots, l_k) = \mathcal{E}^{(i)}_s(l_{k-s+1}, \dots, l_k)$. In other words, as soon as the double point occurs in the sequence, we increase the complexity of the space. We are extending the graph in this space until the next double point occurs, and so on.

In addition to the extension operator, we need also *a trimming operator,* defined below. Given a rooted tree $\mathcal{G}$ and a vertex $v \in V(\mathcal{G})$, we denote by $\mathcal{C}(\mathcal{G}; v)$ (the trimming of $\mathcal{G}$ at the vertex v) the tree obtained from $\mathcal{G}$ by eliminating the subtree $\mathcal{G}(v)$ (along with the edge that connects it to the rest of the graph). If $V \subset V(\mathcal{G})$, $\mathcal{C}(\mathcal{G}; V)$ denotes the tree obtained from $\mathcal{G}$ by eliminating all the subtrees $\mathcal{G}(v)$, $v \in V$ (and the edges connecting the vertices $v \in V$ to the rest of the graph). Given a labeled tree $\mathcal{T}$, we use quite similarly the notations $\mathcal{C}(\mathcal{T}; v)$ and $\mathcal{C}(\mathcal{T}; V)$.

In what follows, a class $\mathcal{F}$ of measurable functions from $(S, \mathcal{A})$ into $[0,1]$ will be used as the label set $\mathcal{L}$, so we will deal with $\mathcal{F}$-labeled trees. An extension operator $\mathcal{E}$ on the set of all such trees is supposed to be given and fixed. We will use the notation $f_v := L(v)$.

Let $\delta > 0$ and let $\{t_k : k \geq 1\}$ be a nondecreasing sequence of nonnegative numbers. We define an $\mathcal{F}$-labeled tree $\mathcal{T}_0$ with one vertex, say, (v_0, f_0), set $\mathcal{F}_0 := \{f_v : v \in V(\mathcal{E}(\mathcal{T}_0))\}$, and then define recursively, for $k = 0, 1, 2, \dots$

$$\mathcal{T}_{k+1} := \begin{cases} \mathcal{C}(\mathcal{E}(\mathcal{T}_k); T_k)) & \text{if } \mathcal{E}(\mathcal{T}_k) \neq \mathcal{T}_k \\ \mathcal{T}_k & \text{otherwise,} \end{cases}$$

$$T_k := \{v \in A(\mathcal{E}(\mathcal{T}_k)) : P(f_v) \geq \min_{\mathcal{F}_k} P + \delta\},$$

$$\mathcal{F}_{k+1} := \mathcal{F}_k \cup \{f_v : v \in V(\mathcal{E}(\mathcal{T}_{k+1}))\},\ \tau_k := t_{h(\mathcal{T}_k)}.$$

In other words, the iterative process of risk minimization starts with the initial function f_0 (the label of the root of the tree). Then we use extension operator to create the pool of iterations $\mathcal{F}_0$ and to select from this pool the functions (labels) with too large value of risk. We trim the extended tree $\mathcal{E}(\mathcal{T}_0)$ to get rid of these functions (and the vertices they label) and we are getting the tree $\mathcal{T}_1$ as the result. Then the iterative process continues recursively.

Clearly, the trees $\mathcal{T}_k$ and the pools of iterations $\mathcal{F}_k$ can not be computed unless the distribution P is known precisely. Therefore, we define below the empirical versions of these objects. We set

$$\hat{\mathcal{T}}_0^+ := \mathcal{T}_0, \hat{\mathcal{T}}_0^- := \mathcal{T}_0,\ \hat{\mathcal{F}}_0^+ := \{f_v : v \in V(\mathcal{E}(\mathcal{T}_0))\}, \hat{\mathcal{F}}_0^- := \{f_v : v \in V(\mathcal{E}(\mathcal{T}_0))\},$$

and then define recursively

$$\hat{\mathcal{T}}_{k+1}^+ := \begin{cases} \mathcal{C}(\mathcal{E}(\hat{\mathcal{T}}_k^+); \hat{T}_k^+)) & \text{if } \mathcal{E}(\hat{\mathcal{T}}_k^+) \neq \hat{\mathcal{T}}_k^+ \\ \hat{\mathcal{T}}_k^+ & \text{otherwise,} \end{cases} \quad \hat{\mathcal{T}}_{k+1}^- := \begin{cases} \mathcal{C}(\mathcal{E}(\hat{\mathcal{T}}_k^-); \hat{T}_k^-)) & \text{if } \mathcal{E}(\hat{\mathcal{T}}_k^-) \neq \hat{\mathcal{T}}_k^- \\ \hat{\mathcal{T}}_k^- & \text{otherwise,} \end{cases}$$

$$\hat{T}_k^+ := \{v \in A(\mathcal{E}(\hat{\mathcal{T}}_k^+)) : P_n(f_v) \geq \min_{\hat{\mathcal{F}}_k^-} P_n + \delta + 4R_n(\hat{\mathcal{F}}_k^+) + \frac{6\hat{\tau}_k^+}{\sqrt{n}}\},$$

$$\hat{T}_k^- := \{v \in A(\mathcal{E}(\hat{\mathcal{T}}_k^-)) : P_n(f_v) \geq \min_{\hat{\mathcal{F}}_k^+} P_n + \delta - 4R_n(\hat{\mathcal{F}}_k^+) - \frac{6\hat{\tau}_k^+}{\sqrt{n}}\},$$

$$\hat{\mathcal{F}}_{k+1}^+ := \hat{\mathcal{F}}_k^+ \cup \{f_v : v \in V(\mathcal{E}(\hat{\mathcal{T}}_{k+1}^+))\},\ \hat{\mathcal{F}}_{k+1}^- := \hat{\mathcal{F}}_k^- \cup \{f_v : v \in V(\mathcal{E}(\hat{\mathcal{T}}_{k+1}^-))\},$$

$$\hat{\tau}_k^+ := t_{h(\hat{\mathcal{T}}_k^+)}, \hat{\tau}_k^- := t_{h(\hat{\mathcal{T}}_k^-)}.$$

These empirical rooted trees and the corresponding pools of functions (labels) can be computed recursively, given the empirical data.

Finally, we define

$$\mathcal{T}_0^+ := \mathcal{T}_0, \mathcal{T}_0^- := \mathcal{T}_0,$$

$$\mathcal{F}_0^+ := \{f_v : v \in V(\mathcal{E}(\mathcal{T}_0))\}, \mathcal{F}_0^- := \{f_v : v \in V(\mathcal{E}(\mathcal{T}_0))\},$$

$$\mathcal{T}_{k+1}^+ := \begin{cases} \mathcal{C}(\mathcal{E}(\mathcal{T}_k^+); T_k^+)) & \text{if } \mathcal{E}(\mathcal{T}_k^+) \neq \mathcal{T}_k^+ \\ \mathcal{T}_k^+ & \text{otherwise,} \end{cases} \quad \mathcal{T}_{k+1}^- := \begin{cases} \mathcal{C}(\mathcal{E}(\mathcal{T}_k^-); T_k^-)) & \text{if } \mathcal{E}(\mathcal{T}_k^-) \neq \mathcal{T}_k^- \\ \mathcal{T}_k^- & \text{otherwise,} \end{cases}$$

$$\mathcal{T}_{k+1}^+ := \mathcal{C}(\mathcal{E}(\mathcal{T}_k^+); T_k^+), \mathcal{T}_{k+1}^- := \mathcal{C}(\mathcal{E}(\mathcal{T}_k^-); T_k^-),$$

$$T_k^+ := \{v \in A(\mathcal{E}(\mathcal{T}_k^+)) : P(f_v) \geq \min_{\mathcal{F}_k^-} P + \delta + 10\mathbb{E}\Delta_n(\mathcal{F}_k^+) + \frac{12\tau_k^+ + 4}{\sqrt{n}}\},$$

$$T_k^- := \{v \in A(\mathcal{E}(\mathcal{T}_k^-)) : P(f_v) \geq \min_{\mathcal{F}_k^+} P + \delta - 10\mathbb{E}\Delta_n(\mathcal{F}_k^+) - \frac{12\tau_k^+ + 4}{\sqrt{n}}\},$$

$$\mathcal{F}_{k+1}^+ := \mathcal{F}_k^+ \cup \{f_v : v \in V(\mathcal{E}(\mathcal{T}_{k+1}^+))\},\ \mathcal{F}_{k+1}^- := \mathcal{F}_k^- \cup \{f_v : v \in V(\mathcal{E}(\mathcal{T}_{k+1}^-))\}.$$

$$\tau_k^+ := t_{h(\hat{T}_k^+)}, \tau_k^- := t_{h(\hat{T}_k^-)}.$$

Note that the choice of the parameter δ in the definitions above could pose a practical problem. A reasonable approach could be to use a prior upper bound on the accuracy of empirical approximation (for instance, in terms of VC-dimensions) as the value of δ.

We consider below two approaches to the problem of empirical risk minimization, based on the iterative pools of functions defined above. In the first approach, the number of iterations N is given in advance and we define $\tilde{f}_N := \text{argmin}_{\hat{\mathcal{F}}_N^-} P_n$. In the second approach, the Rademacher penalty is used to determine the number of iterations in a way close to optimal. Namely, we define, for $1 \leq N \leq \infty$ and for $\sigma \geq 0$, a random number $\hat{k}$ such that

$$\min_{f \in \hat{\mathcal{F}}_{\hat{k}}^-} P_n + 2R_n(\hat{\mathcal{F}}_{\hat{k}}^+) + \frac{3\hat{\tau}_{\hat{k}}^+}{\sqrt{n}} \leq \inf_{1 \leq k \leq N} [\min_{f \in \hat{\mathcal{F}}_k^-} P_n + 2R_n(\hat{\mathcal{F}}_k^+) + \frac{3\hat{\tau}_k^+}{\sqrt{n}}] + \sigma$$

and set $\hat{f}_N := \text{argmin}_{\hat{\mathcal{F}}_{\hat{k}}^-} P_n$.

The following theorems give probabilistic oracle inequalities for the empirical risk minimizers defined above.

3.1. Theorem. *For all* $N \geq 1$

$$\sup_{P \in \mathcal{P}(S)} \mathbb{P}\Big\{P(\tilde{f}_N) > \min_{\hat{\mathcal{F}}_N^-} P_n + 2R_n(\hat{\mathcal{F}}_N^+) + \frac{3\hat{\tau}_N^+}{\sqrt{n}}\Big\} \leq 6 \sum_{k=1}^{N} \exp\{-t_k^2/2\} \tag{3.1}$$

and

$$\sup_{P \in \mathcal{P}(S)} \mathbb{P}\Big\{P(\tilde{f}_N) > \min_{\mathcal{F}_N^-} P + 2\mathbb{E}\Delta_n(\mathcal{F}_N^+) + \frac{2\tau_N^+}{\sqrt{n}}\Big\} \leq 6 \sum_{k=1}^{N} \exp\{-t_k^2/2\}. \tag{3.2}$$

3.2. Theorem. *For all* $N,\ 1 \le N \le \infty$,

$$\sup_{P\in\mathcal{P}(S)} \mathbb{P}\Big\{P(\hat f_N) > \inf_{1\le k\le N}[\min_{\hat{\mathcal{F}}_k^-} P_n + 2R_n(\hat{\mathcal{F}}_k^+) + \frac{3\hat\tau_k^+}{\sqrt{n}}] + \sigma\Big\} \le 6\sum_{k=1}^{N}\exp\{-t_k^2/2\} \tag{3.3}$$

and

$$\sup_{P\in\mathcal{P}(S)} \mathbb{P}\Big\{P(\hat f_N) > \inf_{1\le k\le N}[\min_{\mathcal{F}_k^-} P + 5\mathbb{E}\Delta_n(\mathcal{F}_k^+) + \frac{6\tau_k^+ + 2}{\sqrt{n}}] + \sigma\Big\} \le 6\sum_{k=1}^{N}\exp\{-t_k^2/2\}. \tag{3.4}$$

The meaning of these results can be explained as follows. Define

$$\delta_k(P) := \min_{\mathcal{F}_k^-} P - \inf_{j\ge 1}\min_{\mathcal{F}_j^-} P.$$

This quantity gives the accuracy of approximation of the "minimal" theoretical risk at kth iteration. If we use P_n instead of P in the iteration process (but, miraculously, we are getting the correct pool of functions $\mathcal{F}_k^-$) and we stop at the k-th iteration, the error could become as much as $\delta_k(P) + \Delta_n(\mathcal{F}_k^-)$, which is less than $\delta_k(P) + \Delta_n(\mathcal{F}_k^+)$. If there were an oracle who could tell us what is the value of $\delta_k(P)$ and what is the average accuracy of empirical approximation $\mathbb{E}\Delta_n(\mathcal{F}_k^+)$ for the "theoretical" outer iteration pool, then, by choosing the number of iterations properly, we could achieve the average accuracy of the empirical risk minimization of the order $\inf_{1\le k\le N}[\delta_k(P) + \mathbb{E}\Delta_n(\mathcal{F}_k^+)]$.

Let $\gamma > \sqrt{2}$ and define $C_\gamma := 6\sum_{m\ge 1} m^{-\gamma^2/2}$. We set $t_k := \gamma\sqrt{\log k} + t$ with $t \ge \sqrt{2\log\frac{C_\gamma}{\alpha}}$. For simplicity, assume that $\sigma = 0$. Then, it follows from the bound (3.4) that for all $P \in \mathcal{P}(S)$ with probability at least $1-\alpha$

$$P(\hat f_N) - \inf_{j\ge 1}\min_{\mathcal{F}_j^-} P \le \inf_{1\le k\le N}[\delta_k(P) + 5\mathbb{E}\Delta_n(\mathcal{F}_k^+) + \frac{6\gamma\sqrt{\log k} + t + 2}{\sqrt{n}}].$$

Despite the fact that the oracle is not involved, the iterative structural risk minimization method allows us to achieve almost the same accuracy as with the help of the oracle (up to constants and some extra terms, that are relatively small) with guaranteed probability. This bound, of course, is more of theoretical interest, it demonstrates a form of optimality of the method. On the other hand, it follows from the bound (3.3) that for all $P \in \mathcal{P}(S)$ with probability at least $1-\alpha$

$$P(\hat f_N) \le \inf_{1\le k\le N}[\min_{\hat{\mathcal{F}}_k^-} P_n + 2R_n(\hat{\mathcal{F}}_k^+) + \frac{3\gamma\sqrt{\log k}}{\sqrt{n}}].$$

The expression in the right hand side can be computed based on the data, providing a conservative, but quite reasonable, confidence bound for the risk of the estimator $\hat f_N$.

4. Proofs of the oracle inequalities for iterative structural risk minimization. The following lemmas describe the properties of iterative trees and pools and they are the key ingredients of the proofs of Theorems 3.1–3.2.

4.1. Lemma. *Define*

$$l := \max\{j : h(\mathcal{T}_k) = k,\ k \le j\},$$

$$l^- := \max\{j : h(\mathcal{T}_k^-) = k,\ k \le j\}, l^+ := \max\{j : h(\mathcal{T}_k^+) = k,\ k \le j\},$$

$$\hat l^- := \max\{j : h(\hat{\mathcal{T}}_k^-) = k,\ k \le j\}, \hat l^+ := \max\{j : h(\hat{\mathcal{T}}_k^+) = k,\ k \le j\}.$$

Then

$$\mathcal{T}_k = \mathcal{T}_l,\ k \ge l, \mathcal{T}_k^- = \mathcal{T}_{l^-}^-,\ k \ge l^-, \mathcal{T}_k^+ = \mathcal{T}_{l^+}^+,\ k \ge l^+,$$

$$\hat{\mathcal{T}}_k^- = \hat{\mathcal{T}}_{\hat l^-}^-,\ k \ge \hat l^-, \hat{\mathcal{T}}_k^+ = \hat{\mathcal{T}}_{\hat l^+}^+,\ k \ge \hat l^+.$$

Proof. Let us prove that

$$\mathcal{T}_k^- = \mathcal{T}_{l^-}^-, \ k \geq l^-. \tag{4.1}$$

Clearly, this is the case when $\mathcal{E}(\mathcal{T}_{l^-}^-) = \mathcal{T}_{l^-}^-$ (in this case the tree does not grow further). Suppose that $\mathcal{E}(\mathcal{T}_{l^-}^-) \neq \mathcal{T}_{l^-}^-$. It easily follows from the definitions of the trees that, for all j, $h(\mathcal{T}_j^-) \leq j$ (since each iteration increases the height of the tree at most by 1). Clearly, we have $h(\mathcal{T}_{l^-+1}^-) = h(\mathcal{T}_{l^-}^-) = l$, which means that $T_{l^-}^- = A(\mathcal{E}(\mathcal{T}_{l^-}^-))$ (all alive vertices of $\mathcal{E}(\mathcal{T}_{l^-}^-)$ are going to be trimmed and the tree is going to stop growing) and hence $\mathcal{T}_{l^-+1}^- = \mathcal{T}_{l^-}^-$. Then, since $\mathcal{F}_{l^-}^+ \subset \mathcal{F}_{l^-+1}^+$, we get

$$T_{l^-+1}^- = \{v \in A(\mathcal{E}(\mathcal{T}_{l^-+1}^-)) : P(f_v) > \min_{\mathcal{F}_{l^-+1}^+} P + \delta - 10\mathbb{E}\Delta_n(\mathcal{F}_{l^-+1}^+) - \frac{12\tau_{l^-+1}^+ + 4}{\sqrt{n}}\} =$$

$$= \{v \in A(\mathcal{E}(\mathcal{T}_{l^-}^-)) : P(f_v) > \min_{\mathcal{F}_{l^-+1}^+} P + \delta - 10\mathbb{E}\Delta_n(\mathcal{F}_{l^-+1}^+) - \frac{12\tau_{l^-+1}^+ + 4}{\sqrt{n}}\} \supset$$

$$\supset \{v \in A(\mathcal{E}(\mathcal{T}_{l^-}^-)) : P(f_v) > \min_{\mathcal{F}_{l^-}^+} P + \delta - 10\mathbb{E}\Delta_n(\mathcal{F}_{l^-}^+) - \frac{12\tau_{l^-}^+ + 4}{\sqrt{n}}\} = T_{l^-}^-.$$

It follows that $T_{l^-+1}^- = A(\mathcal{E}(\mathcal{T}_{l^-}^-)) = T_{l^-}^-$, which implies $\mathcal{T}_{l^-+2}^- = \mathcal{T}_{l^-+1}^-$, and (4.1) follows by a simple induction.

The proofs of other relationships are quite similar, only with slight modifications. For instance, to prove that

$$\hat{\mathcal{T}}_k^+ = \hat{\mathcal{T}}_{\hat{l}^+}^+, \ k \geq \hat{l}^+ \tag{4.2}$$

assume that $\mathcal{E}(\hat{\mathcal{T}}_{\hat{l}^+}^+) \neq \hat{\mathcal{T}}_{\hat{l}^+}^+$ (otherwise (4.2) is obvious) and note that $h(\hat{\mathcal{T}}_{\hat{l}^++1}^+) = h(\hat{\mathcal{T}}_{\hat{l}^+}^+) = \hat{l}^+$, which means that $\hat{T}_{\hat{l}^+}^+ = A(\mathcal{E}(\hat{\mathcal{T}}_{\hat{l}^+}^+))$ and hence $\hat{\mathcal{T}}_{\hat{l}^++1}^+ = \hat{\mathcal{T}}_{\hat{l}^+}^+$. This implies that $\hat{\mathcal{F}}_{\hat{l}^+}^+ = \hat{\mathcal{F}}_{\hat{l}^++1}^+$ and since $\hat{\mathcal{F}}_{\hat{l}^+}^- \subset \hat{\mathcal{F}}_{\hat{l}^++1}^-$, we get

$$\hat{T}_{\hat{l}^++1}^+ = \{v \in A(\mathcal{E}(\hat{\mathcal{T}}_{\hat{l}^++1}^+)) : P_n(f_v) > \min_{\hat{\mathcal{F}}_{\hat{l}^++1}^-} P_n + \delta + 4R_n(\hat{\mathcal{F}}_{\hat{l}^++1}^+) + \frac{6\hat{\tau}_{\hat{l}^++1}^+}{\sqrt{n}}\} =$$

$$= \{v \in A(\mathcal{E}(\hat{\mathcal{T}}_{\hat{l}^+}^+)) : P_n(f_v) > \min_{\hat{\mathcal{F}}_{\hat{l}^++1}^-} P_n + \delta + 4R_n(\hat{\mathcal{F}}_{\hat{l}^+}^+) + \frac{6\hat{\tau}_{\hat{l}^++1}^+}{\sqrt{n}}\} \supset$$

$$\supset \{v \in A(\mathcal{E}(\hat{\mathcal{T}}_{\hat{l}^+}^+)) : P(f_v) > \min_{\hat{\mathcal{F}}_{\hat{l}^+}^-} P_n + \delta + 4R_n(\hat{\mathcal{F}}_{\hat{l}^+}^+) + \frac{6\hat{\tau}_{\hat{l}^+}^+}{\sqrt{n}}\} = \hat{T}_{\hat{l}^+}^+ = A(\mathcal{E}(\hat{\mathcal{T}}_{\hat{l}^+}^+)).$$

Hence $\hat{T}_{\hat{l}^++1}^+ = A(\mathcal{E}(\hat{\mathcal{T}}_{\hat{l}^+}^+)) = \hat{T}_{\hat{l}^+}^+$, which implies $\hat{\mathcal{T}}_{\hat{l}^++2}^+ = \hat{\mathcal{T}}_{\hat{l}^++1}^+$, and (4.2) again follows by a simple induction argument.

□

Lemma 4.1 immediately implies that

$$h(\mathcal{T}_k) = k \wedge l, h(\mathcal{T}_k^+) = k \wedge l^+, h(\mathcal{T}_k^-) = k \wedge l^-, h(\hat{\mathcal{T}}_k^+) = k \wedge \hat{l}^+, h(\hat{\mathcal{T}}_k^-) = k \wedge \hat{l}^-$$

and hence

$$\tau_k = t_{k \wedge l}, \tau_k^+ = t_{k \wedge l^+}, \tau_k^- = t_{k \wedge l^-}, \hat{\tau}_k^+ = t_{k \wedge \hat{l}^+}, \hat{\tau}_k^- = t_{k \wedge \hat{l}^-}.$$

4.2. Lemma. *Let $N \in \{1, 2, \dots, \infty\}$. Define*

$$E := \Bigl\{\omega : \forall k = 1, \dots, N : |R_n(\mathcal{F}_k^+) - \mathbb{E}R_n(\mathcal{F}_k^+)| < \frac{\tau_k^+}{\sqrt{n}}, |\Delta_n(\mathcal{F}_k^+) - \mathbb{E}\Delta_n(\mathcal{F}_k^+)| < \frac{\tau_k^+}{\sqrt{n}},$$

$$|R_n(\mathcal{F}_k) - \mathbb{E}R_n(\mathcal{F}_k)| < \frac{\tau_k}{\sqrt{n}}\Bigr\}.$$

Then

$$\mathbb{P}(E^c) \le 6 \sum_{k=1}^{N} \exp\{-t_k^2/2\}. \tag{4.3}$$

and on the event E

$$\forall k = 1, \dots, N : \mathcal{T}_k^- \prec \hat{\mathcal{T}}_k^- \prec \mathcal{T}_k \prec \hat{\mathcal{T}}_k^+ \prec \mathcal{T}_k^+, \tag{4.4}$$

$$\forall k = 1, \dots, N : \mathcal{F}_k^- \subset \hat{\mathcal{F}}_k^- \subset \mathcal{F}_k \subset \hat{\mathcal{F}}_k^+ \subset \mathcal{F}_k^+ \tag{4.5}$$

Proof. Lemma 4.1 implies that for $k = 1, \dots, l$ $\tau_k = t_k$ and for $k > l$ $\tau_k = t_l$ and $\mathcal{F}_k = \mathcal{F}_l$. Similarly, for $k = 1, \dots, l^+$ $\tau_k^+ = t_k$ and for $k > l^+$ $\tau_k^+ = t_{l^+}$ and $\mathcal{F}_k = \mathcal{F}_{l^+}$. Therefore,

$$E^c := \bigcup_{k=1}^{l^+} \Bigl\{|\Delta_n(\mathcal{F}_k^+) - \mathbb{E}\Delta_n(\mathcal{F}_k^+)| \ge t_k n^{-1/2}\Bigr\} \bigcup \bigcup_{k=1}^{l^+} \Bigl\{|R_n(\mathcal{F}_k^+) - \mathbb{E}R_n(\mathcal{F}_k^+)| \ge t_k n^{-1/2}\Bigr\} \bigcup$$

$$\bigcup \bigcup_{k=1}^{l} \Bigl\{|R_n(\mathcal{F}_k) - \mathbb{E}R_n(\mathcal{F}_k)| \ge t_k n^{-1/2}\Bigr\}.$$

It follows from lemmas 2.2, 2.3 that for all $k = 1, \dots, l^+$

$$\mathbb{P}\{|\Delta_n(\mathcal{F}_k^+) - \mathbb{E}\Delta_n(\mathcal{F}_k^+)| \ge t_k n^{-1/2}\} \le 2\exp\{-2t_k^2\}, \tag{4.6}$$

$$\mathbb{P}\{|R_n(\mathcal{F}_k^+) - \mathbb{E}R_n(\mathcal{F}_k^+)| \ge t_k n^{-1/2}\} \le 2\exp\{-t_k^2/2\}, \tag{4.7}$$

and for all $k = 1, \dots, l$

$$\mathbb{P}\{|R_n(\mathcal{F}_k) - \mathbb{E}R_n(\mathcal{F}_k)| \ge t_k n^{-1/2}\} \le 2\exp\{-t_k^2/2\}. \tag{4.8}$$

Then (4.6)–(4.8) imply the bound (4.3).

We will show that on the event E, for any $0 \le k \le N - 1$, the conditions

$$\mathcal{T}_j^- \prec \hat{\mathcal{T}}_j^- \prec \mathcal{T}_j \prec \hat{\mathcal{T}}_j^+ \prec \mathcal{T}_j^+, \ j \le k \tag{4.9}$$

and

$$\mathcal{F}_j^- \subset \hat{\mathcal{F}}_j^- \subset \mathcal{F}_j \subset \hat{\mathcal{F}}_j^+ \subset \mathcal{F}_j^+, j \le k \tag{4.10}$$

imply that

$$\mathcal{T}_{k+1}^- \prec \hat{\mathcal{T}}_{k+1}^- \prec \mathcal{T}_{k+1} \prec \hat{\mathcal{T}}_{k+1}^+ \prec \mathcal{T}_{k+1}^+ \tag{4.11}$$

and

$$\mathcal{F}^-_{k+1} \subset \hat{\mathcal{F}}^-_{k+1} \subset \mathcal{F}_{k+1} \subset \hat{\mathcal{F}}^+_{k+1} \subset \mathcal{F}^+_{k+1}. \tag{4.12}$$

Since $\mathcal{T}^-_0 = \hat{\mathcal{T}}^-_0 = \mathcal{T}_0 = \hat{\mathcal{T}}^+_0 = \mathcal{T}^+_0$ and $\mathcal{F}^-_0 = \hat{\mathcal{F}}^-_0 = \mathcal{F}_0 = \hat{\mathcal{F}}^+_0 = \mathcal{F}^+_0$ this would imply that

$$E \subset \left\{\forall k = 1, \dots, N : \mathcal{T}^-_k \prec \hat{\mathcal{T}}^-_k \prec \mathcal{T}_k \prec \hat{\mathcal{T}}^+_k \prec \mathcal{T}^+_k\right\} \bigcap$$

$$\bigcap \left\{\forall k = 1, \dots, N : \mathcal{F}^-_k \subset \hat{\mathcal{F}}^-_k \subset \mathcal{F}_k \subset \hat{\mathcal{F}}^+_k \subset \mathcal{F}^+_k\right\}$$

and the result would follow from the bound (4.3).

First we establish (4.11). To this end, we only prove that $\mathcal{T}^-_{k+1} \prec \hat{\mathcal{T}}^-_{k+1}$ (the proof of other relations is quite similar). If $l^- < k$, the conditions (4.9) and Lemma 4.1 imply that $\mathcal{T}^-_{k+1} = \mathcal{T}^-_{l^-} \prec \hat{\mathcal{T}}^-_{l^-} \prec \hat{\mathcal{T}}^-_{k+1}$ (the relation $\hat{\mathcal{T}}^-_j \prec \hat{\mathcal{T}}^-_{k+1}$ holds, obviously, for all $j \leq k+1$). Thus it is enough to consider the case when $l^- \geq k$ and, hence, $h(\mathcal{T}^-_k) = h(\hat{\mathcal{T}}^-_k) = k$. In this case the assumption $\mathcal{T}^-_k \prec \hat{\mathcal{T}}^-_k$ immediately implies that $\mathcal{E}(\mathcal{T}^-_k) \prec \mathcal{E}(\hat{\mathcal{T}}^-_k)$. If $\mathcal{E}(\mathcal{T}^-_k) = \mathcal{T}^-_k$, then obviously $\mathcal{T}_{k+1} \prec \hat{\mathcal{T}}^-_{k+1}$. Otherwise, it follows that the set $A(\mathcal{E}(\mathcal{T}^-_k))$ can be identified with a subset of $A(\mathcal{E}(\hat{\mathcal{T}}^-_k))$, so that the labels coincide, i.e. there exists a one-to-one map φ from $A(\mathcal{E}(\mathcal{T}^-_k))$ onto $V \subset A(\mathcal{E}(\hat{\mathcal{T}}^-_k))$ such that $f_v = f_{\varphi(v)}$.

Note that (4.9) implies $h(\hat{\mathcal{T}}^+_k) \leq h(\mathcal{T}^+_k)$, which in turn implies

$$\hat{\tau}^+_k = t_{h(\hat{\mathcal{T}}^+_k)} \leq t_{h(\mathcal{T}^+_k)} = \tau^+_k.$$

If $v \in A(\mathcal{E}(\mathcal{T}^-_k))$ and

$$P(f_v) < \min_{\mathcal{F}^+_k} P + \delta - 10\mathbb{E}\Delta_n(\mathcal{F}^+_k) - \frac{12\tau^+_k + 4}{\sqrt{n}},$$

then on the event E (using the fact that $\hat{\mathcal{F}}^+_k \subset \mathcal{F}^+_k$), we get

$$P_n(f_{\varphi(v)}) = P_n(f_v) < \min_{\mathcal{F}^+_k} P_n + \delta - 10\mathbb{E}\Delta_n(\mathcal{F}^+_k) - \frac{12\tau^+_k + 4}{\sqrt{n}} + 2\Delta_n(\mathcal{F}^+_k) \leq$$

$$\leq \min_{\mathcal{F}^+_k} P_n + \delta - 8\mathbb{E}\Delta_n(\mathcal{F}^+_k) - \frac{10\tau^+_k + 4}{\sqrt{n}} \leq$$

$$\leq \min_{\mathcal{F}^+_k} P_n + \delta - 4\mathbb{E}R_n(\mathcal{F}^+_k) - \frac{10\tau^+_k}{\sqrt{n}} \leq \min_{\mathcal{F}^+_k} P_n + \delta - 4R_n(\mathcal{F}^+_k) - \frac{6\tau^+_k}{\sqrt{n}} \leq$$

$$\leq \min_{\hat{\mathcal{F}}^+_k} P_n + \delta - 4R_n(\hat{\mathcal{F}}^+_k) - \frac{6\hat{\tau}^+_k}{\sqrt{n}}.$$

Hence, on the event E, $\varphi(A(\mathcal{E}(\mathcal{T}^-_k)) \setminus T^-_k) \subset A(\mathcal{E}(\hat{\mathcal{T}}^-_k)) \setminus \hat{T}^-_k$, which implies that

$$\mathcal{T}^-_{k+1} = \mathcal{C}(\mathcal{E}(\mathcal{T}^-_k); T^-_k) \prec \mathcal{C}(\mathcal{E}(\hat{\mathcal{T}}^-_k); \hat{T}^-_k) = \hat{\mathcal{T}}^-_{k+1}.$$

Next we prove that (4.9) and (4.10) imply (4.12). Since the proofs of all inclusions are similar, let us prove only that $\hat{\mathcal{F}}^+_{k+1} \subset \mathcal{F}^+_{k+1}$. Clearly, $h(\hat{\mathcal{T}}^+_j) \leq h(\mathcal{T}^+_j) \leq j, j \leq k+1$. If $\hat{l}^+ \geq k+1$, then $h(\hat{\mathcal{T}}^+_{k+1}) = h(\mathcal{T}^+_{k+1}) = k+1$ and the fact (previously proved) that $\hat{\mathcal{T}}^+_{k+1} \prec \mathcal{T}^+_{k+1}$ implies $\mathcal{E}(\hat{\mathcal{T}}^+_{k+1}) \prec \mathcal{E}(\mathcal{T}^+_{k+1})$. It follows that

$$\{f_v : v \in V(\mathcal{E}(\hat{\mathcal{T}}^+_{k+1}))\} \subset \{f_v : v \in V(\mathcal{E}(\mathcal{T}^+_{k+1}))\}.$$

By the definition of the classes $\hat{\mathcal{F}}^+_{k+1}, \mathcal{F}^+_{k+1}$ and the induction assumption, it follows that $\hat{\mathcal{F}}^+_{k+1} \subset \mathcal{F}^+_{k+1}$.

Otherwise, if $\hat{l}^+ < k+1$, we have by Lemma 4.3, that $\hat{\mathcal{T}}_j^+ = \hat{\mathcal{T}}_{\hat{l}^+}^+,\ j \geq \hat{l}^+$. Therefore,

$$\mathcal{E}(\hat{\mathcal{T}}_{k+1}^+) = \mathcal{E}(\hat{\mathcal{T}}_k^+) = \ldots = \mathcal{E}(\hat{\mathcal{T}}_{\hat{l}^+}^+),$$

which implies $\hat{\mathcal{F}}_{k+1}^+ = \hat{\mathcal{F}}_k^+ = \ldots = \hat{\mathcal{F}}_l^+$. By the induction assumption, we have $\hat{\mathcal{F}}_l^+ \subset \mathcal{F}_l^+$. Also, $\mathcal{F}_l^+ \subset \mathcal{F}_{k+1}^+$ for $l < k+1$, so, we conclude that $\hat{\mathcal{F}}_{k+1}^+ = \hat{\mathcal{F}}_l^+ \subset \mathcal{F}_l^+ \subset \mathcal{F}_{k+1}^+$.

□

Proof of Theorem 3.1. On the event E (see Lemma 4.2), in view of (4.5), we have $\hat{\mathcal{F}}_N^- \subset \mathcal{F}_N \subset \hat{\mathcal{F}}_N^+ \subset \mathcal{F}_N^+$ and

$$\tau_N = t_{h(\mathcal{T}_N)} \leq \hat{\tau}_N^+ = t_{h(\hat{\mathcal{T}}_N^+)} \leq \tau_N^+ = t_{h(\mathcal{T}_N^+)}.$$

Therefore, the following bounds hold (recall Lemma 2.5):

$$P(\tilde{f}_N) \leq P_n(\tilde{f}_N) + \Delta_n(\hat{\mathcal{F}}_N^-) \leq \min_{\hat{\mathcal{F}}_N^-} P_n + \Delta_n(\mathcal{F}_N) \leq$$

$$\leq \min_{\hat{\mathcal{F}}_N^-} P_n + \mathbb{E}\Delta_n(\mathcal{F}_N) + \frac{\tau_N}{\sqrt{n}} \leq \min_{\hat{\mathcal{F}}_N^-} P_n + 2\mathbb{E}R_n(\mathcal{F}_N) + \frac{\tau_N}{\sqrt{n}} \leq$$

$$\leq \min_{\hat{\mathcal{F}}_N^-} P_n + 2R_n(\hat{\mathcal{F}}_N) + \frac{3\tau_N}{\sqrt{n}} \leq \min_{\hat{\mathcal{F}}_N^-} P_n + 2R_n(\hat{\mathcal{F}}_N^+) + \frac{3\hat{\tau}_N^+}{\sqrt{n}},$$

which, by Lemma 4.2, implies (3.1). Similarly, we have

$$P(\tilde{f}_N) \leq P_n(\tilde{f}_N) + \Delta_n(\hat{\mathcal{F}}_N^-) \leq \min_{\hat{\mathcal{F}}_N^-} P_n + \Delta_n(\hat{\mathcal{F}}_N^-) \leq$$

$$\leq \min_{\hat{\mathcal{F}}_N^-} P + 2\Delta_n(\hat{\mathcal{F}}_N^-) \leq \min_{\mathcal{F}_N^-} P + 2\Delta_n(\mathcal{F}_N) \leq \min_{\hat{\mathcal{F}}_N^-} P + 2\mathbb{E}\Delta_n(\mathcal{F}_N^+) + \frac{2\tau_N^+}{\sqrt{n}},$$

which implies (3.2) by Lemma 4.2.

□

Proof of Theorem 3.2. Again we claim that on the event E for all $k = 1, \ldots, N$ $\hat{\mathcal{F}}_k^- \subset \mathcal{F}_k \subset \hat{\mathcal{F}}_k^+ \subset \mathcal{F}_k^+$ and

$$\tau_k = t_{h(\mathcal{T}_k)} \leq \hat{\tau}_k^+ = t_{h(\hat{\mathcal{T}}_k^+)} \leq \tau_k^+ = t_{h(\mathcal{T}_k^+)}.$$

Hence, we get

$$P(\hat{f}_N) \leq P_n(\hat{f}_N) + \Delta_n(\hat{\mathcal{F}}_{\hat{k}}^-) \leq \min_{\hat{\mathcal{F}}_{\hat{k}}^-} P_n + \Delta_n(\mathcal{F}_{\hat{k}}).$$

Since also on the event E for all $k = 1, \ldots, N$,

$$\Delta_n(\mathcal{F}_k) \leq \mathbb{E}\Delta_n(\mathcal{F}_k) + \frac{\tau_k}{\sqrt{n}} \leq 2\mathbb{E}R_n(\mathcal{F}_k) + \frac{\tau_k}{\sqrt{n}} \leq 2R_n(\mathcal{F}_k) + \frac{3\tau_k}{\sqrt{n}},$$

we get

$$P(\hat{f}_N) \leq \min_{\hat{\mathcal{F}}_{\hat{k}}^-} P_n + 2R_n(\mathcal{F}_{\hat{k}}) + \frac{3\tau_{\hat{k}}}{\sqrt{n}} \leq \min_{\hat{\mathcal{F}}_{\hat{k}}^-} P_n + 2R_n(\hat{\mathcal{F}}_{\hat{k}}^+) + \frac{3\hat{\tau}_{\hat{k}}^+}{\sqrt{n}} \leq$$

$$\leq \inf_{1 \leq k \leq N} [\min_{\hat{\mathcal{F}}_k^-} P_n + 2R_n(\hat{\mathcal{F}}_k^+) + \frac{3\hat{\tau}_k^+}{\sqrt{n}}] + \sigma,$$

and (3.3) follows by Lemma 4.2.

To prove (3.4), note that on the event E

$$\inf_{1\le k\le N}[\min_{\hat{\mathcal{F}}_k^-} P_n + 2R_n(\hat{\mathcal{F}}_k^+) + \frac{3\hat{\tau}_k^+}{\sqrt{n}}] \le \inf_{1\le k\le N}[\min_{\mathcal{F}_k^-} P + \Delta_n(\mathcal{F}_k^+) + 2R_n(\mathcal{F}_k^+) + \frac{3\tau_k^+}{\sqrt{n}}] \le$$

$$\le \inf_{1\le k\le N}[\min_{\mathcal{F}_k^-} P + \mathbb{E}\Delta_n(\mathcal{F}_k^+) + 2\mathbb{E}R_n(\mathcal{F}_k^+) + \frac{6\tau_k^+}{\sqrt{n}}] \le$$

$$\le \inf_{1\le k\le N}[\min_{\mathcal{F}_k^-} P + 5\mathbb{E}\Delta_n(\mathcal{F}_k^+) + \frac{6\tau_k^+ + 2}{\sqrt{n}}],$$

where we used the bound of Lemma 2.5.

□

Acknowledgment. The author is thankful to Chaouki Abdallah, Don Hush and Clint Scovel for interesting discussions. Clint Scovel and Don Hush also suggested several improvements and corrections. Fernando Lozano pointed out at several typos.

References

Azuma, K. (1967) Weighted sums of certain dependent random variables. *Tokuku Math. J.* 19, 357–367.

Barron, A., Birgé, L. and Massart, P. (1999) Risk Bounds for Model Selection via Penalization. *Probability Theory and Its Applications*, to appear.

Birgé, L. and Massart, P. (1997) From Model Selection to Adaptive Estimation. In: Festschrift for L. Le Cam. Research Papers in Probability and Statistics, D. Pollard, E. Torgersen and G. Yang (Eds.), 55-87, Springer, New York.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M.K. (1989) Learnability and the Vapnik-Chervonenkis Dimension. *J. of Assoc. for Computing Machinery*, 36, 929–965.

Buescher, K. and Kumar, P.R. (1996) Learning by Canonical Smooth Estimation - Part II: Learning and Choice of Model Complexity. *IEEE Trans. on Automatic Control*, 41, 557–569.

Devroye, L., Györfi, L. and Lugosi, G. (1996) A probabilistic theory of pattern recognition. Springer-Verlag, New York.

Dudley, R.M. (1999) Uniform Central Limit Theorems, Cambridge University Press, to appear.

Freund, Y. (1998) Self Bounding Learning Algorithms. Preprint.

Giné, E. and Zinn, J. (1984) Some limit theorems for empirical processes. *Ann. Probab.*, 12, 929–989.

Hush, D. and Scovel, C. (1999) On a result of Koltchinskii. Preprint.

Johnstone, I. M. (1998) Oracle Inequalities and Nonparametric Function Estimation. In: Documenta Mathematica, *Journal der Deutschen Mathematiker Vereinigung*, Proc. of the International Congress of Mathematicians, Berlin 1998, v.III, 267–278.

Koltchinskii, V. I. (1981) On the central limit theorem for empirical measures. *Probab. Theory and Math. Statist.* 24, 71–82.

Koltchinskii, V. I. (1985) Functional limit theorems and empirical entropy. I. *Probab. Theory and Math. Statist.* 33, 31–42.

Koltchinskii, V. I. (1986) Functional limit theorems and empirical entropy. II. *Probab▷ Theory and Math. Statist.* 34, 73–85.

Koltchinskii, V., Abdallah, C.T., Ariola, M., Dorato, P. and Panchenko, D. (1999) Statistical Learning Control of Uncertain Systems: It is better than it seems. Preprint, UNM

Ledoux, M. (1996) On Talagrand's deviation inequalities for product measures. *ESAIM*: *Probability and Statistics*, 1, 63-87, http://www.emath.fr/ps/

Ledoux, M. and Talagrand, M. (1991) Probability in Banach spaces. Springer-Verlag, New York.

Lugosi, G. and Zeger, K. (1996) Concept Learning Using Complexity Regularization. *IEEE Transactions on Information Theory*, 42, 48-54.

Massart, P. (1998) About the constants in Talagrand's concentration inequalities for empirical processes. Preprint, Université Paris-Sud.

McDiarmid, C. (1989) On the method of bounded differences. In: Surveys in Combinatorics, J. Siemons (Ed.), London Mathematical Soc. Lecture Notes, 141, Cambridge Univ. Press, p. 148–188.

Milman, V. and Schechtman, G. (1986) Asymptotic theory of finite dimensional normed spaces. Lecture Notes in Mathematics, 1200, Springer-Verlag, New York.

Pollard, D. (1982) A central limit theorem for empirical processes. *Journal of the Austral. Math. Soc.* A39, 235–248.

Shawe-Taylor, J., Bartlett, P., Williamson, R.C., and Anthony, M. (1996) Structural Risk Minimization over Data-Dependent Hierarchies, *NeuroCOLT*, Technical Report, NC-TR-96-053.

Rhee, W. T. and Talagrand, M. (1987) Martingale inequalities and NP-complete problems. *Math. Oper-ation Research*, 12, 177-181.

Talagrand, M. (1996a) A new look at independence. *Ann. Probab.* 24, 1–34.

Talagrand, M. (1996b) New concentration inequalities in product spaces. *Invent. Math.* 126, 505-563.

van der Vaart, A. and Wellner, J. (1996) Weak convergence and empirical processes. With Applications to Statistics. Springer-Verlag, New York.

Valiant, L. (1984) A theory of learnable. *Comm. of the ACM*, 27, 1134-1142.

Vapnik, V. (1982) Estimation of Dependencies Based on Empirical Data. Springer-Verlag, New York.

Vapnik, V. (1995) The nature of statistical learning theory. Springer-Verlag, New York.

Vapnik, V. (1998) Statistical Learning Theory. John Wiley & Sons, New York.

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

Vapnik, V.N. and Chervonenkis, A.Ya. (1974) Theory of Pattern Recognition. Nauka, Moscow.

Vidyasagar, M. (1997) A theory of learning and generalization. Springer-Verlag, New York.

Yurinski, V. (1974) Exponential bounds for large deviations. *Theory of Probability and its Applications*, 19, 154–155.