Rademacher Penalties and Structural Risk Minimization

Vladimir Koltchinskii

Abstract—**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—**Classification, empirical process, iterative structural risk minimization, oracle inequalities, Rademacher penalty, structural risk minimization.**

I. 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

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

To determine the set $C_{\rm opt}$ 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 $((X_1, Y_1), \ldots, (X_n, Y_n))$ 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 $\hat{C}_n := \arg\min\{L_n(C): C \in \mathcal{C}\}$, 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 [1], [2] (see also [3]–[7]). It plays also an important role in computational learning theory [8], [9].

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 $C_{\rm opt}$. In an attempt to make the minimal risk $\min_{C\in\mathcal{C}} L(C)$ smaller, one can try to choose a 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 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

$$\Delta^{\mathcal{C}}(F) := \operatorname{card}(\{F \cap C: C \in \mathcal{C}\})$$

and

$$m(\mathcal{C}, n) := \sup\left\{\Delta^{\mathcal{C}}(F): F \subset S, \operatorname{card}(F) = n\right\}, \qquad n \ge 1.$$

Then the VC-dimension of the class $\mathcal{C}$ is defined as

$$V(\mathcal{C}) := \sup\{n \ge 1: m(\mathcal{C}, n) = 2^n\}.$$

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

$$\hat{C} := \arg\min\left\{L_n(C): C \in \hat{\mathcal{C}}_{\hat{N}}\right\}$$
$$\hat{N} := \arg\min\left\{N \ge 1: \min_{C\in\mathcal{C}_N} L_n(C) + \operatorname{pen}(n; N)\right\}$$

Manuscript received July 22, 1999; revised January 9, 2001. This work was supported in part by NSA under Grant MDA904-99-1-0031 and by a Humboldt Research Fellowship.

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

Communicated by S. R. Kulkarni, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(01)04429-7.

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

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

which is, roughly,

$$\mathrm{const}\sqrt{(V(\mathcal{C}_N)\log n + N)/n}$$

(see, e.g., [10]). 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

$$\mathbb{P}\left\{\sup_{C\in\mathcal{C}} |L_n(C) - L(C)| \ge \varepsilon\right\} \le 4e^8 m\big(\mathcal{C},\, n^2\big)\, e^{-2n\varepsilon^2}. \tag{1.2}$$

Lugosi and Zeger [10] established the following bounds for the estimator $\hat{C}$:

$$\mathbb{P}\left\{L\big(\hat{C}\big) - \inf_{C\in\mathcal{C}_N} L(C) \ge \varepsilon\right\} \le e^{-n\varepsilon^2/2} + 4e^8 m\big(\mathcal{C}_N;\, n^2\big)\, e^{-n\varepsilon^2/8} \tag{1.3}$$

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

$$\mathbb{E}L\big(\hat{C}\big) - L_0 \le \inf_{N\ge 1}\left[\inf_{C\in\mathcal{C}_N} L(C) - L_0 + \sqrt{\frac{16V(\mathcal{C}_N)\log n + 8(N+11)}{n}}\right] \tag{1.4}$$

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

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 $L(C) \ge L_0$ for all $C \in \mathcal{C}$. Suppose that $V(\mathcal{C}) \ge 2$. Devroye, Györfi, and Lugosi [6] 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 [6, 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

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

for all

$$n \ge (2L_0)^{-1}((1-2L_0)^{-2} \vee 9)(V(\mathcal{C}) - 1)$$

(here and in what follows $a \vee b := \max(a,\, b)$ and $a \wedge b := \min(a,\, b)$).

Let $\{\mathcal{C}_N\}$ be an increasing sequence of VC-classes such that $\{V(\mathcal{C}_N)\}$ is strictly increasing and for some constant $D > 0$,

$$V(\mathcal{C}_{N+1}) \le DV(\mathcal{C}_N), \qquad N \ge 1.$$

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

$$0 \le \inf_{C\in\mathcal{C}_N} L(C) - L_0 \le \delta_N, \qquad N \ge 1.$$

It follows from (1.4) and (1.5) that with some constants $A, B > 0$

$$\sup_{\mathcal{P}} \mathbb{E}L\big(\hat{C}\big) - L_0 \le A \inf_{N\ge 1}\left[\delta_N + \sqrt{\frac{V(\mathcal{C}_N)\log n}{n}}\right] \tag{1.6}$$

and

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

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

$$||L_n - L||_{\mathcal{C}} := \sup_{C\in\mathcal{C}} |L_n(C) - L(C)|$$

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 the best possible and the VC-dimension of the class $\mathcal{C}$ of decision rules is often unknown and has to be replaced by an upper bound (this is the case, for instance, for some classes of neural networks). This hierarchy of nonoptimal upper bounds leads to the fact that the penalty function $\mathrm{pen}(n,\, N)$, defined by (1.1), is often much larger than the "ideal" penalty $\mathbb{E}||L_n - L||_{\mathcal{C}_N}$. The “ideal” penalty, however, cannot 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 [11], Barron, Birgé, and Massart [12] 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 [13]–[16] 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 [17]). 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

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

$\{r_n\}_{n\geq 1}$ 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*. Quantities similar to $R_n(\mathcal{C})$ have been frequently used in the so-called symmetrization inequalities for empirical processes (see Lemma 2.5 later). The method of Rademacher symmetrization, known in many areas of Analysis and Probability, was brought to the empirical processes theory by Koltchinskii [18], Pollard [19], and, especially, Giné and Zinn [20]. It allowed them to simplify substantially the proofs of the original Vapnik and Chervonenkis [1], [2] 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 [21], [22] for a 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 [22]). Recently, Koltchinskii, Abdallah, Ariola, Dorato, and Panchenko [23] used similar quantities in statistical learning problems that occur in control theory. Lozano [24] studied our method of Rademacher penalization (see Section II) in experiments with simulated data. Bartlett, Boucheron, and Lugosi [25] introduced the Rademacher penalties independently, studied them along with some other data-driven penalties (earlier, in [26] data-driven penalties based on random shatter coefficients were studied), and obtained several inequalities (close to the inequalities considered in Section II), justifying the method of Rademacher penalization. They also did some experiments with simulated data studying the Rademacher penalization. Koltchinskii and Panchenko [27] developed a localized version of Rademacher penalties and used it to bound the risk of function learning in the zero error case. Their bounds give optimal convergence rates of the risk to 0 in several learning problems. This might be of some interest, for instance, in support vector learning, where positive and negative examples are frequently linearly separable in the feature space, and also in bounding the generalization error of voting methods (although more specialized margin type bounds might be better in these examples).

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

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

so it is enough to compute separately the supremum and the infimum above. An easy computation shows that, for instance, finding the supremum can be reduced to minimizing

$$\sum_{j=1}^{n} I_{\{\tilde{Y}_j \neq I_C(X_j)\}}, \qquad \text{where } \tilde{Y}_j := \frac{1+r_j}{2} - Y_j r_j$$

and the computation of the infimum can be dealt with similarly. Thus, as soon as there exists a reliable algorithm of *precise* minimization of the training error (with arbitrary labels), one can use it to compute the Rademacher penalty as well (with no substantial increase of the computational complexity of the method; see [24] for an example of such situation). 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 a 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, leads to an optimal value of the 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 an 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 a hierarchy of function classes given in advance, this method allows one to determine finite data-dependent pools of functions in the data-driven iterative 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 [28]; "structural risk minimization over data-dependent hierarchies" of Shawe-Taylor, Bartlett, Williamson, and Anthony [29]; "self-bounding learning" of Freund [30]). 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.

II. 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 independent and identically distributed (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 $(X_1, \ldots, X_n)$

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

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

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

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

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 [21], [22] for the discussion of these issues).

Consider a family $\{\mathcal{F}_m\colon m \in \mathcal{M}\}$ 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 $\{n_m\colon m \in \mathcal{M}\}$. Let $\{t_m\colon m \in \mathcal{M}\}$ be a set of positive real numbers. We define an "ideal" penalty function by

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

and an empirical Rademacher penalty function by

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

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 $\{X_j\colon j = 1, \ldots, n_m\}_{m\in\mathcal{M}}$), such that

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

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

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

Theorem 2.1: Let $\hat{f}$ be chosen to satisfy (2.3). Then the following inequalities hold:

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

and

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

The proof uses the so-called McDiarmid Inequality (see [6, Theorem 9.2]) which is a corollary of Azuma's inequality [31] for martingale difference sequences. In fact, the argument that leads to the McDiarmid Inequality goes back to Yurinski [32] (see also [33]–[37], [6] for various versions and applications of the inequalities of this type).

We use this inequality to prove the following lemmas.

Lemma 2.2: For all $\varepsilon > 0$

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

and

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

Lemma 2.3: For all $\varepsilon > 0$

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

and

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

Lemma 2.4: For all $\varepsilon > 0$

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

and

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

Note that inequalities similar to the ones of Lemma 2.4 can be also obtained by combining the bounds of Lemmas 2.2 and 2.3, but this leads to worse values of the constants. (This improvements of the constants was suggested to the author by Don Hush and Clint Scovel [38].)

Lemma 2.5: The following inequalities hold:

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

Lemma 2.5 gives symmetrization inequalities for empirical processes. The proofs of the last two inequalities can be found, for example, in [22, Lemmas 2.3.1 and 2.3.6]. The proof of the first inequality is obvious: using Cauchy–Schwarz inequality, we get

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

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

$$P\left(\hat{f}\right)\leq P_{n_{\hat{m}}}\left(\hat{f}\right)+\Delta_{n_{\hat{m}}}(\mathcal{F}_{\hat{m}})\qquad(2.6)$$

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

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

This implies that with probability at least

$$1-\sum_{m\in\mathcal{M}}\exp\left\{-\frac{2}{3}t_m^2\right\}$$

we have (by the definition of $\hat{f}$, $\hat{m}$)

$$\begin{aligned}P(\hat{f})\leq&P_{n_{\hat{m}}}\left(\hat{f}\right)+2R_{n_{\hat{m}}}(\mathcal{F}_{\hat{m}})+3t_{\hat{m}}n_{\hat{m}}^{-1/2}\\ =&P_{n_{\hat{m}}}\left(\hat{f}\right)+\mathcal{E}(\hat{m})\\ \leq&\inf_{m\in\mathcal{M}}\left[\inf_{f\in\mathcal{F}_m}P_{n_m}(f)+\mathcal{E}(m)\right]+\delta\qquad(2.8)\end{aligned}$$

and the inequality (2.4) follows.

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

$$\begin{aligned}\mathbb{P}\Bigg(\bigcup_{m\in\mathcal{M}}&\Big\{\Delta_{n_m}(\mathcal{F}_m)+2R_{n_m}(\mathcal{F}_m)\\ &\geq\mathbb{E}[\Delta_{n_m}(\mathcal{F}_m)+2R_{n_m}(\mathcal{F}_m)]+3t_m n_m^{-1/2}\Big\}\Bigg)\\ &\leq\sum_{m\in\mathcal{M}}\exp\left\{-\frac{2}{3}t_m^2\right\}.\end{aligned}$$

We also have (using the bounds of Lemma 2.5)

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

Therefore,

$$\begin{aligned}\mathbb{P}\Bigg(\bigcup_{m\in\mathcal{M}}&\Big\{\Delta_{n_m}(\mathcal{F}_m)+2R_{n_m}(\mathcal{F}_m)\\ &\geq 5\mathbb{E}\Delta_{n_m}(\mathcal{F}_m)+3t_m n_m^{-1/2}+2n_m^{-1/2}\Big\}\Bigg)\\ &\leq\sum_{m\in\mathcal{M}}\exp\left\{-\frac{2}{3}t_m^2\right\}.\qquad(2.10)\end{aligned}$$

Using (2.8)–(2.10), we conclude that with probability at least

$$1-2\sum_{m\in\mathcal{M}}\exp\left\{-\frac{2}{3}t_m^2\right\}$$

we have

$$\begin{aligned}P\left(\hat{f}\right)\leq&\inf_{m\in\mathcal{M}}\left[\inf_{f\in\mathcal{F}_m}P_{n_m}(f)+\mathcal{E}(m)\right]+\delta\\ \leq&\inf_{m\in\mathcal{M}}\Big[\inf_{f\in\mathcal{F}_m}P(f)+\Delta_{n_m}(\mathcal{F}_m)+2R_{n_m}(\mathcal{F}_m)\\ &+3t_m n_m^{-1/2}\Big]+\delta\\ \leq&\inf_{m\in\mathcal{M}}\Big[\inf_{f\in\mathcal{F}_m}P(f)+5\mathbb{E}\Delta_{n_m}(\mathcal{F}_m)+2n_m^{-1/2}\\ &+6t_m n_m^{-1/2}\Big]+\delta\\ =&\inf_{m\in\mathcal{M}}\left[\inf_{f\in\mathcal{F}_m}P(f)+\mathcal{I}(m)\right]+\delta\qquad(2.10')\end{aligned}$$

which completes the proof. □

Let

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

Corollary 2.6: If $\hat{f}$ is chosen to satisfy (2.3), then the following inequality holds for all $P\in\mathcal{P}(S)$:

$$\begin{aligned}\mathbb{E}P\left(\hat{f}\right)\leq&\inf_{m\in\mathcal{M}}\left[\inf_{f\in\mathcal{F}_m}P(f)+\mathcal{I}(m)\right]+\delta\\ &+6\sqrt{6\pi}\sum_{m\in\mathcal{M}}\frac{1}{\sqrt{n_m}}\left[1-\Phi\left(\frac{2}{\sqrt{3}}\,t_m\right)\right].\qquad(2.11)\end{aligned}$$

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

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

and

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

Define

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

and $\xi^+ := \xi \vee 0$. It follows from (2.5′) that for all $\varepsilon > 0$

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

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

$$\begin{aligned}\mathbb{E}\xi \leq \mathbb{E}\xi^+ &= \int_0^{+\infty} \mathbb{P}\{\xi^+ \geq \varepsilon\}\, d\varepsilon \\ &\leq 2\sum_{m\in\mathcal{M}} \int_0^{+\infty} \exp\Big\{-\frac{2}{3}\Big(t_m + \varepsilon n_m^{1/2}\Big)^2\Big\}\, d\varepsilon \\ &= 2\sum_{m\in\mathcal{M}} \frac{1}{\sqrt{n_m}} \int_0^{+\infty} \exp\Big\{-\frac{2}{3}(t_m + v)^2\Big\}\, dv \\ &= \sqrt{6\pi} \sum_{m\in\mathcal{M}} \frac{1}{\sqrt{n_m}} \Big[1 - \Phi\Big(\frac{2}{\sqrt{3}}\, t_m\Big)\Big]\end{aligned}$$

and (2.11) easily follows. □

In particular, assume that $n_m \equiv n$ and let

$$C := 2\sum_{m\in\mathcal{M}} \exp\Big\{-\frac{2}{3} t_m^2\Big\} < +\infty.$$

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

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

and

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

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

$$\sup_{P\in\mathcal{P}(S)} \mathbb{P}\Big\{P\Big(\hat{f}\Big) \geq \inf_{m\in\mathbb{N}}\Big[\inf_{f\in\mathcal{F}_m} P(f) + \frac{5D_m(P) + 6\gamma\sqrt{\log m} + 2}{\sqrt{n}}\Big] + \delta + 6\varepsilon\Big\} \leq C_\gamma \exp\Big\{-\frac{2}{3}\varepsilon^2 n\Big\} \qquad (2.12)$$

and

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

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

$$\tilde{m} := \arg\min\Big[\delta_m(P) + \frac{D_m(P)}{\sqrt{n}}\Big].$$

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 inequalities 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 $1 - \alpha$

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

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

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

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 $\mathcal{F}_m := \{I_C\colon C \in \mathcal{C}_m\}$, 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., [22, 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 I (see (1.6), (1.7)), we have

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

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 cannot 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, see (2.10′)).

III. 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 [30] and Langford and Blum [39] 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 $\{\mathcal{F}_k^-\}$, $\{\mathcal{F}_k\}$, and $\{\mathcal{F}_k^+\}$. 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 $\mathcal{V}$ be a countable set and $\mathcal{L}$ be a set. 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 $\mathcal{G}$. 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 $v_0, \ldots, V_k(\mathcal{G})$ the set of all vertices of $\mathcal{G}$ connected to V_0 with a path of length k. Let $h(\mathcal{G}) := \max\{k: V_k(\mathcal{G}) \neq \emptyset\}$ be *the height* of the rooted tree $\mathcal{G}$. Denote $A(\mathcal{G}) := V_{h(\mathcal{G})}(\mathcal{G})$. We call the vertices in this set alive. For a rooted $\mathcal{L}$-labeled tree $\mathcal{T} = (\mathcal{G}; L)$, we use the notations $V(\mathcal{T}), V_k(\mathcal{T}), \ldots$ 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 $\mathcal{T}(v)$.

We use in what follows an ordering on the set of all rooted labeled trees, defined recursively 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) $\operatorname{card}(V_1(\mathcal{T}_1)) \leq \operatorname{card}(V_1(\mathcal{T}_2))$ 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.

Next we introduce *an extension operator* $\mathcal{E}$ on the set of all $\mathcal{L}$-labeled rooted trees. Let $\operatorname{Fin}(\mathcal{L})$ be the class of all finite subsets of $\mathcal{L}$. For all $k \geq 1$, define a mapping $\mathcal{E}_k: \mathcal{L}^k \mapsto \operatorname{Fin}(\mathcal{L})$. Suppose that $\pi = ((v_0, l_0), \ldots, (v_k, l_k))$, where $l_j := L(v_j)$ is a path through the labeled tree from the root to a terminal vertex v_k. Let $F := \mathcal{E}_{k+1}(l_0, \ldots, l_k)$. Given π, $\mathcal{E}(\pi)$ is obtained by adding $\operatorname{card}(F)$ 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 extensions $\mathcal{E}(\mathcal{T})$, 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 neighborhood $N(l) \subset \mathcal{L}$. For $k \geq 0$, define $\mathcal{E}_{k+1}(l_0, \ldots, l_k) := N(l_k)$. This defines an extension operator $\mathcal{E}$, which is closely related 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: $\mathcal{T}_k := \mathcal{E}(\mathcal{T}_{k-1})$, $k = 1, 2, \ldots$.

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 neighborhood $N(l) \subset \mathbb{Z}^d$. We define $\mathcal{E}_1(l_0) := N(l_0)$ and for $k \geq 1$

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

This defines another extension operator, which can be used rather naturally to construct minimization algorithms of steepest descent type. Some more sophisticated versions of this example (in which, for instance, the size of the iterative steps is being changed) can be easily defined in order to model more complicated iterative minimization and search techniques.

Example 3.3: 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 by adding neurons). Assume that $\mathcal{L} := \bigcup_{j=1}^{\infty} \mathcal{L}_j$. Denote

$$c(l) := \inf\{j \geq 1 \colon l \in \mathcal{L}_j\}, \qquad l \in \mathcal{L}$$

($c(l)$ 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 $(l_0, \ldots, l_k)$ such that $l_k = l_{k-1}$, we set $\mathcal{E}_{k+1}(l_0, \ldots, l_k) = \{\Gamma(l_k)\}$. Otherwise, if $l_k \neq l_{k-1}$, let $s \geq 1$ be the smallest integer such that

$$c(l_k) = c(l_{k-1}) = \cdots = c(l_{k-s+1}) = i \neq c(l_{k-s}).$$

We define

$$\mathcal{E}_{k+1}(l_0, \ldots, l_k) = \mathcal{E}_s^{(i)}(l_{k-s+1}, \ldots, l_k).$$

In other words, as soon as the double point occurs in the sequence, we increase the complexity of the search 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*. Given a rooted tree $\mathcal{G}$ and a set $V \subset A(\mathcal{G})$, we denote $\mathcal{C}(\mathcal{G}; V)$ (the trimming of $\mathcal{G}$ at V), the tree obtained from $\mathcal{G}$ by eliminating all the vertices of the set V and all the edges connecting them to the rest of the tree. Given a rooted labeled tree $\mathcal{T}$, we use quite similarly the notation $\mathcal{C}(\mathcal{T}; V)$.

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 $f_v := L(v)$.

Given a labeled tree $\mathcal{T}$ and $T \subset A(\mathcal{E}(\mathcal{T}))$, we denote

$$\mathcal{T} \triangleright T := \begin{cases} \mathcal{C}(\mathcal{E}(\mathcal{T}); T), & \text{if } \mathcal{E}(\mathcal{T}) \neq \mathcal{T} \\ \mathcal{T}, & \text{otherwise.} \end{cases}$$

Also set $\mathcal{F}[\mathcal{T}] := \{f_v \colon v \in A(\mathcal{E}(\mathcal{T}))\}$. In what follows, we deal with sequences of growing trees obtained by applying the extension operator recursively. Given such a sequence of labeled trees $\{\mathcal{T}_k\}$, we define a sequence of function classes $\{\mathcal{F}_k\}$ as follows:

$$\mathcal{F}_{k+1} := \mathcal{F}_k \cup \mathcal{F}[\mathcal{T}_{k+1}] \qquad \mathcal{F}_0 := \{f_v \colon v \in V(\mathcal{E}(\mathcal{T}_0))\}.$$

We call $\{\mathcal{F}_k\}$ *the sequence of function pools* of $\{\mathcal{T}_k\}$.

We now turn to the definition of labeled trees and function pools related to the process of minimization of a target function $J \colon \mathcal{F} \mapsto \mathbb{R}$ (it will be either P or P_n in what follows). Let $\delta > 0$ be a fixed parameter. We start with an $\mathcal{F}$-labeled tree $\mathcal{T}_0(J)$ with one vertex, say (v_0, f_0), and then define recursively, for $k = 0, 1, 2, \ldots\ \mathcal{T}_{k+1}(J) := \mathcal{T}_k(J) \triangleright T_k(J)$, where[1]

$$T_k(J) := \left\{ v \in A(\mathcal{E}(\mathcal{T}_k(J))) \colon J(f_v) > \min_{\mathcal{F}_k(J)} J + \delta \right\}$$

$\{\mathcal{F}_k(J)\}$ being the sequence of function pools of $\{\mathcal{T}_k(J)\}$. 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 the extension operator to create the pool of iterations $\mathcal{F}_0(J)$ and to select from this pool the functions (labels) with too large value of J. We trim the extended tree $\mathcal{E}(\mathcal{T}_0(J))$ to get rid of these functions (and the vertices they label) and we are getting the tree $\mathcal{T}_1(J)$ as the result. Then the iterative process continues recursively. The parameter δ controls the threshold level of trimming and determines how exhaustive the search for minimum is. The value $\delta = 0$ results in an algorithm of steepest descent type. Larger values of δ would lead to more exhaustive search.

In this paper, we deal with the situation when the target function J is not known precisely and has to be replaced by an approximation $\tilde{J} \colon \mathcal{F} \mapsto \mathbb{R}$. In this case, it is important to define the function pools for $\tilde{J}$ in such a way that they approximate in some sense the function pools for J. To this end, let $k \mapsto \Pi(k) \geq 0$ be a nondecreasing "penalty function" (that will bound in what follows the "distance" of $\tilde{J}$ to J at different iterations). We define sequences $\{\mathcal{T}_k^+(J; \Pi)\}$ and $\{\mathcal{T}_k^-(J; \Pi)\}$ of labeled trees starting with $\mathcal{T}_0^+(J; \Pi) = \mathcal{T}_0^-(J; \Pi) := \mathcal{T}_0(J)$ and then continuing recursively

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

where

$$T_k^+(J, \Pi) := \Bigg\{ v \in A\big(\mathcal{E}\big(\mathcal{T}_k^+(J; \Pi)\big)\big) \colon J(f_v) > \min_{\mathcal{F}_k^-(J, \Pi)} J + \delta + \Pi(k) \Bigg\}$$

$$T_k^-(J, \Pi) := \Bigg\{ v \in A\big(\mathcal{E}\big(\mathcal{T}_k^-(J; \Pi)\big)\big) \colon J(f_v) > \min_{\mathcal{F}_k^+(J, \Pi)} J + \delta - \Pi(k) \Bigg\}$$

$\{\mathcal{F}_k^+(J; \Pi)\}$ and $\{\mathcal{F}_k^-(J; \Pi)\}$ being the sequences of function pools of the trees $\{\mathcal{T}_k^+(J; \Pi)\}$ and $\{\mathcal{T}_k^-(J; \Pi)\}$, respectively. Note that in the above definition, the penalty $\Pi(k)$ might depend (and will depend below) on the trees $\mathcal{T}_j^+(J; \Pi)$, $\mathcal{T}_j^-(J; \Pi)$, $j \leq k$, already defined at the kth iteration.

The following properties of the iterative trees and pools easily follow from the definitions by a simple induction argument.

Property A: For all k

$$\mathcal{T}_k^-(J; \Pi) \prec \mathcal{T}_k(J) \prec \mathcal{T}_k^+(J; \Pi)$$

and

$$\mathcal{F}_k^-(J; \Pi) \subset \mathcal{F}_k(J) \subset \mathcal{F}_k^+(J; \Pi).$$

[1] Here and in what follows we use the notation $\min_{\mathcal{G}} J := \min_{g \in \mathcal{G}} J(g)$.

Moreover, if $\Pi(k) \equiv 0$, then

$$\mathcal{T}_k^+(J, \Pi) = \mathcal{T}_k^-(J, \Pi) = \mathcal{T}_k(J), \qquad k \geq 0.$$

Property B: Suppose that $\Pi(k) := \pi(\mathcal{T}_k^+)$, $k \geq 0$, where $\mathcal{T} \mapsto \pi(\mathcal{T}) \geq 0$ is a monotone functional on rooted labeled trees (i.e., $\mathcal{T}_1 \prec \mathcal{T}_2$ implies that $\pi(\mathcal{T}_1) \leq \pi(\mathcal{T}_2)$). Let

$$\begin{aligned} l := l(J) &:= \min\{j\colon \mathcal{T}_{j+1}(J) = \mathcal{T}_j(J)\} \\ l^- := l^-(J; \Pi) &:= \min\{j\colon \mathcal{T}_{j+1}^-(J, \Pi) = \mathcal{T}_j^-(J, \Pi)\} \\ l^+ := l^+(J; \Pi) &:= \min\{j\colon \mathcal{T}_{j+1}^+(J, \Pi) = \mathcal{T}_j^+(J, \Pi)\}. \end{aligned}$$

Then $l^- \leq l \leq l^+$ and

$$\begin{aligned} \mathcal{T}_k(J) &= \mathcal{T}_l(J), && k \geq l \\ \mathcal{T}_k^-(J, \Pi) &= \mathcal{T}_{l^-}^-(J, \Pi), && k \geq l^- \\ \mathcal{T}_k^+ &= \mathcal{T}_{l^+}^+(J, \Pi), && k \geq l^+. \end{aligned}$$

This property means that each of the trees stops growing as soon as it remains the same for two consecutive iterations. The numbers l, l^-, l^+ will be called *the stopping times* of the corresponding trees (note that they can be equal to $+\infty$). Clearly, we have the following formulas for the heights of the trees:

$$\begin{aligned} h(\mathcal{T}_k(J)) &= k \wedge l \\ h(\mathcal{T}_k^-(J, \Pi)) &= k \wedge l^- \end{aligned}$$

and

$$h(\mathcal{T}_k^+(J, \Pi)) = k \wedge l^+.$$

Property C: If $\tilde{J}\colon \mathcal{F} \mapsto \mathbb{R}$ is such that

$$\left\| \tilde{J} - J \right\|_{\mathcal{F}_k^+(\tilde{J}; \Pi)} \leq \Pi(k)/2, \qquad k \geq 0$$

then

$$\mathcal{T}_k^-\left(\tilde{J}; \Pi\right) \prec \mathcal{T}_k(J) \prec \mathcal{T}_k^+\left(\tilde{J}; \Pi\right)$$

and

$$\mathcal{F}_k^-\left(\tilde{J}; \Pi\right) \subset \mathcal{F}_k(J) \subset \mathcal{F}_k^+\left(\tilde{J}; \Pi\right), \qquad k \geq 0.$$

In addition,

$$\min_{\mathcal{F}_k^+(\tilde{J}; \Pi)} \tilde{J} - \Pi(k)/2 \leq \min_{\mathcal{F}_k(J)} J \leq \min_{\mathcal{F}_k^-(\tilde{J}; \Pi)} \tilde{J} + \Pi(k)/2, \qquad k \geq 0.$$

This property shows that, as soon as a "small enough" penalty Π bounds the accuracy of approximation of J by $\tilde{J}$, the "iterative minima" of $\tilde{J}$ approximate the "iterative minima" of J.

We use the above construction to define proper iterative function pools for the empirical risk P_n and to use them to approximate the iterative process of minimization of the true risk P.

Let $\mathcal{T}_k := \mathcal{T}_k(P)$ and $\mathcal{F}_k := \mathcal{F}_k(P)$, l being the corresponding stopping time. Clearly, the trees $\mathcal{T}_k$ and the pools of functions $\mathcal{F}_k$ cannot be computed unless the distribution P is known precisely and the empirical (data-dependent) versions of these objects are needed. Let $\{t_k\colon k \geq 1\}$ be a nondecreasing sequence of nonnegative numbers. We now define the trees $\hat{\mathcal{T}}_k^+ := \mathcal{T}_k^+(P_n, \hat{\Pi})$, $\hat{\mathcal{T}}_k^- := \mathcal{T}_k^-(P_n, \hat{\Pi})$ and pools $\hat{\mathcal{F}}_k^+ := \mathcal{F}_k^+(P_n, \hat{\Pi})$, $\hat{\mathcal{F}}_k^- := \mathcal{F}_k^-(P_n, \hat{\Pi})$ (the corresponding stopping times being $\hat{l}^+$ and $\hat{l}^-$), where

$$\hat{\Pi}(k) := 4R_n\left(\hat{\mathcal{F}}_k^+\right) + \frac{6\hat{\tau}_k^+}{\sqrt{n}}, \qquad \hat{\tau}_k^+ := t_{k \wedge \hat{l}^+}.$$

These empirical rooted trees and the corresponding pools of functions can be computed recursively, given the empirical data. It is also easy to see that the minima of P_n and the Rademacher penalties can be updated iteratively when the extension operator adds new functions to the pools.

We also define the trees $\mathcal{T}_k^+ := \mathcal{T}_k^+(P, \overline{\Pi})$, $\mathcal{T}_k^- := \mathcal{T}_{\underline{k}}^-(P, \overline{\Pi})$ and the pools $\mathcal{F}_k^+ := \mathcal{F}_k^+(P, \overline{\Pi})$, $\mathcal{F}_k^- := \mathcal{F}_k^-(P, \overline{\Pi})$ (with stopping times l^+, l^-), where

$$\overline{\Pi}(k) := 10\mathbb{E}\Delta_n(\mathcal{F}_k^+) + \frac{12\tau_k^+ + 4}{\sqrt{n}}, \qquad \tau_k^+ := t_{k \wedge l^+}.$$

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 $\tilde{f}_N := \arg\min_{\hat{\mathcal{F}}_N^-} P_n$. 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 $2 \leq N \leq \infty$ and for $\sigma \geq 0$, a random number $\hat{k}$ such that

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

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

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

Theorem 3.1: For all $N \geq 2$

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

and

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

Theorem 3.2:

For all $2 \leq N \leq \infty$

$$\begin{aligned} &\sup_{P \in \mathcal{P}(S)} \mathbb{P}\left\{ P\left(\hat{f}_N\right) > \inf_{1 \leq k < N} \left[\min_{\hat{\mathcal{F}}_k^-} P_n + 2R_n\left(\hat{\mathcal{F}}_k^+\right) \right.\right. \\ &\qquad \left.\left. + \frac{3\hat{\tau}_k^+}{\sqrt{n}} \right] + \sigma \right\} \\ &\qquad \leq 6 \sum_{k<N} \exp\{-t_k^2/2\} \qquad (3.3) \end{aligned}$$

and

$$\begin{aligned} &\sup_{P \in \mathcal{P}(S)} \mathbb{P}\left\{ P\left(\hat{f}_N\right) > \inf_{1 \leq k < N} \left[\min_{\mathcal{F}_k^-} P + 5\mathbb{E}\Delta_n(\mathcal{F}_k^+) \right.\right. \\ &\qquad \left.\left. + \frac{6\tau_k^+ + 2}{\sqrt{n}} \right] + \sigma \right\} \\ &\qquad \leq 6 \sum_{k<N} \exp\{-t_k^2/2\}. \qquad (3.4) \end{aligned}$$

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

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

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 kth 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 $\inf_{1\leq k<N}[\delta_k(P) + \mathbb{E}\Delta_n(\mathcal{F}_k^+)]$.

Let $\gamma > \sqrt{2}$ and define

$$C_\gamma := 6\sum_{m\geq 1} m^{-\gamma^2/2}.$$

We set $t_k := \gamma\sqrt{\log k} + t_\alpha$ with $t_\alpha := \sqrt{2\log\frac{C_\gamma}{\alpha}}$. 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 $1-\alpha$

$$P\left(\hat{f}_N\right) - \inf_{j\geq 1}\min_{\mathcal{F}_j^-} P \leq \inf_{1\leq k<N}\left[\delta_k(P) + 5\mathbb{E}\Delta_n(\mathcal{F}_k^+) + \frac{6\gamma\sqrt{\log k} + 6t_\alpha + 2}{\sqrt{n}}\right].$$

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 $1-\alpha$

$$P\left(\hat{f}_N\right) \leq \inf_{1\leq k<N}\left[\min_{\hat{\mathcal{F}}_k^-} P_n + 2R_n\left(\hat{\mathcal{F}}_k^+\right) + \frac{3\gamma\sqrt{\log k} + 3t_\alpha}{\sqrt{n}}\right].$$

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 $\hat{f}_N$.

We conclude this section with some remarks about the computational complexity of the iterative structural risk minimization and about the choice of the parameter δ in the above algorithms.

Let us set $\Pi(k) \equiv \delta/2$ (δ being the threshold parameter introduced above) and define sequences of trees

$$\mathcal{T}_{k,\delta}^+ := \mathcal{T}_k^+(P, \delta/2) \quad \text{and} \quad \mathcal{T}_{k,\delta}^- := \mathcal{T}_k^-(P, \delta/2).$$

Clearly, $\mathcal{F}_{k,\delta}^+$ and $\mathcal{F}_{k,\delta}^-$ denote the function pools. The next proposition easily follows from Theorem 3.2 (using a well-known bound

$$\mathbb{E}\Delta_n\left(\mathcal{F}_{k,\delta}^+\right) \leq C\sqrt{\frac{\log\operatorname{card}\left(\mathcal{F}_{k,\delta}^+\right)}{n}}$$

with some numerical constant C; see, e.g., [22, Secs. 2.2 and 2.4]).

Proposition: Suppose that for some $\delta > 0$ there exist $\beta > 0$ and $\gamma > 0$ such that

$$\operatorname{card}\left(\mathcal{F}_{k,\delta}^+\right) = O\left(k^\beta\right), \qquad \text{as } k \to \infty$$

and

$$\min_{\mathcal{F}_{k,\delta}^-} P - \inf_{j\geq 1}\min_{\mathcal{F}_{j,\delta}^-} P = O\left(k^{-\gamma}\right), \qquad \text{as } k \to \infty.$$

Then there exist constants $C_1 > 0$ and $C_2 > 0$ such that for all $P \in \mathcal{P}(S)$ with probability at least $1-\alpha$

$$P\left(\hat{f}_N\right) - \inf_{j\geq 1}\min_{\mathcal{F}_{j,\delta}^-} P \leq \varepsilon$$

as soon as $n \geq C_1\varepsilon^{-2}[\log\frac{1}{\varepsilon} \vee \log\frac{1}{\alpha}]$ and $N \geq C_2\varepsilon^{-1/\gamma}$.

The conditions of the proposition actually mean that, for a given extension operator and for some δ, the minimization algorithm for P is "polynomial time" (the computational time is proportional to the size of the function pool needed to achieve certain accuracy ε of approximate minimization; the conditions imply that the cardinality of the pool is of the order $O(\varepsilon^{-\beta/\gamma})$). The proposition shows that the data-driven iterative structural minimization of P_n is then also "polynomial time." Moreover, if computation of the value of a function f at a point x is viewed as an elementary operation, the number of such operations needed in order to achieve accuracy ε with a guaranteed probability is $O(\varepsilon^{-2-\beta/\gamma}\log\frac{1}{\varepsilon})$.

Next we turn to the problem of choosing the value of the parameter δ in the definitions of the trees and iterative pools above. On the one hand, the parameter δ controls the size of distribution-dependent iterative function pools $\mathcal{F}_k$. The larger values of δ would result in more exhaustive and computationally involved search and, in principle, could give a better solution of the risk minimization problem at a higher cost. The study of this aspect of the problem definitely goes beyond the scope of the current paper and could be accomplished only under specific conditions on the extension operator in special learning problems. On the other hand, it is clear from the definitions of the data-driven pools $\hat{\mathcal{F}}_k^-$ and $\hat{\mathcal{F}}_k^+$ that the value of δ should be larger than the penalty term $\hat{\Pi}(k)$ involved in the definition. As soon as this condition is violated, the growth of the data-driven pools $\hat{\mathcal{F}}_k^-$ would stop. Thus, one might think about using the value $\delta := \overline{\delta} + B$, where $\overline{\delta}$ is a parameter that determines how exhaustive the search should be (one can set $\overline{\delta} = 0$ if one is happy with an algorithm of steepest descent type) and B is a prior upper bound on the accuracy of empirical approximation (for instance, in terms of the VC-dimensions). Note that the value of δ is not involved explicitly in the bounds of Theorems 3.1 and 3.2, so, even in the case when the prior bound B is conservative, the bounds of the theorems can, in principle, be better than the prior bound. Of course, a large value of B would result in exhaustive and computationally involved search in the process of risk minimization. An alternative to this approach is to replace the bound B by data-driven quantities (based again on the Rademacher penalties). For instance, given $C > 10$, one can replace δ in the definition of the trees $\hat{\mathcal{T}}_k^-$ and $\hat{\mathcal{T}}_k^+$ by the quantities

$$\hat{\delta}_k^- := \overline{\delta} + 2CR_n\left(\hat{\mathcal{F}}_k^-\right) + \frac{2C\hat{\tau}_k^-}{\sqrt{n}}$$

and

$$\hat{\delta}_k^+ := \overline{\delta} + 8CR_n\left(\hat{\mathcal{F}}_k^+\right) + \frac{12C\hat{\tau}_k^+ + 2C}{\sqrt{n}}$$

respectively. Similarly, in the definitions of $\mathcal{T}_k$, $\mathcal{T}_k^-$, and $\mathcal{T}_k^+$, one uses now

$$\delta_k := \overline{\delta} + 4C\mathbb{E}\Delta_n(\mathcal{F}_k) + \frac{4C\tau_k + 2C}{\sqrt{n}}$$
$$\delta_k^- := \overline{\delta} + C\mathbb{E}\Delta_n\left(\mathcal{F}_k^-\right)$$

and

$$\delta_k^+ := \overline{\delta} + 20C\mathbb{E}\Delta_n\left(\mathcal{F}_k^+\right) + \frac{20C\tau_k^+ + 10C}{\sqrt{n}}$$

respectively (we used the notations $\hat{\tau}_k^- := t_{k\wedge \hat{l}^-}$ and $\tau_k := t_{k\wedge l}$). With such a choice of deltas one can prove results similar to Theorems 3.1 and 3.2. However, relatively large values of the constants involved in the current version of this procedure do not make it very practical.

We would like to emphasize that we do not possess at the moment any ready to be used methodology of dealing with the problem of choosing δ as well as with some other aspects of the implementation of the iterative structural risk minimization method. The aim of the above remarks is to attract attention to this problem to be addressed in the future.

IV. Proofs of the Oracle Inequalities for Iterative Structural Risk Minimization

The following lemma describe the properties of iterative trees and pools and it is the key ingredient of the proofs of Theorems 3.1 and 3.2.

Recall that l, l^-, l^+, $\hat{l}^-$, $\hat{l}^+$ denote the stopping times of the corresponding trees and

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

Lemma 4.1: Let $N \in \{1, \ldots, \infty\}$. Define

$$\begin{aligned} E := \Bigg\{\omega \in \Omega\colon \forall\, k = 1, \ldots, N\colon & \\ & \left|R_n\left(\mathcal{F}_k^+\right) - \mathbb{E}R_n\left(\mathcal{F}_k^+\right)\right| < \frac{\tau_k^+}{\sqrt{n}}, \\ & \left|\Delta_n\left(\mathcal{F}_k^+\right) - \mathbb{E}\Delta_n\left(\mathcal{F}_k^+\right)\right| < \frac{\tau_k^+}{\sqrt{n}}, \\ & \left|R_n(\mathcal{F}_k) - \mathbb{E}R_n(\mathcal{F}_k)\right| < \frac{\tau_k}{\sqrt{n}}\Bigg\}. \end{aligned}$$

Then

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

and on the event E

$$\forall\, k = 1, \ldots, N, \qquad \mathcal{T}_k^- \prec \hat{\mathcal{T}}_k^- \prec \mathcal{T}_k \prec \hat{\mathcal{T}}_k^+ \prec \mathcal{T}_k^+ \tag{4.4}$$
$$\forall\, k = 1, \ldots, N, \qquad \mathcal{F}_k^- \subset \hat{\mathcal{F}}_k^- \subset \mathcal{F}_k \subset \hat{\mathcal{F}}_k^+ \subset \mathcal{F}_k^+. \tag{4.5}$$

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

$$\begin{aligned} E^c := & \bigcup_{k=1}^{l^+}\left\{\left|\Delta_n\left(\mathcal{F}_k^+\right) - \mathbb{E}\Delta_n\left(\mathcal{F}_k^+\right)\right| \ge t_k n^{-1/2}\right\} \\ & \bigcup \bigcup_{k=1}^{l^+}\left\{\left|R_n\left(\mathcal{F}_k^+\right) - \mathbb{E}R_n\left(\mathcal{F}_k^+\right)\right| \ge t_k n^{-1/2}\right\} \\ & \bigcup \bigcup_{k=1}^{l}\left\{\left|R_n(\mathcal{F}_k) - \mathbb{E}R_n(\mathcal{F}_k)\right| \ge t_k n^{-1/2}\right\}. \end{aligned}$$

It follows from Lemmas 2.2 and 2.3 that for all $k = 1, \ldots, l^+$

$$\mathbb{P}\left\{\left|\Delta_n\left(\mathcal{F}_k^+\right) - \mathbb{E}\Delta_n\left(\mathcal{F}_k^+\right)\right| \ge t_k n^{-1/2}\right\} \le 2\exp\left\{-2t_k^2\right\} \tag{4.6}$$
$$\mathbb{P}\left\{\left|R_n\left(\mathcal{F}_k^+\right) - \mathbb{E}R_n\left(\mathcal{F}_k^+\right)\right| \ge t_k n^{-1/2}\right\} \le 2\exp\left\{-t_k^2/2\right\} \tag{4.7}$$

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

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

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

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

and

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

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

$$\begin{aligned} E \subset & \left\{\forall\, k = 1, \ldots, N\colon \mathcal{T}_k^- \prec \hat{\mathcal{T}}_k^- \prec \mathcal{T}_k \prec \hat{\mathcal{T}}_k^+ \prec \mathcal{T}_k^+\right\} \\ & \bigcap \left\{\forall\, k = 1, \ldots, N\colon \mathcal{F}_k^- \subset \hat{\mathcal{F}}_k^- \subset \mathcal{F}_k \subset \hat{\mathcal{F}}_k^+ \subset \mathcal{F}_k^+\right\} \end{aligned}$$

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 Property B 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 $j \le k+1$). Thus, it is enough to consider the case when $l^- \ge k$ and, hence, $k \wedge l^- = k \wedge \hat{l}^- = k$. 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 $k \wedge \hat{l}^{+} \le k \wedge l^{+}$, which in turn implies

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

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

$$P(f_v) \le \min_{\mathcal{F}_k^{+}} P + \delta - 10\mathbb{E}\Delta_n\left(\mathcal{F}_k^{+}\right) - \frac{12\tau_k^{+} + 4}{\sqrt{n}}$$

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

$$\begin{aligned} P_n(f_{\varphi(v)}) &= P_n(f_v) \\ &\le \min_{\mathcal{F}_k^{+}} P_n + \delta - 10\mathbb{E}\Delta_n\left(\mathcal{F}_k^{+}\right) - \frac{12\tau_k^{+} + 4}{\sqrt{n}} \\ &\quad + 2\Delta_n\left(\mathcal{F}_k^{+}\right) \\ &\le \min_{\mathcal{F}_k^{+}} P_n + \delta - 8\mathbb{E}\Delta_n\left(\mathcal{F}_k^{+}\right) - \frac{10\tau_k^{+} + 4}{\sqrt{n}} \\ &\le \min_{\mathcal{F}_k^{+}} P_n + \delta - 4\mathbb{E}R_n\left(\mathcal{F}_k^{+}\right) - \frac{10\tau_k^{+}}{\sqrt{n}} \\ &\le \min_{\mathcal{F}_k^{+}} P_n + \delta - 4R_n\left(\mathcal{F}_k^{+}\right) - \frac{6\tau_k^{+}}{\sqrt{n}} \\ &\le \min_{\hat{\mathcal{F}}_k^{+}} P_n + \delta - 4R_n\left(\hat{\mathcal{F}}_k^{+}\right) - \frac{6\hat{\tau}_k^{+}}{\sqrt{n}}. \end{aligned}$$

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

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

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, $j \wedge \hat{l}^{+} \le j \wedge l^{+} \le j$, $j \le k+1$. If $\hat{l}^{+} \ge k+1$, then $l^{+} \ge k+1$, 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

$$\left\{f_v \colon v \in A\left(\mathcal{E}\left(\hat{\mathcal{T}}_{k+1}^{+}\right)\right)\right\} \subset \left\{f_v \colon v \in A\left(\mathcal{E}\left(\mathcal{T}_{k+1}^{+}\right)\right)\right\}.$$

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 Property B, that $\hat{\mathcal{T}}_j^{+} = \hat{\mathcal{T}}_{\hat{l}^{+}}^{+}$, $j \ge \hat{l}^{+}$. Therefore,

$$\mathcal{E}\left(\hat{\mathcal{T}}_{k+1}^{+}\right) = \mathcal{E}\left(\hat{\mathcal{T}}_k^{+}\right) = \cdots = \mathcal{E}\left(\hat{\mathcal{T}}_{\hat{l}^{+}}^{+}\right)$$

which implies

$$\hat{\mathcal{F}}_{k+1}^{+} = \hat{\mathcal{F}}_k^{+} = \cdots = \hat{\mathcal{F}}_l^{+}.$$

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

$$\hat{\mathcal{F}}_{k+1}^{+} = \hat{\mathcal{F}}_l^{+} \subset \mathcal{F}_l^{+} \subset \mathcal{F}_{k+1}^{+}. \qquad \square$$

Proof of Theorem 3.2: On the event E (see Lemma 4.1), 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

$$\tau_N = t_{N \wedge l} \le \hat{\tau}_N^{+} = t_{N \wedge \hat{l}^{+}} \le \tau_N^{+} = t_{N \wedge l^{+}}.$$

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

$$\begin{aligned} P\left(\tilde{f}_N\right) &\le P_n\left(\tilde{f}_N\right) + \Delta_n\left(\hat{\mathcal{F}}_N^{-}\right) \\ &\le \min_{\hat{\mathcal{F}}_N^{-}} P_n + \Delta_n(\mathcal{F}_N) \\ &\le \min_{\hat{\mathcal{F}}_N^{-}} P_n + \mathbb{E}\Delta_n(\mathcal{F}_N) + \frac{\tau_N}{\sqrt{n}} \\ &\le \min_{\hat{\mathcal{F}}_N^{-}} P_n + 2\mathbb{E}R_n(\mathcal{F}_N) + \frac{\tau_N}{\sqrt{n}} \\ &\le \min_{\hat{\mathcal{F}}_N^{-}} P_n + 2R_n\left(\hat{\mathcal{F}}_N\right) + \frac{3\tau_N}{\sqrt{n}} \\ &\le \min_{\hat{\mathcal{F}}_N^{-}} P_n + 2R_n\left(\hat{\mathcal{F}}_N^{+}\right) + \frac{3\hat{\tau}_N^{+}}{\sqrt{n}} \end{aligned}$$

which, by Lemma 4.1, implies (3.1). Similarly, we have

$$\begin{aligned} P\left(\tilde{f}_N\right) &\le P_n\left(\tilde{f}_N\right) + \Delta_n\left(\hat{\mathcal{F}}_N^{-}\right) \\ &\le \min_{\hat{\mathcal{F}}_N^{-}} P_n + \Delta_n\left(\hat{\mathcal{F}}_N^{-}\right) \\ &\le \min_{\hat{\mathcal{F}}_N^{-}} P + 2\Delta_n\left(\hat{\mathcal{F}}_N^{-}\right) \\ &\le \min_{\mathcal{F}_N^{-}} P + 2\Delta_n(\mathcal{F}_N) \\ &\le \min_{\hat{\mathcal{F}}_N^{-}} P + 2\mathbb{E}\Delta_n\left(\mathcal{F}_N^{+}\right) + \frac{2\tau_N^{+}}{\sqrt{n}} \end{aligned}$$

which implies (3.2) by Lemma 4.1.

Proof of Theorem 3.3: 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

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

Hence, we get

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

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

$$\begin{aligned} \Delta_n(\mathcal{F}_k) &\le \mathbb{E}\Delta_n(\mathcal{F}_k) + \frac{\tau_k}{\sqrt{n}} \\ &\le 2\mathbb{E}R_n(\mathcal{F}_k) + \frac{\tau_k}{\sqrt{n}} \\ &\le 2R_n(\mathcal{F}_k) + \frac{3\tau_k}{\sqrt{n}} \end{aligned}$$

we get

$$\begin{aligned} P\left(\hat{f}_N\right) &\le \min_{\hat{\mathcal{F}}_{\hat{k}}^{-}} P_n + 2R_n(\mathcal{F}_{\hat{k}}) + \frac{3\tau_{\hat{k}}}{\sqrt{n}} \\ &\le \min_{\hat{\mathcal{F}}_{\hat{k}}^{-}} P_n + 2R_n\left(\hat{\mathcal{F}}_{\hat{k}}^{+}\right) + \frac{3\hat{\tau}_{\hat{k}}^{+}}{\sqrt{n}} \\ &\le \inf_{1 \le k \le N}\left[\min_{\hat{\mathcal{F}}_k^{-}} P_n + 2R_n\left(\hat{\mathcal{F}}_k^{+}\right) + \frac{3\hat{\tau}_k^{+}}{\sqrt{n}}\right] + \sigma \end{aligned}$$

and (3.3) follows by Lemma 4.1.

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

$$\begin{aligned}
&\inf_{1\le k\le N}\left[\min_{\hat{\mathcal{F}}_k^-} P_n + 2R_n\left(\hat{\mathcal{F}}_k^+\right) + \frac{3\hat{\tau}_k^+}{\sqrt{n}}\right]\\
&\quad\le \inf_{1\le k\le N}\left[\min_{\mathcal{F}_k^-} P + \Delta_n(\mathcal{F}_k^+) + 2R_n(\mathcal{F}_k^+) + \frac{3\tau_k^+}{\sqrt{n}}\right]\\
&\quad\le \inf_{1\le k\le N}\left[\min_{\mathcal{F}_k^-} P + \mathbb{E}\Delta_n(\mathcal{F}_k^+) + 2\mathbb{E}R_n(\mathcal{F}_k^+) + \frac{6\tau_k^+}{\sqrt{n}}\right]\\
&\quad\le \inf_{1\le k\le N}\left[\min_{\mathcal{F}_k^-} P + 5\mathbb{E}\Delta_n(\mathcal{F}_k^+) + \frac{6\tau_k^+ + 2}{\sqrt{n}}\right]
\end{aligned}$$

where we used the bound of Lemma 2.5. □

Acknowledgment

The author wishes to thank C. Abdallah, D. Hush, and C. Scovel for interesting discussions. C. Scovel and D. Hush also suggested several improvements and corrections. F. Lozano pointed out several typos. The anonymous referees made a number of comments and suggestions that substantially improved the paper.

References

[1] V. N. Vapnik and A. Ya. Chervonenkis, " On the uniform convergence of relative frequencies of events to their probabilities.," *Theory Probab. its Applic.*, vol. 16, pp. 264–280, 1971.

[2] ——, *Theory of Pattern Recognition* Nauka, U.S.S.R., 1974.

[3] ——, *Estimation of Dependencies Based on Empirical Data*. New York: Springer-Verlag, 1982.

[4] ——, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[5] ——, *Statistical Learning Theory*. New York: Wiley, 1998.

[6] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[7] M. Vidyasagar, *A Theory of Learning and Generalization*. New York: Springer-Verlag, 1997.

[8] L. Valiant, "A theory of learnable," *Commun. ACM*, vol. 27, pp. 1134–1142, 1984.

[9] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the Vapnik–Chervonenkis dimension," *J. Assoc. Comput. Mach.*, vol. 36, pp. 929–965, 1989.

[10] G. Lugosi and K. Zeger, "Concept learning using complexity regularization," *IEEE Trans. Inform. Theory*, vol. 42, pp. 48–54, Jan. 1996.

[11] L. Birgé and P. Massart, "From model selection to adaptive estimation," in *Festschrift for L. Le Cam*, D. Pollard, E. Torgersen, and G. Yang, Eds. New York: Springer, 1997, Research Papers in Probability and Statistics, pp. 55–87.

[12] A. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Probab. Theory Related Fields*, vol. 113, pp. 301–413, 1999.

[13] M. Talagrand, "A new look at independence," *Ann. Probab.*, vol. 24, pp. 1–34, 1996.

[14] ——, "New concentration inequalities in product spaces," *Invent. Math.*, vol. 126, pp. 505–563, 1996.

[15] M. Ledoux. (1996) On Talagrand's deviation inequalities for product measures. *ESAIM: Probab. Statist.* [Online], pp. 63–87, vol. 1, 1996. Available: http://www.emath.fr/ps/

[16] P. Massart, "About the constants in Talagrand's concentration inequalities for empirical processes," *Ann. Probab.*, vol. 28, no. 2, pp. 863–884, 2000.

[17] I. M. Johnstone, "Oracle inequalities and nonparametric function estimation," in *Documenta Mathematica, Journal der Deutschen Mathematiker Vereinigung (Proc. Int. Congr. Mathematicians)*, vol. III, Berlin, Germany, 1998, pp. 267–278.

[18] V. I. Koltchinskii, "On the central limit theorem for empirical measures," *Probab. Theory Math. Statist.*, vol. 24, pp. 71–82, 1981.

[19] D. Pollard, "A central limit theorem for empirical processes," *J. Austral. Math. Soc.*, vol. A39, pp. 235–248, 1982.

[20] E. Giné and J. Zinn, "Some limit theorems for empirical processes," *Ann. Probab.*, vol. 12, pp. 929–989, 1984.

[21] R. M. Dudley, *Uniform Central Limit Theorems*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[22] A. van der Vaart and J. Wellner, *Weak Convergence and Empirical Processes. With Applications to Statistics*. New York: Springer-Verlag, 1996.

[23] V. Koltchinskii, C. T. Abdallah, M. Ariola, P. Dorato, and D. Panchenko, "Improved sample complexity estimates for statistical learning control of uncertain systems," *IEEE Trans. Automat. Contr.*, to be published.

[24] F. Lozano, "Model Selection using Rademacher Penalization," in *Proc. 2nd ICSC Symp. Neural Computation NC2000*. Berlin, Germany: ICSC Academic, 2000.

[25] P. Bartlett, S. Boucheron, and G. Lugosi, "Model selection and error estimation," preprint, 2000.

[26] S. Boucheron, G. Lugosi, and P. Massart, "A sharp concentration inequality with applications in random combinatorics and learning," *Random Structures and Algorithms*, vol. 16, pp. 277–292, 2000.

[27] V. Koltchinskii and D. Panchenko, "Rademacher processes and bounding the risk of function learning," in *High Dimensional Probability II*, E. Giné, D. Mason, and J. Wellner, Eds. Boston, MA: Birkhäuser, 2000, vol. 47, Progress in Probability, pp. 443–457.

[28] K. Buescher and P. R. Kumar, "Learning by canonical smooth estimation—Part II: Learning and choice of model complexity," *IEEE Trans. Automat. Contr.*, vol. 41, pp. 557–569, 1996.

[29] J. Shawe-Taylor, P. Bartlett, R. C. Williamson, and M. Anthony, "Structural risk minimization over data-dependent hierarchies," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1926–1940, Sept. 1998.

[30] Y. Freund, "Self bounding learning algorithms," in *Proc. 11th Annu. Conf. Computational Learning Theory, COLT 98*, 1998, pp. 247–258.

[31] K. Azuma, "Weighted sums of certain dependent random variables," *Tokuku Math. J.*, vol. 19, pp. 357–367, 1967.

[32] V. Yurinski, "Exponential bounds for large deviations," *Theory Probab. its Applic.*, vol. 19, pp. 154–155, 1974.

[33] V. I. Koltchinskii, "Functional limit theorems and empirical entropy. I," *Probab. Theory Math. Statist.*, vol. 33, pp. 31–42, 1985.

[34] ——, "Functional limit theorems and empirical entropy. II," *Probab. Theory Math. Statist.*, vol. 34, pp. 73–85, 1986.

[35] V. Milman and G. Schechtman, "Asymptotic theory of finite dimensional normed spaces," in *Lecture Notes in Mathematics*. New York: Springer-Verlag, 1986, vol. 1200.

[36] W. T. Rhee and M. Talagrand, "Martingale inequalities and NP-complete problems," *Math. Oper. Res.*, vol. 12, pp. 177–181, 1987.

[37] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics (London Mathematical Soc. Lecture Notes)*, J. Siemons, Ed. Cambridge, U.K.: Cambridge Univ. Press, 1989, vol. 141, pp. 148–188.

[38] D. Hush and C. Scovel, private communication, 1999.

[39] J. Langford and A. Blum, "Microchoice bounds and self bounding learning algorithms," in *Proc. 12th Annu. Conf. Computational Learning Theory, COLT 99*, 1999, pp. 209–214.