Journal of Machine Learning Research 00 (2010) 00-00 Submitted 00/00; Published 00/00

Rademacher Complexities and Bounding the Excess Risk in Active Learning

Vladimir Koltchinskii * VLAD@MATH.GATECH.EDU
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0160, USA

Editor: Gabor Lugosi

Abstract

Sequential algorithms of active learning based on the estimation of the level sets of the empirical risk are discussed in the paper. Localized Rademacher complexities are used in the algorithms to estimate the sample sizes needed to achieve the required accuracy of learning in an adaptive way. Probabilistic bounds on the number of active examples have been proved and several applications to binary classification problems are considered.

Keywords: active learning, excess risk, Rademacher complexities, capacity function, disagreement coefficient

1. Introduction

Let $(S, \mathcal{A})$ be a measurable space and $T \subset \mathbb{R}$. Consider a standard **prediction problem** in which (X, Y) is a random couple in $S \times T$ with unknown distribution P. Here X is a **design point** whose distribution will be denoted Π and Y is a **response variable** with conditional distribution (given X) $P_{Y|X}(\cdot|X = x)$. The response variable Y is to be predicted based on an observation of X. This class of problems includes many versions of regression and classification. For instance, in the binary classification, $T = \{-1, 1\}$ and the conditional distribution of Y given X is completely characterized by the regression function

$$\eta(x) := \mathbb{E}(Y|X = x).$$

In the framework of **passive learning**, a learning algorithm inputs the training data $(X_1, Y_1), \ldots (X_n, Y_n)$ that consists of n independent examples sampled from the distribution P. The goal is to construct a data dependent prediction rule $\hat{g} : S \mapsto T$ whose risk with respect to a properly chosen loss function is "close" to the minimal possible risk. More precisely, given a loss function $\ell : T \times T \mapsto \mathbb{R}$, the risk of a prediction rule $g : S \mapsto T$ is defined as

$$P(\ell \bullet g) = \mathbb{E}\ell(Y; g(X)),$$

where we used the notation $(\ell \bullet g)(x, y) := \ell(y; g(x))$. For instance, in the binary classification setting, the binary loss $\ell(y, u) := I(y \neq u)$ is usually used. In this case, the risk of a classification rule $g : S \mapsto \{-1, 1\}$ is its generalization error:

$$P(\ell \bullet g) = \mathbb{P}\{Y \neq g(X)\}.$$

∗. Partially supported by NSF grants MSPA-MCS-0624841, DMS-0906880 and CCF-0808863

Suppose $\mathcal{G}$ is a class of prediction rules $g : S \mapsto T$. The quantity

$$\mathcal{E}_P(\ell \bullet g) := P(\ell \bullet g) - \inf_{g \in \mathcal{G}} P(\ell \bullet g)$$

is called the **excess risk** of g. One of the main goals of statistical analysis of learning algorithms is to understand how the excess risk $\mathcal{E}_P(\ell \bullet \hat{g})$ of a data dependent decision rule $\hat{g}$ output by such an algorithm depends on the sample size n, on the "complexity" of the class $\mathcal{G}$ of prediction rules and on the underlying complexity of the prediction problem itself.

In the recent years, there has been a lot of interest in **active learning** algorithms. In this framework, the algorithm can modify the design distribution in the process of learning. More precisely, suppose that the training examples (X_j, Y_j) are sampled sequentially. At each iteration (say, iteration number k), the algorithm requests a design point X_{k+1} sampled from a distribution $\hat{\Pi}_k$ that depends on the training data $(X_1, Y_1), \dots, (X_k, Y_k)$. Given $X_{k+1} = x$, the response variable (the label) Y_{k+1} is sampled from the conditional distribution $P_{Y|X}(\cdot | X = x)$. The question is whether there are such active learning algorithms for which the excess risk after n iterations is provably smaller than for the passive prediction rules based on the same number n of training examples. It happens that the answer depends on the type of learning problem. A minimax analysis by Castro, Willett and Nowak (2005) and Castro and Nowak (2008) shows that such an improvement is possible in classification problems and in some special classes of regression problems with non-smooth regression function (for instance, if the regression function is a step function). In such cases, the improvement can be very significant. In some classification problems, the excess risk of active learning algorithms can converge to zero with an **exponential rate** as $n \to \infty$ (comparing with the rate $O(n^{-1})$ in the case of passive learning). Castro and Nowak (2008) studied several examples of binary classification problems in which the active learning approach is beneficial and suggested nice active learning algorithms in these problems. However, the drawback of these algorithms is that they are not adaptive in the sense that they require the prior knowledge of distribution dependent parameters of the problem, such as noise characteristics in classification. The development of active learning methods that are adaptive and, at the same time, computationally tractable remains a challenge. There has been a progress in the design of active learning algorithms that possess some degree of adaptivity, in particular, see Dasgupta, Hsu and Monteleoni (2007), Balcan, Beygelzimer and Langford (2009), Balcan, Hanneke and Wortman (2008) and Hanneke (2009a, 2009b). In the last two interesting papers by Hanneke, some versions of the algorithms of Balcan, Beygelzimer and Langford (2009) and by Dasgupta, Hsu and Monteleoni (2007) were studied using the technique of **Rademacher complexities** that much earlier proved to be very useful in the analysis of passive learning (see Bartlett, Boucheron and Lugosi (2002), Koltchinskii (2001), Koltchinskii and Panchenko (2000), Bartlett, Bousquet and Mendelson (2005), Koltchinskii (2006, 2008) and references therein). Hanneke showed that incorporating Rademacher complexities in active learning algorithms allows one to develop rather general versions of such algorithms that are adaptive under broad assumptions on the underlying distributions.

In the current paper, we continue this line of research. We consider the following model of active learning. At each iteration, a learning algorithm has to choose a set $\hat{A} \subset S$ of "good" design points and also the number of training examples needed at the current iteration. Both the set $\hat{A}$ and the required number of the examples might depend on the training data that is already available. The algorithm has an access to an oracle that is asked to provide the required number of examples (X, Y) sampled from the conditional distribution $P(\cdot|x \in \hat{A})$. Alternatively, it can be described as follows. The oracle provides training examples (X, Y) sampled from an unknown probability distribution P. At each iteration, the algorithm chooses a set $\hat{A}$ of "good" design points and asks the oracle whether the next design point X_k is "good". If it is "good", the algorithm accepts the point, the oracle provides it with a label Y_k and returns the couple (X_k, Y_k). Testing whether the example is "good" costs the algorithm nothing, but each "good" labeled training example costs \$ 1. Thus, only the number of "good" examples matters for determining the total cost of learning. The question is how many "good" examples are needed for the excess risk $\mathcal{E}_P(\ell \bullet \hat{g})$ of the resulting classifier $\hat{g}$ to become smaller than δ with a guaranteed probability at least $1 - \alpha$.

We will develop active learning algorithms that are somewhat akin to what is done in Hanneke (2009a, 2009b), but they are more closely related to the construction of localized Rademacher complexities used in the definitions of distribution dependent and data dependent excess risk bounds in empirical risk minimization (see Koltchinskii (2006, 2008)). The main idea of this construction is to characterize the complexity of the problem by the sup-norm of a special Rademacher process indexed by the level sets of the risk. To be more specific, suppose that we are dealing with a binary classification problem and that the empirical risk (with respect to the binary loss) is being minimized over a class $\mathcal{G}$ of binary functions. Then what matters is the collection of δ-minimal sets

$$\mathcal{G}(\delta) := \Big\{ g \in \mathcal{G} : \mathcal{E}_P(\ell \bullet g) \leq \delta \Big\}, \ \delta > 0.$$

These sets can be estimated based on the empirical data and Rademacher complexities of such estimated sets for small enough values of δ are used to define reasonably tight bounds on the excess risk. In many learning problems, the δ-minimal sets become small as $\delta \to 0$, for instance, in the sense that their $L_2(\Pi)$-diameter is small. It turns out that an important role in the development of active classification algorithms is played by the sets of the following type

$$A(\delta) := \Big\{ x \in S | \exists g_1, g_2 \in \mathcal{G}(\delta) : g_1(x) \neq g_2(x) \Big\}.$$

Such a set is called a **disagreement set** since it consists of the points for which there are two classifiers in $\mathcal{G}(\delta)$ whose predictions at point x disagree with each other. If the δ-minimal sets are small for small enough values of δ, one can expect that the corresponding disagreement sets are also small. This is not always the case, but there are natural examples in which indeed the measure $\Pi(A(\delta))$ tends to 0 as $\delta \to 0$ (sometimes, even $\Pi(A(\delta)) = O(\delta)$). Note that if the empirical risk is being minimized over the δ-minimal set $\mathcal{G}(\delta)$, one can eliminate from the sample all the design points X_j such that $X_j \notin A(\delta)$: the minimizers of the empirical risk are not going to change since the value of the binary loss at such training examples (X_j, Y_j) is a constant on the δ-minimal set. So, only the examples for which X_j belongs to a small disagreement set $A(\delta)$ are really needed. This simple observation opens a possibility of reducing the sample size in the process of active learning, and this has already been exploited in several algorithms described in the literature (see Dasgupta, Hsu and Monteleoni (2007), Balcan, Beygelzimer and Langford (2009), Hanneke (2009a, 2009b) and references therein). It is interesting to mention that some notions similar to the "disagreement sets" were used much earlier in the study of ratio type empirical processes (for instance, in the work of Alexander in the 80s; see, Giné and Koltchinskii (2006) and references therein). Moreover, it was used in Giné and Koltchinskii (2006) to obtain refined excess risk bounds in binary classification (in the passive learning case). This will be discussed in some detail in Section 4.

Our approach is based on iterative estimation of the δ-minimal sets for a decreasing sequence $\{\delta_j\}$ of values of δ. It happens that, for larger values of δ, it is possible to construct a rough estimate of the δ-minimal sets based on a relatively small number of training examples. The required sample sizes $\bar{n}(\delta)$ can be estimated using Rademacher complexities. For smaller values of δ, more examples are needed, but, at the same time, for the smaller values of δ the disagreement sets are also small, and these sets again can be estimated based on the training examples that have been already sampled. Thus, there is a possibility to come up with an active learning strategy that, at each iteration, computes an estimate $\hat{A}$ of the disagreement set and determines the required sample size, and then samples the required number of design points from the conditional distribution $\Pi(\cdot|x \in \hat{A})$. Each of these points $X_j = x$ is provided with a label Y_j sampled from the conditional distribution $P_{Y|X}(\cdot|X = x)$. The algorithm stops as soon as δ_j becomes smaller than the required accuracy of learning δ. At this stage, it outputs an estimate of the δ-minimal set $\mathcal{G}(\delta)$. The number of labeled examples needed to achieve this goal is rougly

$$\sum_{\delta_j \geq \delta} \bar{n}(\delta_j)\Pi(A(\delta_j))$$

(see theorems 7, 8, 9 for more precise formulations). For binary classification problems with a VC-class $\mathcal{G}$ such that $\Pi(A(\delta)) = O(\delta)$, this leads to the bound on the number of labeled examples of the order $\log(1/\delta)\log\log(1/\delta)$.

The need to estimate the whole δ-minimal set in this learning strategy rather than simply minimizing the empirical risk might look like too strong of an assumption. However, in the alternative general versions of adaptive strategies of active learning due to Hanneke (2009a, 2009b) this is also needed. Hanneke uses previously suggested agnostic learning methods of Dasgupta, Hsu and Monteleoni (2007) and of Balcan, Beygelzimer and Langford (2009) in combination with Rademacher complexities that are based on estimated level sets of the empirical risk. So, currently, this seems to be unavoidable in general adaptive methods and our approach is just based on a more direct use of the δ-minimal sets.

To give a precise description of a version of active learning method considered in this paper and to study its statistical properties, several facts from the general theory of empirical risk minimization will be needed. In particular, in Section 2, we describe a construction of distribution dependent and data dependent bounds on the excess risk based on localized Rademacher complexities (see Koltchinskii (2006, 2008)). In Section 3, we describe our active learning algorithms. These algorithms are sequential in the sense that the training data is being sampled until the desired accuracy of learning is achieved. We prove several bounds on the number of active examples needed to achieve this goal with a specified probability. In sections 2 and 3, it is convenient to study the problem in a more abstract framework, in which we suppress the labels Y_j and write S instead of $S \times \{-1, 1\}$, X instead of (X, Y),

f instead of $\ell \bullet g$, etc. This allows us to simplify the notations and the description of the algorithms, and, at the same time, it makes the results a little more general. In principle, it should be possible to apply these results to more general classes of learning problems than binary classification (for instance, to special regression models with a non-smooth regression function or to the problem of estimation of the level sets of an unknown probability density). However, we do not pursue this possibility here and, instead, we concentrate in Section 4 on the binary classification problems, which still remains the most interesting class of learning problems where the active learning approach leads to faster convergence rates.

2. Empirical Risk Minimization: Bounds on the Excess Risk

Let $(S, \mathcal{A})$ be a measurable space, let P be a probability measure in $(S, \mathcal{A})$ and let $X, X_1, X_2, \ldots$ be i.i.d. random variables in $(S, \mathcal{A})$ with common distribution P. Let $\mathcal{F}$ be a class of $\mathcal{A}$-measurable functions $f : S \mapsto [0,1]$. The values of functions $f \in \mathcal{F}$ will be interpreted as "losses" associated with some decisions and the integral

$$Pf := \int_S f dP = \mathbb{E} f(X)$$

represents the expected loss, or the (true) risk. The optimization problem

$$Pf \longrightarrow \min, \ f \in \mathcal{F} \tag{1}$$

is interpreted as a "risk minimization" problem and the quantity

$$\mathcal{E}_P(f) := Pf - \inf_{g \in \mathcal{F}} Pg$$

is called the excess risk of f. The δ-minimal set of the true risk is defined as

$$\mathcal{F}_P(\delta) := \Big\{ f \in \mathcal{F} : \mathcal{E}_P(f) \leq \delta \Big\}, \ \delta \geq 0.$$

In learning theory problems, the distribution P is usually unknown and the risk Pf is to be estimated by the empirical risk. The empirical measure based on the sample $(X_1, \ldots, X_n)$ of size n is defined as

$$P_n := n^{-1} \sum_{j=1}^{n} \delta_{X_j}$$

and the problem of risk minimization is replaced by the "empirical risk minimization":

$$P_n f \longrightarrow \min, \ f \in \mathcal{F}. \tag{2}$$

Naturally, this also leads to the definitions of the "excess empirical risk" $\mathcal{E}_{P_n}(f)$ and of the δ-minimal sets of the empirical risk $\mathcal{F}_{P_n}(\delta), \ \delta > 0$.

Given a solution $\hat{f}$ of the empirical risk minimization problem (2), a basic question is to provide reasonably tight upper confidence bounds on the excess risk $\mathcal{E}_P(\hat{f})$ that depend on complexity characteristics of the class $\mathcal{F}$. It is also of importance to understand when the δ-minimal sets of the empirical risk are reasonably good estimates of the δ-minimal sets of the true risk. We will need below several results of this type that can be found in Koltchinskii (2006, 2008).

First of all, we will need an upper confidence bound on the size of the empirical process

$$\sup_{f,g\in\mathcal{F}_P(\delta)} |(P_n - P)(f-g)|.$$

To construct such a bound, we use Talagrand's famous concentration inequalities. Suppose $\rho_P : L_2(P)\times L_2(P) \mapsto [0,+\infty)$ and

$$\rho_P^2(f,g) \geq P(f-g)^2 - (P(f-g))^2,\ \ f,g\in L_2(P).$$

Define the diameter of $\mathcal{F}_P(\delta)$ as

$$D(\delta) := D_P(\delta) := \sup_{f,g\in\mathcal{F}(\delta)} \rho_P(f,g).$$

It provides a measure of the size of the δ-minimal sets. We will also use the following quantity that characterizes the accuracy of "empirical approximation" of P by P_n on the δ-minimal sets:

$$\phi_n(\delta) := \mathbb{E} \sup_{f,g\in\mathcal{F}(\delta)} |(P_n - P)(f-g)|.$$

Given a decreasing sequence $\{\delta_j\}$ of positive numbers with $\delta_0 := 1$ and a sequence $\{t_j\}$ of positive numbers, define a step function $U_n(\delta),\ \delta\in(0,1]$ as follows:

$$\bar{U}_n(\delta) := 2\sum_{j\geq 0}\left[\phi_n(\delta_j) + D(\delta_j)\sqrt{\frac{t_j}{n}} + \frac{t_j}{n}\right] I_{(\delta_{j+1},\delta_j]}(\delta).$$

A version of Talagrand's concentration inequality with explicit constants due to Bousquet implies that, for all $j\geq 0$ and for all $\delta\in(\delta_{j+1},\delta_j]$, with probability at least $1-e^{-t_j}$

$$\sup_{f,g\in\mathcal{F}_P(\delta)} |(P_n - P)(f-g)| \leq \bar{U}_n(\delta).$$

Given $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$, define

$$\psi^\flat(\delta) := \sup_{\sigma\geq\delta} \frac{\psi(\sigma)}{\sigma}$$

and

$$\psi^\sharp(\varepsilon) := \inf\Big\{\delta > 0 : \psi^\flat(\delta) \leq \varepsilon\Big\}.$$

Let

$$\delta_n(\mathcal{F};P) := \sup\Big\{\delta\in(0,1] : \delta \leq \bar{U}_n(\delta)\Big\}.$$

The following bounds were proved in Koltchinskii (2006, 2008).

Theorem 1 *For all* $\delta \geq \delta_n(\mathcal{F}; P)$,

$$\mathbb{P}\Big\{\mathcal{E}_P(\hat{f}) > \delta\Big\} \leq \sum_{\delta_j \geq \delta} e^{-t_j}$$

and

$$\mathbb{P}\left\{\sup_{f \in \mathcal{F}, \mathcal{E}_P(f) \geq \delta} \left|\frac{\mathcal{E}_{P_n}(f)}{\mathcal{E}_P(f)} - 1\right| > \bar{U}_n^\flat(\delta)\right\} \leq \sum_{\delta_j \geq \delta} e^{-t_j}.$$

Thus, the quantity $\delta_n(\mathcal{F}; P)$ is a distribution dependent upper bound on the excess risk $\mathcal{E}_P(\hat{f})$ that holds with a guaranteed probability. Moreover, for all $\delta \geq \delta_n(\mathcal{F}; P)$ and for all $f \in \mathcal{F}$ with $\mathcal{E}_P(f) \geq \delta$ it is possible to control the size of the ratio $\frac{\mathcal{E}_{P_n}(f)}{\mathcal{E}_P(f)}$ in terms of the quantity $\bar{U}_n^\flat(\delta)$. This ratio bound for the excess risk immediately implies the following statement showing that for all the values of δ above certain threshold the δ-minimal sets of empirical risk provide estimates of the δ-minimal sets of the true risk.

Proposition 2 *Let* $\bar{\delta}_n := \bar{U}_n^\sharp\Big(\frac{1}{2}\Big)$. *For all* $\delta \geq \bar{\delta}_n$, *with probability at least*

$$1 - \sum_{\delta_j \geq \delta} e^{-t_j}$$

the following inclusions hold:

$$\forall \sigma \geq \delta \quad \mathcal{F}_P(\sigma) \subset \mathcal{F}_{P_n}(3/2\sigma) \quad \text{and} \quad \mathcal{F}_{P_n}(\sigma) \subset \mathcal{F}_P(2\sigma).$$

Data dependent upper confidence bounds on the excess risk can be constructed using localized sup-norms of Rademacher processes that provide a way to estimate the size of the empirical process. Given i.i.d. Rademacher random variables $\{\varepsilon_i\}$ independent of $\{X_i\}$, the Rademacher process is defined as

$$R_n(f) := n^{-1} \sum_{j=1}^{n} \varepsilon_j f(X_j).$$

We will assume that

$$\rho_P^2(f, g) := P(f - g)^2.$$

Define

$$\hat{\phi}_n(\delta) := \sup_{f, g \in \mathcal{F}_{P_n}(\delta)} |R_n(f - g)|$$

and

$$\hat{D}_n(\delta) := \sup_{f, g \in \mathcal{F}_{P_n}(\delta)} \rho_{P_n}(f, g).$$

These quantities are empirical versions of $\phi_n(\delta)$ and $D_P(\delta)$ and they can be used to define an empirical version of the function $\bar{U}_n$:

$$\hat{U}_n(\delta) := \hat{K} \sum_{j \geq 0} \left[\hat{\phi}_n(\hat{c}\delta_j) + \hat{D}_n(\hat{c}\delta_j)\sqrt{\frac{t_j}{n}} + \frac{t_j}{n}\right] I_{(\delta_{j+1}, \delta_j]}(\delta),$$

where $\hat{K}, \hat{c}$ are numerical constants. We will also define

$$\tilde{U}_n(\delta) := \tilde{K}\sum_{j\geq 0}\left[\phi_n(\tilde{c}\delta_j) + D(\tilde{c}\delta_j)\sqrt{\frac{t_j}{n}} + \frac{t_j}{n}\right]I_{(\delta_{j+1},\delta_j]}(\delta)$$

with some numerical constants $\tilde{K}, \tilde{c}$.

It can be shown (see Koltchinskii (2006, 2008)) that for large enough numerical constants $\hat{K}, \tilde{K}, \hat{c}, \tilde{c}$ and for all $\delta \geq \bar{\delta}_n$, $\bar{U}_n(\delta) \leq \hat{U}_n(\delta) \leq \tilde{U}_n(\delta)$ with a high probability. More precisely, the following statement holds. Denote

$$\hat{\delta}_n := \hat{U}_n^{\sharp}\left(\frac{1}{2}\right),\ \tilde{\delta}_n := \tilde{U}_n^{\sharp}\left(\frac{1}{2}\right).$$

Theorem 3 *There exists a choice of numerical constants $\hat{K}, \tilde{K}, \hat{c}, \tilde{c}$ in the definitions of the functions $\hat{U}_n, \tilde{U}_n$ such that the following holds. For all $\delta \geq \bar{\delta}_n$, there exists an event E of probability*

$$\mathbb{P}(E) \geq 1 - 3\sum_{\delta_j \geq \delta} e^{-t_j}$$

such that on this event

$$\bar{U}_n(\sigma) \leq \hat{U}_n(\sigma) \leq \tilde{U}_n(\sigma),\ \sigma \geq \delta.$$

As a consequence,

$$\mathbb{P}\left\{\bar{\delta}_n \leq \hat{\delta}_n \leq \tilde{\delta}_n\right\} \geq 1 - 3\sum_{\delta_j \geq \bar{\delta}_n} e^{-t_j}.$$

The proof is based on the following "statistical version" of Talagrand's concentration inequality (which, in turn, follows from the usual Talagrand's concentration inequality for empirical processes and standard symmetrization and contraction arguments, see Koltchinskii (2008)). Suppose that $\mathcal{F}$ is a class of measurable functions on S uniformly bounded by $U > 0$. Denote

$$\sigma_P^2(\mathcal{F}) := \sup_{f\in\mathcal{F}} Pf^2 \text{ and } \sigma_n^2(\mathcal{F}) := \sup_{f\in\mathcal{F}} P_n f^2.$$

For a function $Y : \mathcal{F} \mapsto \mathbb{R}$, denote

$$\|Y\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |Y(f)|.$$

Theorem 4 *There exists a numerical constant $K > 0$ such that for all $t \geq 1$ with probability at least $1 - e^{-t}$ the following bounds hold:*

$$\left|\|R_n\|_{\mathcal{F}} - \mathbb{E}\|R_n\|_{\mathcal{F}}\right| \leq K\left[\sqrt{\frac{t}{n}\left(\sigma_n^2(\mathcal{F}) + U\|R_n\|_{\mathcal{F}}\right)} + \frac{tU}{n}\right],$$

$$\mathbb{E}\|R_n\|_{\mathcal{F}} \leq K\left[\|R_n\|_{\mathcal{F}} + \sigma_n(\mathcal{F})\sqrt{\frac{t}{n}} + \frac{tU}{n}\right],$$

$$\sigma_P^2(\mathcal{F}) \leq K\biggl(\sigma_n^2(\mathcal{F}) + U\|R_n\|_{\mathcal{F}} + \frac{tU}{n}\biggr)$$

and

$$\sigma_n^2(\mathcal{F}) \leq K\biggl(\sigma_P^2(\mathcal{F}) + U\|R_n\|_{\mathcal{F}} + \frac{tU}{n}\biggr).$$

Also,

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}} \leq K\biggl[\|R_n\|_{\mathcal{F}} + \sigma_n(\mathcal{F})\sqrt{\frac{t}{n}} + \frac{tU}{n}\biggr]$$

and

$$\biggl|\|P_n - P\|_{\mathcal{F}} - \mathbb{E}\|P_n - P\|_{\mathcal{F}}\biggr| \leq K\biggl[\sqrt{\frac{t}{n}\biggl(\sigma_n^2(\mathcal{F}) + U\|R_n\|_{\mathcal{F}}\biggr)} + \frac{tU}{n}\biggr].$$

In what follows, it will be of interest to consider sequential learning algorithms in which the sample size is being gradually increased until the excess risk becomes smaller than a given level δ. The following quantities are used in the analysis of such algorithms. Let us fix a set $M \subset \mathbb{N}$. A possible choice is $M = \mathbb{N}$, but, usually, we will take $M = \{2^k : k \geq 0\}$. Denote

$$\bar{n}(\delta) := \inf\biggl\{n \in M : \bar{\delta}_n \leq \delta\biggr\} = \inf\biggl\{n \in M : \bar{U}_n^\flat(\delta) \leq \frac{1}{2}\biggr\},$$

$$\hat{n}(\delta) := \inf\biggl\{n \in M : \hat{\delta}_n \leq \delta\biggr\} = \inf\biggl\{n \in M : \hat{U}_n^\flat(\delta) \leq \frac{1}{2}\biggr\}$$

and

$$\tilde{n}(\delta) := \inf\biggl\{n \in M : \tilde{\delta}_n \leq \delta\biggr\} = \inf\biggl\{n \in M : \tilde{U}_n^\flat(\delta) \leq \frac{1}{2}\biggr\}.$$

If

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}} \to 0 \text{ as } n \to \infty,$$

which is true for so called Glivenko-Cantelli classes of functions with respect to P (see, e.g., van der Vaart and Wellner (1996)), then it is easy to see that

$$\bar{\delta}_n \to 0 \text{ and } \tilde{\delta}_n \to 0 \text{ as } n \to \infty.$$

In this case, we have

$$\tilde{n}(\delta) < +\infty, \bar{n}(\delta) < +\infty, \delta \in (0,1].$$

It is also easy to see that the functions $n \mapsto \bar{U}_n(\delta)$ and $n \mapsto \tilde{U}_n(\delta)$ are noincreasing (it follows from the well known reverse supermartingale properties of empirical processes; see, van der Vaart and Wellner (1996), Lemma 2.4.5). This implies that, for all $n \geq \bar{n}(\delta)$, $\bar{\delta}_n \leq \delta$ and, for all $n \geq \tilde{n}(\delta)$, $\tilde{\delta}_n \leq \delta$. Since $\bar{U}_n(\delta) \leq \tilde{U}_n(\delta),\ \delta \in (0,1]$, it is also clear that

$$\bar{n}(\delta) \leq \tilde{n}(\delta),\ \delta \in (0,1].$$

The next proposition immediately follows from the definition of $\bar{n}(\delta)$ (it is, in fact, just a reformulation of the statements of Proposition 2 and Theorem 3):

Proposition 5 *(i) For all $n \geq \bar{n}(\delta)$,*

$$\mathbb{P}\Big\{\mathcal{E}_P(\hat{f}_n) > \delta\Big\} \leq \sum_{\delta_j \geq \delta} e^{-t_j};$$

(ii) For all $n \geq \bar{n}(\delta)$, with probability at least

$$1 - \sum_{\delta_j \geq \delta} e^{-t_j}$$

the following inclusions hold:

$$\forall \sigma \geq \delta \ \ \mathcal{F}_P(\sigma) \subset \mathcal{F}_{P_n}(3/2\sigma) \ \text{ and } \ \mathcal{F}_{P_n}(\sigma) \subset \mathcal{F}_P(2\sigma).$$

(iii) For all $n \geq \bar{n}(\delta)$, there exists an event E of probability

$$\mathbb{P}(E) \geq 1 - 3\sum_{\delta_j \geq \delta} e^{-t_j}$$

such that on this event

$$\bar{U}_n(\sigma) \leq \hat{U}_n(\sigma) \leq \tilde{U}_n(\sigma), \ \sigma \geq \delta.$$

We will also need a version of the statements of Proposition 5 that are uniform in $n \in M$. To this end, assume that the numbers t_j in the definitions of the functions $\bar{U}_n, \hat{U}_n, \tilde{U}_n$ depend also on n. We will denote these numbers $t_j^{(n)}$. Assume that, for all j, $\frac{t_j^{(n)}}{n}$ is a nonincreasing function of n. Then the next statement immediately follows from Theorem 3 and the union bound.

Proposition 6 *There exists an event H of probability*

$$\mathbb{P}(H) \geq 1 - \sum_{n \in M}\sum_{j \geq 0} e^{-t_j^{(n)}}$$

such that on this event, for all $n \in M$,

$$\mathcal{E}_P(\hat{f}_n) \leq \bar{\delta}_n,$$

$$\forall \delta \geq \bar{\delta}_n \ \ \mathcal{F}_P(\delta) \subset \mathcal{F}_{P_n}(3/2\delta) \ \text{ and } \ \mathcal{F}_{P_n}(\delta) \subset \mathcal{F}_P(2\delta).$$

Moreover, there exists an event E of probability

$$\mathbb{P}(E) \geq 1 - 3\sum_{n \in M}\sum_{j \geq 0} e^{-t_j^{(n)}}$$

such that on this event, for all $n \in M$,

$$\bar{U}_n(\delta) \leq \hat{U}_n(\delta) \leq \tilde{U}_n(\delta), \ \sigma \geq \bar{\delta}_n$$

and

$$\bar{\delta}_n \leq \hat{\delta}_n \leq \tilde{\delta}_n.$$

As a consequence, we also have that for all $\delta \in (0, 1]$

$$\bar{n}(\delta) \leq \hat{n}(\delta) \leq \tilde{n}(\delta).$$

The simplest choice of the numbers $t_j^{(n)}$, in the case when $M = \{2^k : k \geq 0\}$ and $\delta_j = 2^{-j},\ j \geq 0$, is

$$t_j^{(n)} = 2\log(\log_2 n + 1) + 2\log(j+1) + \log\frac{1}{\alpha} + \log(12),\ j \geq 0,\ n \in M,$$

where $\alpha \in (0,1)$. With this choice, all the claims of Proposition 6 hold with a guaranteed probability at least $1-\alpha$.

The main conclusion one can draw from Proposition 6 is that the sample size needed to achieve the desired "accuracy of learning" δ (i.e., to "learn" a function for which the excess risk is smaller than δ) can itself be learned from the data. More precisely, the estimator $\hat{n}(\delta)$ of the required sample size can be computed sequentially by increasing the sample size n gradually, computing for each n the data dependent excess risk bound $\hat{\delta}_n$ and stopping as soon as $\hat{\delta}_n \leq \delta$. With a high probability, the stopping time $\hat{n}(\delta)$ provides a "correct" estimate of the sample size (up to a numerical constant) in the sense that it is between two distribution dependent estimates ($\bar{n}(\delta)$ and $\tilde{n}(\delta)$) that are typically of the same order of magnitude (up to numerical constants). At the same time, the sample size $\hat{n}(\delta)$ is sufficient for estimating the σ-minimal sets of the true risk by the σ-minimal sets of the empirical risk for all $\sigma \geq \delta$ (in the sense of the inclusions of Proposition 6). These facts will play a crucial role in our design of active learning methods in the next section.

3. Sequential Active Learning

We first describe a simplified (non-adaptive) version of active learning in which it is assumed that the minimal sample size $\bar{n}(\delta)$ needed to achieve the desired "accuracy of learning" of the order δ is given. As before, suppose that $\{\delta_k\}_{k\geq 0}$ is a nonincreasing sequence of positive numbers with $\delta_0 = 1$. Denote $\bar{n}_k := \bar{n}(\delta_k),\ k \geq 1$.

Algorithm 1

```
$\hat{\mathcal{F}}_0 := \mathcal{F}$;
for $k = 1, 2, \dots,$
  $\hat{A}_k := \Big\{x : \sup_{f,g\in\hat{\mathcal{F}}_{k-1}} |f(x)-g(x)| > \delta_k\Big\}$;
  $\hat{P}_k := \frac{1}{\bar{n}_k}\sum_{j=1}^{\bar{n}_k} I_{\hat{A}_k}(X_j)\delta_{X_j}$;
  $\hat{\mathcal{F}}_k := \hat{\mathcal{F}}_{k-1}\bigcap\mathcal{F}_{\hat{P}_k}(3\delta_k)$;
end
```

The set $\hat{A}_k$ defined at each iteration of the algorithm is viewed as a set of "active examples" (or "active set"). The examples $X_j \in \hat{A}_k$ are needed to compute the "active empirical measure" $\hat{P}_k$. The underlying assumption is that there exists a "base algorithm" that computes the δ-minimal set

$$\mathcal{F}_Q(\delta) := \Big\{f : \mathcal{E}_Q(f) := Qf - \inf_{f\in\mathcal{F}} Qf \leq \delta\Big\}$$

for an arbitrary discrete measure Q with a finite number of atoms. This algorithm is used to compute the set $\mathcal{F}_{\hat{P}_k}(3\delta_k)$. In principle, it would be enough only to ensure that, given

$\delta > 0$ and measure Q, the "base algorithm" outputs a set $\bar{\mathcal{F}}_Q(\delta)$ such that

$$\mathcal{F}_Q(c_1\delta) \subset \bar{\mathcal{F}}_Q(\delta) \subset \mathcal{F}_Q(c_2\delta)$$

for some numerical constants $0 < c_1 < c_2$. However, to simplify the notations, we will assume that $c_1 = c_2 = 1$.

Of course, in reality, **Algorithm 1** stops after a finite number of iterations. A possible choice could be

$$L := \sum_{j\geq 0} I(\delta_j \geq \delta),$$

which can be viewed as the number of iterations needed to achieve the "desired accuracy" of learning $\delta \in (0,1)$. In other words, the algorithm stops when δ_j becomes smaller than δ. For instance, if $\delta_j = 2^{-j}, j \geq 0$, then the number of iterations L is of the order $\log_2(1/\delta)$. Let $\nu(\delta)$ denote the total number of active examples utilized by the algorithm in the first L iterations. Then

$$\nu(\delta) \leq \sum_{\delta_k \geq \delta} \sum_{j=1}^{\bar{n}_k} I_{\hat{A}_k}(X_j).$$

Denote

$$A(\delta) := \Big\{x : \sup_{f,g\in\mathcal{F}(8\delta)} |f(x) - g(x)| > \delta\Big\}$$

and

$$\pi(\delta) := P(A(\delta)).$$

The following statement will be easily proved by induction.

Theorem 7 *With probability at least*

$$1 - \sum_{n\in M}\sum_{j\geq 0} e^{-t_j^{(n)}},$$

the following inclusions hold for the classes $\hat{\mathcal{F}}_k$ output by **Algorithm 1***: for all $k \geq 0$*

$$\mathcal{F}_P(\delta_k) \subset \hat{\mathcal{F}}_k \subset \mathcal{F}_P(8\delta_k). \tag{3}$$

Also, for all $t \geq 1$ and all $\delta \in (0,1]$, with probability at least

$$1 - \sum_{n\in M}\sum_{j\geq 0} \exp\{-t_j^{(n)}\} - \sum_{\delta_j\geq\delta} \exp\{-\bar{n}(\delta_j)\pi(\delta_{j-1})t\log t\}$$

the following bound holds:

$$\nu(\delta) \leq et \sum_{\delta_j\geq\delta} \bar{n}(\delta_j)\pi(\delta_{j-1}).$$

Proof The inclusions (3) obviously hold for $k = 0$. Assuming that, for all $j < k$,

$$\mathcal{F}_P(\delta_j) \subset \hat{\mathcal{F}}_j \subset \mathcal{F}_P(8\delta_j),$$

we will prove that the same inclusions hold also for k. Let H be the event of probability at least

$$1-\sum_{n\in M}\sum_{j\geq 0} e^{-t_j^{(n)}}$$

defined in Proposition 6. By the induction assumption,

$$\mathcal{F}_P(\delta_k)\subset\mathcal{F}_P(\delta_{k-1})\subset\hat{\mathcal{F}}_{k-1}$$

and, by the definition of $\hat{A}_k$, we have for all $f,g\in\hat{\mathcal{F}}_{k-1}$,

$$|P_{\bar{n}_k}(f-g)-\hat{P}_k(f-g)|=\left|\bar{n}_k^{-1}\sum_{i=1}^{\bar{n}_k}(f-g)(X_i)-\bar{n}_k^{-1}\sum_{i:X_i\in\hat{A}_k}(f-g)(X_i)\right|$$

$$=\left|\bar{n}_k^{-1}\sum_{i:X_i\notin\hat{A}_k}(f-g)(X_i)\right|\leq\delta_k.$$

We can conclude that, for all $f\in\mathcal{F}_P(\delta_k)$,

$$\left|\mathcal{E}_{P_{\bar{n}_k}}(f)-\mathcal{E}_{\hat{P}_k}(f)\right|\leq\delta_k.$$

Also, by the inclusions of Proposition 6 and the definition of $\bar{n}_k=\bar{n}(\delta_k)$, we have on the event H that

$$\mathcal{F}_P(\delta_k)\subset\mathcal{F}_{P_{\bar{n}_k}}(2\delta_k).$$

Hence, for all $f\in\mathcal{F}_P(\delta_k)$,

$$\mathcal{E}_{P_{\bar{n}_k}}(f)\leq 2\delta_k \;\text{ and }\; \mathcal{E}_{\hat{P}_k}(f)\leq 3\delta_k.$$

This implies the inclusion

$$\mathcal{F}_P(\delta_k)\subset\hat{\mathcal{F}}_{k-1}\bigcap\mathcal{F}_{\hat{P}_k}(3\delta_k)=\hat{\mathcal{F}}_k.$$

On the other hand, since $\hat{\mathcal{F}}_k\subset\hat{\mathcal{F}}_{k-1}$, we have, for all $f\in\hat{\mathcal{F}}_k$,

$$\left|\mathcal{E}_{P_{\bar{n}_k}}(f)-\mathcal{E}_{\hat{P}_k}(f)\right|\leq\delta_k.$$

Since for all $f\in\hat{\mathcal{F}}_k$, $\mathcal{E}_{\hat{P}_k}(f)\leq 3\delta_k$, we also have $\mathcal{E}_{P_{\bar{n}_k}}(f)\leq 4\delta_k$. Thus, using again the inclusions of Proposition 6, we get

$$\hat{\mathcal{F}}_k\subset\mathcal{F}_{P_{\bar{n}_k}}(4\delta_k)\subset\mathcal{F}_P(8\delta_k),$$

proving the inclusions (3)

To prove the bound on $\nu(\delta)$, note that on the event H, where the inclusions (3) hold for all k such that $\delta_k\geq\delta$, we have

$$\hat{A}_k\subset A(8\delta_{k-1}).$$

Hence, on the event H,

$$\nu(\delta) \le \sum_{\delta_k \ge \delta} \nu_k,$$

where

$$\nu_k := \sum_{j=1}^{\bar{n}_k} I_{A(8\delta_{k-1})}(X_j).$$

Clearly, ν_k is a binomial random variable with parameters $\bar{n}_k$ and $\pi(\delta_{k-1})$. Therefore, we have

$$\mathbb{P}\{\nu_k \ge s\} \le \left(\frac{e\bar{n}_k \pi(\delta_{k-1})}{s}\right)^s$$

(see, e.g., Dudley (1999), p. 16). Taking $s := et\bar{n}_k\pi(\delta_{k-1})$ yields

$$\mathbb{P}\Bigl\{\nu_k \ge et\bar{n}_k\pi(\delta_{k-1})\Bigr\} \le \exp\{-\bar{n}_k\pi(\delta_{k-1})t\log t\}.$$

Applying the union bound, we get

$$\mathbb{P}\biggl\{\nu(\delta) \ge et\sum_{\delta_k\ge\delta}\bar{n}_k\pi(\delta_{k-1})\biggr\} \le \mathbb{P}(H^c) + \sum_{\delta_k\ge\delta}\exp\{-\bar{n}_k\pi(\delta_{k-1})t\log t\}.$$

Since

$$\mathbb{P}(H^c) \le \sum_{n\in M}\sum_{j\ge 0} e^{-t_j^{(n)}},$$

the result follows. ■

The simplest way to make the method data dependent is to replace in **Algorithm 1** the sample sizes $\bar{n}_k = \bar{n}(\delta_k)$ by their estimates $\hat{n}_k := \hat{n}(\delta_k),\ k \ge 1$ and to redefine $\hat{P}_k$ in the iterative procedure for $\hat{A}_k, \hat{P}_k$ and $\hat{\mathcal{F}}_k$ as follows:

$$\hat{P}_k := \frac{1}{\hat{n}_k}\sum_{j=1}^{\hat{n}_k} I_{\hat{A}_k}(X_j)\delta_{X_j}.$$

This modification of **Algorithm 1** will be called **Algorithm 2**. The following statement can be proved quite similarly to Theorem 7 (using Proposition 6).

Recall the definition of the number of iterations L and also that $\nu(\delta)$ denotes the number of active examples utilized by the algorithm in the first L iterations.

Theorem 8 *With probability at least*

$$1 - 3\sum_{j\ge 0}\sum_{n\in M} e^{-t_j^{(n)}},$$

the following inclusions hold for the classes $\hat{\mathcal{F}}_k$ output by **Algorithm 2***: for all $k \ge 0$,*

$$\mathcal{F}_P(\delta_k) \subset \hat{\mathcal{F}}_k \subset \mathcal{F}_P(8\delta_k).$$

Moreover, for all $t \geq 1$ and for all $\delta \in (0,1]$, with probability at least

$$1 - 3\sum_{j\geq 0}\sum_{n\in M}\exp\{-t_j^{(n)}\} - \sum_{\delta_j\geq\delta}\exp\{-\tilde{n}(\delta_j)\pi(\delta_{j-1})t\log t\}$$

the following bound holds:

$$\nu(\delta) \leq et\sum_{\delta_j\geq\delta}\tilde{n}(\delta_j)\pi(\delta_{j-1}).$$

Note that in this version of the algorithm all the training examples X_j (not only the examples in the active sets $\hat{A}_k$) are used to determine the sample sizes $\hat{n}_k$. So, from this point of view, **Algorithm 2** can not be viewed as really "active". However, it is easy to see that in a more concrete framework of prediction problems (such as, for instance, the binary classification) one can modify the definitions of the localized Rademacher complexities and of the sample sizes $\hat{n}_k$ in such a way that they depend only on the design points, but not on the response variables (labels). Thus, in the cases when sampling the design points is "cheap" and only assigning the labels to them is "expensive" (which is a common motivational assumption in the literature on active learning), the algorithms of this type make some sense (see Section 4 for more details).

It is more interesting, however, that even in the abstract framework of empirical risk minimization it is possible to change the definition of Rademacher complexities and the estimates of the sample sizes based on them so that only the active examples that belong to the sets $\hat{A}_k$ are being used in the computation. We will describe such a data driven algorithm of active learning below.

Let $\delta_j := 2^{-j},\ j \geq 0$. As before, we will define iteratively data dependent function classes $\hat{\mathcal{F}}_k$ beginning with $\hat{\mathcal{F}}_0 := \mathcal{F}$ that provide estimates of the δ-minimal sets $\mathcal{F}_P(\delta)$ for sufficiently small values of δ and we set

$$\hat{A}_k := \Big\{x : \sup_{f,g\in\hat{\mathcal{F}}_{k-1}} |f(x) - g(x)| > c\delta_k\Big\} \text{ with some constant } c > 0.$$

Define

$$\hat{P}_n^{(k)} := n^{-1}\sum_{j=1}^n I_{\hat{A}_k}(X_j)\delta_{X_j}$$

and

$$\hat{R}_n^{(k)}(f) := n^{-1}\sum_{j=1}^n \varepsilon_j f(X_j) I_{\hat{A}_k}(X_j).$$

Denote

$$\hat{U}_n^{(k)} := \hat{K}\Bigg[\sup_{f,g\in\hat{\mathcal{F}}_{k-1}}\Big|\hat{R}_n^{(k)}(f-g)\Big| + D_{\hat{P}_n^{(k)}}(\hat{\mathcal{F}}_{k-1})\sqrt{\frac{t_k^{(n)}}{n}} + \frac{t_k^{(n)}}{n}\Bigg]$$

and define iteratively a nondecreasing data dependent sequence $\hat{n}_k$:

$$\hat{n}_k := \min\Big\{n \in M, n \geq \hat{n}_{k-1} : \hat{U}_n^{(k)} \leq \frac{1}{2}\delta_{k+1}\Big\}$$

with the initial condition $\hat{n}_0 := \inf M$.

Note that the following iterative relationships hold for the distribution dependent sample sizes $\bar{n}_k := \bar{n}(\delta_{k+1})$ and $\tilde{n}_k := \tilde{n}(\delta_{k+1})$:

$$\bar{n}_k = \min\Bigl\{n \in M, n \geq \bar{n}_{k-1} : \bar{U}_n(\delta_k) \leq \frac{1}{2}\delta_{k+1}\Bigr\}, \quad \bar{n}_0 := \inf M$$

and

$$\tilde{n}_k = \min\Bigl\{n \in M, n \geq \tilde{n}_{k-1} : \tilde{U}_n(\delta_k) \leq \frac{1}{2}\delta_{k+1}\Bigr\}, \quad \tilde{n}_0 := \inf M.$$

(which easily follows from the definitions of $\bar{n}(\delta), \tilde{n}(\delta)$).

We will write, for brevity, $\hat{P}_k := \hat{P}^{(k)}_{\hat{n}_k}$. With these definitions and notations, we can define $\hat{\mathcal{F}}_k$ iteratively exactly as before:

$$\hat{\mathcal{F}}_k := \hat{\mathcal{F}}_{k-1} \bigcap \mathcal{F}_{\hat{P}_k}(3\delta_k).$$

In short, the algorithm can be described as follows:

```
**Algorithm 3**
$\hat{\mathcal{F}}_0 := \mathcal{F}$;
**for** $k = 1, 2, \dots,$
  $\hat{A}_k := \Bigl\{x : \sup_{f,g\in\hat{\mathcal{F}}_{k-1}} |f(x) - g(x)| > c\delta_k\Bigr\}$;
  $\hat{n}_k := \min\Bigl\{n \in M, n \geq \hat{n}_{k-1} : \hat{U}^{(k)}_n \leq \frac{1}{2}\delta_{k+1}\Bigr\}$;
  $\hat{\mathcal{F}}_k := \hat{\mathcal{F}}_{k-1} \bigcap \mathcal{F}_{\hat{P}_k}(3\delta_k)$;
**end**
```

As before, we define

$$A(\delta) := \Bigl\{x : \sup_{f,g\in\mathcal{F}(8\delta)} |f(x) - g(x)| > c\delta\Bigr\}$$

and

$$\pi(\delta) := P(A(\delta)).$$

The properties of **Algorithm 3** are summarized in the following theorem.

Theorem 9 *There exist numerical constants c in the definition of the active sets $\hat{A}_k$, $\hat{K}$ in the definition of $\hat{U}^{(k)}_n$ and $\tilde{K}, \tilde{c}$ in the definition of the function $\tilde{U}_n$ such that the following holds. With probability at least*

$$1 - 3\sum_{j\geq 0}\sum_{n\in M} e^{-t^{(n)}_j},$$

the following inequalities and inclusions hold for all $k \geq 0$:

$$\bar{n}_k \leq \hat{n}_k \leq \tilde{n}_k,$$

$$\mathcal{F}_P(\delta_k) \subset \hat{\mathcal{F}}_k \subset \mathcal{F}_P(8\delta_k).$$

Moreover, for all $t \geq 1$, with probability at least

$$1 - 3\sum_{j\geq 0}\sum_{n\in M}\exp\{-t_j^{(n)}\} - \sum_{\delta_j\geq\delta}\exp\{-\tilde{n}(\delta_{j+1})\pi(\delta_{j-1})t\log t\}$$

the following bound holds:

$$\nu(\delta) \leq et\sum_{\delta_j\geq\delta}\tilde{n}(\delta_{j+1})\pi(\delta_{j-1}).$$

Proof There exists an event E of probability at least

$$1 - 3\sum_{n\in M}\sum_{j\geq 0} e^{-t_j^{(n)}}$$

on which the following holds. For all k and for all $n \in M$

$$\hat{K}\left[\sup_{f,g\in\mathcal{F}_P(8\delta_{k-1})}\Big|R_n(f-g)\Big| + D_{P_n}(\mathcal{F}_P(8\delta_{k-1}))\sqrt{\frac{t_k^{(n)}}{n}} + \frac{t_k^{(n)}}{n}\right] \leq \frac{1}{2}\tilde{U}_n(\delta_k) \tag{4}$$

and

$$\hat{K}\left[\sup_{f,g\in\mathcal{F}_P(\delta_{k-1})}\Big|R_n(f-g)\Big| + D_{P_n}(\mathcal{F}_P(\delta_{k-1}))\sqrt{\frac{t_k^{(n)}}{n}} + \frac{t_k^{(n)}}{n}\right] \geq 2\bar{U}_n(\delta_k) \tag{5}$$

with properly chosen constants in the definitions of the functions $\bar{U}_n, \tilde{U}_n$ and constant $\hat{K}$. At the same time, on the same event E, for all $n \in M, n \geq \bar{n}(\delta)$ and all $\sigma \geq \delta$,

$$\mathcal{F}_P(\sigma) \subset \mathcal{F}_{P_n}(2\sigma) \text{ and } \mathcal{F}_{P_n}(\sigma) \subset \mathcal{F}_P(2\sigma). \tag{6}$$

To construct such an event, first consider the event H of Proposition 6 on which the inclusions (6) hold. Then define

$$E_{k,n} := \left\{\hat{K}\left[\sup_{f,g\in\mathcal{F}_P(8\delta_{k-1})}\Big|R_n(f-g)\Big| + D_{P_n}(\mathcal{F}_P(8\delta_{k-1}))\sqrt{\frac{t_k^{(n)}}{n}} + \frac{t_k^{(n)}}{n}\right] \leq \frac{1}{2}\tilde{U}_n(\delta_k)\right\}$$

and

$$E'_{k,n} := \left\{\hat{K}\left[\sup_{f,g\in\mathcal{F}_P(\delta_{k-1})}\Big|R_n(f-g)\Big| + D_{P_n}(\mathcal{F}_P(\delta_{k-1}))\sqrt{\frac{t_k^{(n)}}{n}} + \frac{t_k^{(n)}}{n}\right] \geq 2\bar{U}_n(\delta_k)\right\}.$$

Using the "statistical version" of Talagrand's concentration inequality (Theorem 4) one can show that with a proper choice of $\hat{K}$ and the constants in the definitions of the functions $\bar{U}_n, \tilde{U}_n$,

$$\mathbb{P}(E_{n,k}) \leq 1 - e^{-t_k^{(n)}}, \quad \mathbb{P}(E'_{n,k}) \geq 1 - e^{-t_k^{(n)}}$$

for all $k \geq 0$ and for all $n \in M$. Define

$$E := \bigcap_{k\geq 0, n\in M}(E_{n,k} \cap E'_{n,k}) \cap H.$$

Then

$$\mathbb{P}(E) \geq 1 - 3\sum_{n\in M}\sum_{j\geq 0} e^{-t_j^{(n)}}$$

and all the desired properties hold on the event E.

We will now show by induction that, on the event E for $k = 0, 1, \dots$

$$\bar{n}_k \leq \hat{n}_k \leq \tilde{n}_k,$$

$$\mathcal{F}_P(\delta_k) \subset \hat{\mathcal{F}}_k \subset \mathcal{F}_P(8\delta_k)$$

and also for $k = 1, 2, \dots$ and for all $n \in M$

$$2\bar{U}_n(\delta_k) - \frac{\delta_{k+1}}{2} \leq \hat{U}_n^{(k)} \leq \frac{1}{2}\Big[\tilde{U}_n(\delta_k) + \frac{\delta_{k+1}}{2}\Big]. \tag{7}$$

By the definitions, the claims are obviously true for $k = 0$. Assume that they have been proved up to $k-1$. By this induction assumption, we have $\hat{\mathcal{F}}_{k-1} \subset \mathcal{F}_P(8\delta_{k-1})$ and, by the definition of the set $\hat{A}_k$,

$$\sup_{f,g\in\hat{\mathcal{F}}_{k-1}} \Big|\hat{R}_n^{(k)}(f-g)\Big| \leq \sup_{f,g\in\hat{\mathcal{F}}_{k-1}} \Big|R_n(f-g)\Big| + c\delta_k$$

and

$$D^2_{\hat{P}_n^{(k)}}(\hat{\mathcal{F}}_{k-1}) \leq D^2_{P_n}(\hat{\mathcal{F}}_{k-1}) + c^2\delta_k^2.$$

This implies the following upper bound on $\hat{U}_n^{(k)}$:

$$\hat{U}_n^{(k)} \leq \hat{K}\Big[\sup_{f,g\in\mathcal{F}_P(8\delta_{k-1})}\Big|R_n(f-g)\Big| + D_{P_n}(\mathcal{F}_P(8\delta_{k-1}))\sqrt{\frac{t_k^{(n)}}{n}} + \frac{t_k^{(n)}}{n} + c\delta_k + c\delta_k\sqrt{\frac{t_k^{(n)}}{n}}\Big].$$

Applying to the last term the inequality $ab \leq (a^2+b^2)/2$ and taking into account the fact that $\delta_k \leq 1$, it is easy to deduce from this that with c satisfying the condition

$$\hat{K}c + \hat{K}^2c^2/2 \leq 1/8,$$

we have

$$\hat{U}_n^{(k)} \leq \hat{K}\Big[\sup_{f,g\in\mathcal{F}_P(8\delta_{k-1})}\Big|R_n(f-g)\Big| + D_{P_n}(\mathcal{F}_P(8\delta_{k-1}))\sqrt{\frac{t_k^{(n)}}{n}} + \frac{t_k^{(n)}}{n}\Big] + \delta_{k+1}/4.$$

Quite similarly, using the inclusion $\mathcal{F}_P(\delta_{k-1}) \subset \hat{\mathcal{F}}_{k-1}$ that also holds under the induction assumption, one can show that with a proper choice of constant c in the definition of the set $\hat{A}_k$

$$\hat{U}_n^{(k)} \geq \hat{K}\Big[\sup_{f,g\in\mathcal{F}_P(\delta_{k-1})}\Big|R_n(f-g)\Big| + D_{P_n}(\mathcal{F}_P(\delta_{k-1}))\sqrt{\frac{t_k^{(n)}}{n}} + \frac{t_k^{(n)}}{n}\Big] - \delta_{k+1}/2.$$

Combining this with bounds (4) and (5) immediately implies (7).

Applying (7) to $n = \hat{n}_k$, we get

$$2\bar{U}_{\hat{n}_k}(\delta_k) - \frac{\delta_{k+1}}{2} \leq \hat{U}^{(k)}_{\hat{n}_k} \leq \frac{\delta_{k+1}}{2},$$

which yields

$$\bar{U}_{\hat{n}_k}(\delta_k) \leq \frac{\delta_{k+1}}{2}.$$

We also have $\hat{n}_k \geq \hat{n}_{k-1} \geq \bar{n}_{k-1}$ (by the induction assumption). By the definition of $\bar{n}_k$, this implies that $\hat{n}_k \geq \bar{n}_k$.

On the other hand, denote $\hat{n}_k^-$ the element of the ordered set M preceding $\hat{n}_k$. We will use inequality (7) with $n = \hat{n}_k^-$. It gives

$$\hat{U}^{(k)}_{\hat{n}_k^-} \leq \frac{1}{2}\Big[\tilde{U}_{\hat{n}_k^-}(\delta_k) + \frac{\delta_{k+1}}{2}\Big]. \tag{8}$$

If it happened that $\hat{n}_k^- < \hat{n}_{k-1}$, then we must have $\hat{n}_k = \hat{n}_{k-1}$, which, by the induction assumption, implies that $\hat{n}_k = \hat{n}_{k-1} \leq \tilde{n}_{k-1} \leq \tilde{n}_k$. If $\hat{n}_k^- \geq \hat{n}_{k-1}$, then the definition of $\hat{n}_k$ implies that

$$\hat{U}^{(k)}_{\hat{n}_k^-} > \frac{\delta_{k+1}}{2},$$

which together with (8) implies that

$$\tilde{U}_{\hat{n}_k^-}(\delta_k) > \frac{\delta_{k+1}}{2}.$$

But, if $\hat{n}_k > \tilde{n}_k$, then $\hat{n}_k^- \geq \tilde{n}_k$, which would imply that

$$\tilde{U}_{\tilde{n}_k}(\delta_k) > \frac{\delta_{k+1}}{2}$$

(since for all δ, $\tilde{U}_n(\delta)$ is a nonincreasing function of n). The last inequality contradicts the definition of $\tilde{n}_k$ implying that $\hat{n}_k \leq \tilde{n}_k$.

The proof of the inclusions

$$\mathcal{F}_P(\delta_k) \subset \hat{\mathcal{F}}_k \subset \mathcal{F}_P(8\delta_k)$$

and the derivation of the bound on $\nu(\delta)$ repeat the argument of Theorem 7. ■

Remark. Note that in the bounds on $\nu(\delta)$ of theorems 7, 8 and 9 one can replace functions $\pi(\delta)$, $\bar{n}(\delta)$ and $\tilde{n}(\delta)$ by arbitrary upper bounds (with the same change in the bounds on the probability).

As soon as $\pi(\delta) \to 0$ as $\delta \to 0$, the upper bounds on $\nu(\delta)$ show that, in the case of active learning, there is a reduction of the number of training examples needed to achieve a desired accuracy of learning comparing with passive learning. In the next section, we explore in some detail what is happening in the case of binary classification.

4. Active Learning in Binary Classification

Let (X, Y) be a random couple with values in $S \times \{-1, 1\}$ and with distribution P, where $(S, \mathcal{A})$ is an arbitrary measurable space. In binary classification problems, the first component X is viewed as an observable instance and the second component Y is an unobservable "label". The value of Y is to be predicted based on an observation of X and on the training data $(X_1, Y_1), \dots, (X_n, Y_n)$ consisting of n independent copies of (X, Y). Measurable functions $g : S \mapsto \{-1, 1\}$ are called (binary) classifiers. Let $\ell : \{-1, 1\} \times \{-1, 1\} \mapsto \{0, 1\}$ be the binary loss function $\ell(y, u) := I(y \neq u)$, and, as before, $(\ell \bullet g)(x, y) := \ell(y, g(x))$ be the "loss" of classifier g for the example $(x, y) \in S \times \{-1, 1\}$. The quantity

$$P(\ell \bullet g) = P\{(x, y) : y \neq g(x)\} = \mathbb{P}\{Y \neq g(X)\}$$

is called the generalization error, or the risk of g. We still denote $\eta(x) := \mathbb{E}(Y|X = x)$ the regression function. It is well known that the minimum of the generalization error over the set of all binary classifiers is attained at the Bayes classifier

$$g_*(x) = \operatorname{sign}(\eta(x)).$$

We will assume in what follows that $\mathcal{G}$ is a class of binary classifiers such that $g_* \in \mathcal{G}$.

For a binary classifier g, define its excess risk as

$$\mathcal{E}_P(\ell \bullet g) := P(\ell \bullet g) - P(\ell \bullet g_*).$$

The following formula is well known (see, e.g., Devroye, Györfi and Lugosi (1996), Theorem 2.2).

$$\mathcal{E}_P(\ell \bullet g) = \int_{\{g \neq g_*\}} |\eta(x)| \Pi(dx), \tag{9}$$

where Π is the marginal distribution of X.

A standard approach to learning the Bayes classifier is based on the empirical risk minimization:

$$\hat{g} := \operatorname{argmin}_{g \in \mathcal{G}} P_n(\ell \bullet g) = \operatorname{argmin}_{g \in \mathcal{G}} P_n\{(x, y) : y \neq g(x)\} =$$

$$\operatorname{argmin}_{g \in \mathcal{G}} n^{-1} \sum_{j=1}^{n} I(Y_j \neq g(X_j)),$$

where P_n denotes the empirical distribution based on the training data $(X_1, Y_1), \dots, (X_n, Y_n)$ (we will also use the notation Π_n for the empirical distribution based on $(X_1, \dots, X_n)$).

If $\mathcal{F} := \ell \bullet \mathcal{G} := \{\ell \bullet g : g \in \mathcal{G}\}$ denotes the loss class, then we are in the framework of abstract empirical risk minimization of sections 2,3 and general results of these sections can be now specialized for the classification context.

It is natural to characterize the quality of the classifier $\hat{g}$ in terms of its excess risk $\mathcal{E}_P(\ell \bullet \hat{g})$ and to study how it depends on the complexity of the class $\mathcal{G}$ as well as on the complexity of the classification problem itself. The simplest complexity assumption on the class $\mathcal{G}$ is that it is a VC-class of binary functions of VC-dimension V (in other words, $\mathcal{C} := \Big\{\{x : g(x) = +1 : g \in \mathcal{G}\}\Big\}$ is a VC-class of sets of VC-dimension V). Under this assumption, a well known result, essentially due to Vapnik and Chervonenkis, is that, for some constant $K > 0$ and for all $t > 0$, with probability at least $1 - e^{-t}$

$$\mathcal{E}_P(\ell \bullet \hat{g}) \leq K\left[\sqrt{\frac{V}{n}} + \sqrt{\frac{t}{n}}\right].$$

In principle, this bound is minimax optimal, but it can be significantly improved for special families of distributions P under further assumptions on the complexity of the classification problem. For instance, the following **Massart's low noise assumption** is frequently used: for some constant $h \in (0, 1]$

$$|\eta(x)| \geq h, \ x \in S.$$

The parameter h is a characteristic of the level of noise in binary labels Y_j. In other words, it is a simple measure of complexity of a binary classification problem. The following theorem is a version of the result proved by Massart and Nedelec (2006):

Theorem 10 *There exists a constant $K > 0$ such that, for all $t > 0$, with probability at least $1 - e^{-t}$*

$$\mathcal{E}_P(\ell \bullet \hat{g}) \leq K\left[\frac{V}{nh}\log\Bigl(\frac{nh^2}{V}\Bigr) + \frac{t}{nh}\right] \bigwedge \left[\sqrt{\frac{V}{n}} + \sqrt{\frac{t}{n}}\right].$$

This upper bound on the excess risk is optimal in a minimax sense (as it was also shown by Massart and Nedelec (2006)). However, it still can be refined using the following function τ (which is a local version of Alexander's **capacity function** introduced in the 80s and used in the theory of ratio type empirical processes, see Giné and Koltchinskii (2006) and references therein). Define

$$\mathcal{G}_\delta := \{g \in \mathcal{G} : \Pi\{x : g(x) \neq g_*(x)\} \leq \delta\}$$

and let

$$\tau(\delta) := \frac{\Pi\Bigl\{x | \exists g \in \mathcal{G}_\delta : g(x) \neq g_*(x)\Bigr\}}{\delta}.$$

Clearly, the set $\mathcal{G}_\delta$ consists of the classifiers from $\mathcal{G}$ that are in a neighborhood of size δ of the Bayes classifier g_* and the set

$$D_\delta := \Bigl\{x | \exists g \in \mathcal{G}_\delta : g(x) \neq g_*(x)\Bigr\}$$

consists of all the points x such that there exists a classifier g in the neighborhood $\mathcal{G}_\delta$ that "disagrees" with the Bayes classifier at x. The function $\tau(\delta)$ is always upper bounded by $\frac{1}{\delta}$. However, if it happens that the measure Π of the "disagreement set" D_δ is small when δ is small, then $\tau(\delta)$ might grow slower than $\frac{1}{\delta}$ as $\delta \to 0$, or even it can be bounded by a constant. If $\mathcal{C} := \{\{g = +1\} : g \in \mathcal{G}\}$ and $C_* := \{g_* = +1\}$, then

$$\tau(\delta) = \frac{\Pi\Bigl(\bigcup_{C \in \mathcal{C}, \Pi(C \triangle C_*) \leq \delta} (C \triangle C_*)\Bigr)}{\delta},$$

so, roughly, $\tau(\delta)$ shows how many disjoint sets $C\triangle C_*$ of "size" δ can be "packed" in the union of all such sets. For instance, if $\mathcal{C}$ is a class of convex sets in $[0,1]^d$, Π is the Lebesgue measure in $[0,1]^d$ and $C_* \in \mathcal{C}, \Pi(C_*) > 0$, then it can be shown that τ is uniformly bounded by a constant (see Giné and Koltchinskii (2006)). A more detailed analysis of disagreement sets, capacity functions and their connections to the geometry of the class $\mathcal{G}$ can be found in Friedman (2009).

The following result was proved in Giné and Koltchinskii (2006).

Theorem 11 *There exists a constant* $K > 0$ *such that, for all* $t > 0$, *with probability at least* $1 - e^{-t}$

$$\mathcal{E}_P(\ell \bullet \hat{g}) \le K\left[\frac{V}{nh}\log\tau\Big(\frac{V}{nh^2}\Big) + \frac{t}{nh}\right]\bigwedge\left[\sqrt{\frac{V}{n}} + \sqrt{\frac{t}{n}}\right].$$

Clearly, this result implies the theorem of Massart and Nedelec (since $\tau(\delta) \le \frac{1}{\delta}$). The proof is based on applying subtle bounds for empirical processes discussed in Giné and Koltchinskii (2006) to compute the excess risk bound $\bar{\delta}_n$ of Section 2. Then, the general result of Theorem 1 (see also Proposition 2) can be used to bound the excess risk.

The case when $\tau(\delta)$ is uniformly bounded from above by a constant τ_0 is of special interest. In this case, with probability at least $1 - e^{-t}$,

$$\mathcal{E}_P(\ell \bullet \hat{g}) \le K\left[\frac{V}{nh}\log\tau_0 + \frac{t}{nh}\right],$$

so the main term of the bound is of the order $O(\frac{V}{nh})$ and it does not contain logarithmic factors depending on n and h. It will be convenient for our purposes to phrase this result in a slightly different form. Namely, given $\delta \in (0,1)$ and $\alpha \in (0,1)$, denote

$$n(\delta,\alpha) := \inf\Big\{n : \mathbb{P}\{\mathcal{E}_P(\hat{g}_n) \ge \delta\} \le \alpha\Big\}.$$

Then

$$n(\delta,\alpha) \le K\bigg(\left[\frac{V}{\delta h}\log\tau_0 + \frac{\log(1/\alpha)}{\delta h}\right]\bigwedge\left[\frac{V}{\delta^2} + \frac{\log(1/\alpha)}{\delta^2}\right]\bigg).$$

The quantity $n(\delta,\alpha)$ shows how many training examples are needed to make the excess risk of the classifier $\hat{g}$ smaller than δ with a guaranteed probability of at least $1-\alpha$. It characterizes the sample complexity of passive learning. In the case of empirical risk minimization over a VC-class with a bounded capacity function τ, the sample complexity is of the order $O(\frac{V}{h}\frac{1}{\delta})$.

The role of the capacity function is rather modest in the case of passive learning since it only allows one to refine the excess risk and the sample complexity bounds by making the logarithmic factors more precise. However, the capacity function τ happened to be of crucial importance in the analysis of active learning methods of binary classification. This function was rediscovered in active learning literature and its supremum is being used there under the name of **disagreement coefficient** (see, e.g., Hanneke (2009a, 2009b) and references therein).

We will describe an active learning algorithm that is a specialized version of more abstract **Algorithm 3** of Section 3. As before, we denote $\delta_j := 2^{-j},\ j \geq 0$ and choose a set $M \subset \mathbb{N}$ of natural numbers as well as nonnegative real numbers $t_k^{(n)},\ n \in M, k \geq 0$.

Given a class $\mathcal{G}$ of binary classifiers, denote

$$\mathcal{G}_P(\delta) := \biggl\{ g : \mathcal{E}_P(\ell \bullet g) \leq \delta \biggr\},\ \delta > 0.$$

These sets will be called δ-minimal sets of the true risk. Clearly, if $\mathcal{F} = \ell \bullet \mathcal{G}$, then under the notations of Section 2

$$\mathcal{F}_P(\delta) = \biggl\{ \ell \bullet g : g \in \mathcal{G}_P(\delta) \biggr\}.$$

In principle, one can directly use **Algorithm 3** for the class $\mathcal{F}$. However, we will modify it slightly in order to adapt it to the binary classification framework.

We will define iteratively data dependent function classes $\hat{\mathcal{G}}_k$ that provide estimates of the δ-minimal sets $\mathcal{G}_P(\delta)$. and also a nondecreasing data dependent sequence of estimated sample sizes $\hat{n}_k$. It will be assumed that we have an access to an algorithm that, given a discrete measure Q on $S \times \{-1, 1\}$ and $\delta > 0$, computes the δ-minimal set $\mathcal{G}_Q(\delta)$ of Q.

Several definitions and notations will be needed. Note that for the binary loss ℓ, for all binary classifiers g_1, g_2 and for all $\delta \in (0, 1)$, the condition $|\ell(y, g_1(x)) - \ell(y, g_2(x))| \geq \delta$ is equivalent to the condition $g_1(x) \neq g_2(x)$. This leads to the following definition of sets $\hat{A}_k$ (that are subsets of S, not of $S \times \{-1, 1\}$). Assuming that $\hat{\mathcal{G}}_{k-1}$ has been already defined, let

$$\hat{A}_k := \biggl\{ x : \exists g_1, g_2 \in \hat{\mathcal{G}}_{k-1}, g_1(x) \neq g_2(x) \biggr\}$$

be the set of all the points where at least two classifiers in $\hat{\mathcal{G}}_{k-1}$ disagree with each other. This set will be used as a set of active design points at the k-th iteration.

Next define active empirical distributions based on the unlabeled examples $\{X_j\}$ and on the labeled examples $\{(X_j, Y_j)\}$:

$$\hat{\Pi}_n^{(k)} := n^{-1} \sum_{j=1}^{n} I_{\hat{A}_k}(X_j) \delta_{X_j}$$

and

$$\hat{P}_n^{(k)} := n^{-1} \sum_{j=1}^{n} I_{\hat{A}_k}(X_j) \delta_{(X_j, Y_j)}$$

For simplicity, we will also use the notation $\hat{P}_k := \hat{P}_{\hat{n}_k}^{(k)}$. Let

$$\hat{D}_n^{(k)} := \frac{1}{2} \sup_{g_1, g_2 \in \hat{\mathcal{G}}_{k-1}} \Bigl(\hat{\Pi}_n^{(k)} (g_1 - g_2)^2 \Bigr)^{1/2}$$

be the $L_2(\hat{\Pi}_n^{(k)})$-diameter of the class $\hat{\mathcal{G}}_{k-1}$. Note that, if we literally followed the definitions of Section 3, we would have to define the diameter as

$$\sup_{g_1, g_2 \in \hat{\mathcal{G}}_{k-1}} \Bigl(\hat{P}_n^{(k)} (\ell \bullet g_1 - \ell \bullet g_2)^2 \Bigr)^{1/2}.$$

However, it is easy to check that for all $(x, y) \in S \times \{-1, 1\}$ and all binary classifiers g_1, g_2

$$(\ell \bullet g_1)(x, y) - (\ell \bullet g_2)(x, y) = \frac{1}{2} y(g_2(x) - g_1(x)),$$

which justifies the new definition. This simple identity also implies that the function $\phi_n(\delta)$, defined in Section 2 and used in the construction of the excess risk bounds, can be upper bounded as follows:

$$\phi_n(\delta) \leq 2\mathbb{E} \sup_{f_1, f_2 \in \mathcal{F}(\delta)} |R_n(f_1 - f_2)| = 2\mathbb{E} \sup_{g_1, g_2 \in \mathcal{G}(\delta)} |R_n(\ell \bullet g_1 - \ell \bullet g_2)| =$$

$$\mathbb{E} \sup_{g_1, g_2 \in \mathcal{G}(\delta)} \biggl| n^{-1} \sum_{j=1}^{n} \varepsilon_j Y_j (g_2(X_j) - g_1(X_j)) \biggr|,$$

where at the beginning we used the symmetrization inequality (see. e.g. van der Vaart and Wellner (1996)). Note that, conditionally on $(X_1, Y_1), \dots, (X_n, Y_n)$, the distribution of the random vector $(\varepsilon_1 Y_1, \dots, \varepsilon_n Y_n)$ is the same as the distribution of $(\varepsilon_1, \dots, \varepsilon_n)$. Because of this,

$$\phi_n(\delta) \leq \mathbb{E} \sup_{g_1, g_2 \in \mathcal{G}(\delta)} \biggl| n^{-1} \sum_{j=1}^{n} \varepsilon_j Y_j (g_2(X_j) - g_1(X_j)) \biggr| =$$

$$\mathbb{E}\mathbb{E}\biggl(\sup_{g_1, g_2 \in \mathcal{G}(\delta)} \biggl| n^{-1} \sum_{j=1}^{n} \varepsilon_j Y_j (g_2(X_j) - g_1(X_j)) \biggr| \biggm| (X_1, Y_1), \dots, (X_n, Y_n) \biggr) =$$

$$\mathbb{E}\mathbb{E}\biggl(\sup_{g_1, g_2 \in \mathcal{G}(\delta)} \biggl| n^{-1} \sum_{j=1}^{n} \varepsilon_j (g_2(X_j) - g_1(X_j)) \biggr| \biggm| (X_1, Y_1), \dots, (X_n, Y_n) \biggr) =$$

$$\mathbb{E} \sup_{g_1, g_2 \in \mathcal{G}(\delta)} |R_n(g_1 - g_2)|.$$

This simple observation allows one to replace the Rademacher complexities for the loss class $\mathcal{F} = \ell \bullet \mathcal{G}$ by the Rademacher complexities for the class $\mathcal{G}$ itself (and the proofs of the excess risk bounds and other results cited in Section 2 go through with no changes). Of course, the same applies to all the constructions and the results of Section 3.

Because of this, we now define the Rademacher complexity based only on the "active" examples as

$$\hat{R}_n^{(k)} := \sup_{g_1, g_2 \in \hat{\mathcal{G}}_{k-1}} \biggl| n^{-1} \sum_{j=1}^{n} \varepsilon_j (g_1 - g_2)(X_j) I_{\hat{A}_k}(X_j) \biggr|.$$

Finally, denote

$$\hat{U}_n^{(k)} := \hat{K} \biggl[\hat{R}_n^{(k)} + \hat{D}_n^{(k)} \sqrt{\frac{t_k^{(n)}}{n}} + \frac{t_k^{(n)}}{n} \biggr].$$

With these definitions and notations, we can now introduce the following modification of **Algorithm 3** of Section 3.

Algorithm 4
$\hat{\mathcal{G}}_0 := \mathcal{G}$;

```
for $k = 1, 2, \dots,$
  $\hat{A}_k := \Big\{x : \exists g_1, g_2 \in \hat{\mathcal{G}}_{k-1}, g_1(x) \neq g_2(x)\Big\};$
  $\hat{n}_k := \min\Big\{n \in M, n \geq \hat{n}_{k-1} : \hat{U}_n^{(k)} \leq \frac{1}{2}\delta_{k+1}\Big\};$
  $\hat{\mathcal{G}}_k := \hat{\mathcal{G}}_{k-1} \bigcap \mathcal{G}_{\hat{P}_k}(3\delta_k);$
end
```

Remark. One can also use in **Algorithm 4** the Rademacher complexities defined as follows:

$$\hat{R}_n^{(k)} := \sup_{g_1, g_2 \in \hat{\mathcal{G}}_{k-1}} \left| n^{-1} \sum_{j=1}^{n} \varepsilon_j (g_1 - g_2)(X_j) \right|.$$

In this case, not only the active design points, but all the design points X_j are used to compute the Rademacher complexities and to estimate the sample sizes $\hat{n}_k$. Note, however, that the labels Y_j are not involved in this computation, so, the algorithm still can be viewed as "active". The resulting algorithm is a modification of **Algorithm 3** from Section 3.

In the case when $\mathcal{G}$ is a VC-class of VC-dimension V, we will choose $M := \{2^k : k \geq 0\}$. We will also define

$$t_k^{(n)} := \log(1/\alpha) + 2\log(k+1) + 2\log(\log_2 n + 1) + \log(24). \tag{10}$$

This leads to the following result that is a corollary of Theorem 9.

Corollary 12 *Let $\delta \in (0,1)$. Suppose that Massart's low noise assumption holds with some $h \in (0,1)$. Suppose that*

$$\tau_0 := \sup_{u \in (0,1]} \tau(u) < +\infty.$$

Then there exists an event of probability at least $1 - \alpha$ such that the following inclusions hold for the classes $\hat{\mathcal{G}}_k$ output by **Algorithm 4***: for all $k \geq 0$,*

$$\mathcal{G}_P(\delta_k) \subset \hat{\mathcal{G}}_k \subset \mathcal{G}_P(8\delta_k). \tag{11}$$

Also with probability at least $1-\alpha$, the following bound on the number $\nu(\delta)$ of active training examples used by **Algorithm 4** *in the first $L = \Big[\log_2(1/\delta)\Big]$ iterations holds with some numerical constant $C > 0$:*

$$\nu(\delta) \leq C\frac{\tau_0 \log(1/\delta)}{h^2}\Big[V\log\tau_0 + \log(1/\alpha) + \log\log(1/\delta) + \log\log(1/h)\Big].$$

In particular, it means that with probability at least $1 - \alpha$

$$\mathcal{G}_P(\delta/2) \subset \hat{\mathcal{G}}_L \subset \mathcal{G}_P(16\delta).$$

Proof We only sketch the proof here, the missing details are not very complicated. The result follows from Theorem 9, more precisely, from its modified version that takes into account the slight changes we made in the definition of the Rademacher complexities. The inclusions (11) follow from this theorem in a straightforward way. To prove the bound on $\nu(\delta)$, one has first to bound the quantity $\tilde{\delta}_n$. This computation was essentially done by Giné and Koltchinskii (2006) (it actually leads to the bound of Theorem 11). Namely, with some constant C_1,

$$\bar{\delta}_n \leq C_1 \biggl[\frac{V}{nh} \log \tau\Bigl(\frac{V}{nh^2}\Bigr) + \frac{\log(1/\alpha) + \log\log n}{nh}\biggr] \bigwedge \biggl[\sqrt{\frac{V}{n}} + \sqrt{\frac{\log(1/\alpha) + \log\log n}{n}}\biggr].$$

As a result, the following upper bound on $\tilde{n}(\sigma), \sigma \geq \delta$ holds with some constant K_1 :

$$\tilde{n}(\sigma) \leq K_1 \biggl(\biggl[\frac{V}{\sigma h} \log \tau_0 + \frac{\log(1/\alpha) + \log\log_2(1/\delta) + \log\log(1/h)}{\sigma h}\biggr] \bigwedge$$

$$\biggl[\frac{V}{\sigma^2} + \frac{\log(1/\alpha) + \log\log_2(1/\delta)}{\sigma^2}\biggr]\biggr).$$

Note that, under Massart's low noise assumption, formula (9) for the excess risk implies that for all binary classifiers g

$$\mathcal{E}(\ell \bullet g) \geq h\Pi\{x : g(x) \neq g_*(x)\}.$$

Hence

$$\mathcal{F}(\sigma) \subset \Bigl\{\ell \bullet g : g \in \mathcal{G}_{\sigma/h}\Bigr\}.$$

For the sets $A(\sigma)$ used in Theorems 9, this implies the following:

$$A(\sigma) = \Bigl\{(x,y) : \sup_{f_1, f_2 \in \mathcal{F}(8\sigma)} |f_1(x,y) - f_2(x,y)| > c\sigma\Bigr\} \subset$$

$$\Bigl\{(x,y) : \sup_{g_1, g_2 \in \mathcal{G}(8\sigma/h)} |(\ell \bullet g_1)(x,y) - (\ell \bullet g_2)(x,y)| > c\sigma\Bigr\} =$$

$$\Bigl\{x : \exists g_1, g_2 \in \mathcal{G}(8\sigma/h) : g_1(x) \neq g_2(x)\Bigr\} \times \{-1,1\} = \Bigl\{x : \exists g \in \mathcal{G}(8\sigma/h) : g(x) \neq g_*(x)\Bigr\} \times \{-1,1\}$$

(we used the assumption that $g_* \in \mathcal{G}$ and, hence, $g_* \in \mathcal{G}(8\sigma/h)$). This implies, by the definitions of the functions π and τ, that

$$\pi(\sigma) = P(A(\sigma)) \leq \Pi\Bigl(\Bigl\{x : \exists g \in \mathcal{G}(8\sigma/h) : g(x) \neq g_*(x)\Bigr\}\Bigr) \leq \frac{8\sigma}{h}\tau(8\sigma/h).$$

Using the definition of τ_0, we conclude that for all $\sigma \geq \delta$ $\pi(\sigma) \leq \frac{8\tau_0}{h}\sigma$. It remains to substitute the bounds on $\tilde{n}(\sigma)$ and $\pi(\sigma)$ into the bound on $\nu(\delta)$ of Theorem 9

$$\nu(\delta) \leq et \sum_{\delta_j \geq \delta} \tilde{n}(\delta_{j+1})\pi(\delta_{j-1}),$$

say, with $t = e$. This gives

$$\nu(\delta) \leq e^2 \sum_{\delta_j \geq \delta} K_1 \biggl(\biggl[\frac{V}{\delta_{j+1} h} \log \tau_0 + \frac{\log(1/\alpha) + \log\log_2(1/\delta) + \log\log(1/h)}{\delta_{j+1} h}\biggr] \frac{8\tau_0}{h}\delta_{j-1},$$

which is bounded from above by

$$C\frac{\tau_0\log(1/\delta)}{h^2}\Bigl[V\log\tau_0+\log(1/\alpha)+\log\log(1/\delta)+\log\log(1/h)\Bigr].$$

with a properly chosen numerical constant C. Also, it easily follows from the probability estimates of Theorem 9 that the above bound on $\nu(\delta)$ holds with probability at least $1-\alpha$. ■

Finally, we discuss the properties of **Algorithm 4** under **Tsybakov's low noise assumption**. Namely, we assume that for some $\gamma>0$, for some constant B and for all $t>0$

$$\Pi\{x:|\eta(x)|\le t\}\le Bt^{\gamma}.$$

It is well known that under this assumption the following bound on the excess risk holds for an arbitrary classifier g :

$$\mathcal{E}_P(\ell\bullet g)\ge c\Pi^{\kappa}(\{g\ne g_*\}),$$

where $\kappa=\frac{1+\gamma}{\gamma}$ and c is a constant that depends on B,κ. We will assume in this case that $\mathcal{G}$ is not necessarily a VC-class, but it can be more massive. For instance, denote $N(\mathcal{G};L_2(\Pi_n);\varepsilon)$ the minimal number of $L_2(\Pi_n)$-balls of radius ε needed to cover $\mathcal{G}$ and suppose that these covering numbers satisfy the condition:

$$\log N(\mathcal{G};L_2(\Pi_n);\varepsilon)\le\left(\frac{A}{\varepsilon}\right)^{2\rho},\ \varepsilon>0.$$

for some $\rho\in(0,1]$ and some constant $A>0$. Then, the following upper bound on the excess risk of an empirical risk minimizer $\hat{g}$ holds with probability at least $1-e^{-t}$:

$$\mathcal{E}_P(\ell\bullet\hat{g})\le K\biggl(\left(\frac{1}{n}\right)^{-\kappa/(2\kappa+\rho-1)}+\left(\frac{t}{n}\right)^{\kappa/(2\kappa-1)}\biggr),$$

where K is a constant depending on κ,ρ,A,B. The bounds of this type were first proved by Tsybakov (2004) (see also Koltchinskii (2006, 2008)). It easily follows from this bound that in order to achieve the excess risk of order δ one needs $O\Bigl(\delta^{-2+(1-\rho)/\kappa}\Bigr)$ training examples.

We will now consider **Algorithm 4** with $M:=\{2^k:k\ge0\}$, and with the real numbers $t_k^{(n)}$ defined by (10).

This leads to the following result that is also a corollary of Theorem 9.

Corollary 13 *Let $\delta\in(0,1)$. Suppose that Tsybakov's low noise assumption holds with some $\gamma>0$ and $B>0$. Let $\kappa:=\frac{1+\gamma}{\gamma}$. Suppose that*

$$\tau_0:=\sup_{u\in(0,1]}\tau(u)<+\infty.$$

Then there exists an event of probability at least $1-\alpha$ such that the following inclusions hold for the classes $\hat{\mathcal{G}}_k$ output by **Algorithm 4***: for all k with $\delta_k\ge\delta$,*

$$\mathcal{G}_P(\delta_k)\subset\hat{\mathcal{G}}_k\subset\mathcal{G}_P(8\delta_k).\tag{12}$$

Also with probability at least $1-\alpha$, the following bound on the number $\nu(\delta)$ of active training examples used by **Algorithm 4** *holds with some constant $C>0$ depending on κ, ρ, A, B :*

$$\nu(\delta) \le C\tau_0\Bigl[\delta^{-2+(2-\rho)/\kappa} + \delta^{-2+2/\kappa}(\log(1/\alpha) + \log\log(1/\delta))\Bigr].$$

The proof is similar to that of Corollary 12. In this case, the improvement comparing with passive learning is by a factor $\delta^{1/\kappa}$.

Remark. Alternatively, one can assume that the active learning algorithm stops as soon as the specified number of active examples, say, n has been achieved. If $\hat{L}$ denotes the number of iterations needed to achieve this target, then $8\delta_{\hat{L}}$ is an upper bound on the excess risk of the classifiers from the set $\hat{\mathcal{G}}_{\hat{L}}$. Under the assumptions of Corollary 12, inverting the bound on $\nu(\delta)$ easily gives that $\delta_{\hat{L}}$ is upper bounded by

$$\exp\biggl\{-\beta\frac{nh^2}{C_2\tau_0}\biggr\},$$

where

$$\beta := \frac{1}{V\log\tau_0 \vee \log(1/\alpha) \vee \log(nh^2/C_2\tau_0) \vee \log\log(1/h)}$$

with some numerical constant C_2. Thus, the excess risk of such classifiers tends to zero exponentially fast as $n \to \infty$. This is the form in which the excess risk bounds in active learning are usually stated in the literature (see, e.g., Hanneke (2009a, 2009b)). In fact, this is a refinement of the bounds of Hanneke that were proved for somewhat different active learning algorithms (see Hanneke (2009b), theorems 4, 5). Similarly, under the conditions of Corollary 13, the bound on $\delta_{\hat{L}}$ becomes

$$\biggl(\frac{\tau_0}{n}\biggr)^{\kappa/(2\kappa+\rho-2)} \bigvee \biggl(\frac{\tau_0(\log(1/\alpha) + \log\log n)}{n}\biggr)^{\kappa/(2\kappa-2)}.$$

(compare with Theorem 6 in Hanneke (2009b)).

Remark. Although we concentrated in this section only on binary classification problems, the active learning algorithms described in Section 3 can be also used in the context of multiclass classification and some other problems (e.g., estimation of non-smooth regression function and estimation of level sets of a probability density). Recall that in the framework of prediction with a general loss function ℓ described in the Introduction,

$$\mathcal{F} := \ell \bullet \mathcal{G} := \Bigl\{\ell \bullet g : g \in \mathcal{G}\Bigr\}.$$

Following an idea of Beygelzimer, Dasgupta and Langford (2009), one can now replace the disagreement set $\hat{A}_k$ for the class $\hat{\mathcal{F}}_{k-1} = \ell \bullet \hat{\mathcal{G}}_{k-1}$ involved in **Algorithm 3** by a larger set

$$\hat{A}_k^+ := \biggl\{(x,y) : \exists g_1, g_2 \in \hat{\mathcal{G}}_{k-1} \sup_{y'\in T} |\ell(y', g_1(x)) - \ell(y', g_2(x))| > c\delta_k\biggr\} =$$

$$\biggl\{x : \exists g_1, g_2 \in \hat{\mathcal{G}}_{k-1} \sup_{y\in T} |\ell(y, g_1(x)) - \ell(y, g_2(x))| > c\delta_k\biggr\} \times T.$$

This leads to the following modification of **Algorithm 3**:

```
Algorithm 5
$\hat{\mathcal{G}}_0 := \mathcal{G}$;
for $k = 1, 2, \dots,$
    $\hat{A}_k^+ := \Big\{x : \exists g_1, g_2 \in \hat{\mathcal{G}}_{k-1}\ \sup_{y\in T} |\ell(y, g_1(x)) - \ell(y, g_2(x))| > c\delta_k\Big\} \times T.$
    $\hat{n}_k := \min\Big\{n \in M, n \geq \hat{n}_{k-1} : \hat{U}_n^{(k)} \leq \frac{1}{2}\delta_{k+1}\Big\}$;
    $\hat{\mathcal{G}}_k := \hat{\mathcal{G}}_{k-1} \bigcap \mathcal{G}_{\hat{P}_k}(3\delta_k)$;
end
```

The Rademacher complexities of classes $\hat{\mathcal{F}}_k = \ell \bullet \hat{\mathcal{G}}_k$ (the quantities $\hat{U}_n^{(k)}$) as well as active empirical measures $\hat{P}_k$ involved in this algorithm are now based on active sets $\hat{A}_k^+$. Clearly, only the labels of active examples are used in this version of the algorithm. If now we define

$$\pi(\delta) := \Pi\Big(\Big\{x : \exists g_1, g_2 \in \mathcal{G}_P(8\delta)\ \sup_{y\in T} |\ell(y, g_1(x)) - \ell(y, g_2(x))| > c\delta\Big\}\Big),$$

it is very easy to check that the statement of Theorem 9 still holds for such a modification of the algorithm. At the same time, it is not clear at this point whether a modified definition of disagreement coefficient in the paper by Beygelzimer, Dasgupta and Langford (2009) can be used to analyze the properties of active learning algorithms of this type and whether it is possible to extend such an analysis beyond classification and similar problems.

Acknowledgments

The author is thankful to anonymous referees for a number of helpful comments and corrections.

References

[1] Balcan, M.–F., Hanneke, S. and Wortman, J. (2008) The true sample complexity of active learning. In: *Proc. 21st Annual Conference on Learning Theory (COLT 2008)*

[2] Balcan, M.–F., Beygelzimer, A. and Langford, J. (2009) Agnostic Active Learning. *J. of Computer and System Sciences,* 75, 1, 78–89.

[3] Bartlett, P., Boucheron, S. and Lugosi, G. (2002) Model Selection and Error Estimation, *Machine Learning,* 48, 85–113.

[4] Bartlett, P., Bousquet, O. and Mendelson, S. (2005) Local Rademacher Complexities, *Annals of Statistics,* 33,4, 1497–1537.

[5] Beygelzimer, A., Dasgupta, S. and Langford, J. (2009) Importance Weighted Active Learning, In *Proceedings of the 26th International Conference on Machine Learning* ICML 2009, Montreal, Canada.

[6] Castro, R., Willett, R. and Nowak, R. (2005) Faster rates in regression via active learning. In: *Advances in Neural Information Processing Systems (NIPS 2005).*

[7] Castro, R.M. and Nowak, R.D. (2008) Minimax bounds for active learning. *IEEE Transactions on Information Theory,* 54, 5, 2339–2353.

[8] Dasgupta, S., Hsu, D. and Monteleoni, C. (2007) A general agnostic active learning algorithm. In: *Advances in Neural Information Processing Systems (NIPS 2007).*

[9] Devroye, L., Györfi, L. and Lugosi, G. (1996) A Probabilistic Theory of Pattern Recognition, Springer.

[10] Dudley, R.M. (1999) Uniform Central Limit Theorems, Cambridge University Press.

[11] Giné, E. and Koltchinskii, V. (2006) Concentration Inequalities and Asymptotic Results for Ratio Type Empirical Processes. *Annals of Probability,* 34, 3, 1143–1216.

[12] Friedman, E.J. (2009) Active Learning for Smooth Problems, In *Proceeding of the 22nd Annual Conference on Learning Theory*, COLT 2009, Montreal, Canada.

[13] Hanneke, S. (2009a) Adaptive Rates of Convergence in Active Learning. In: *Proc. 22nd Annual Conference on Learning Theory (COLT 2009).*

[14] Hanneke, S. (2009b) Rates of Convergence in Active Learning. Preprint.

[15] Koltchinskii, V. (2001) Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory,* 47(5), 1902–1914.

[16] Koltchinskii, V. (2006) Local Rademacher Complexities and Oracle Inequalities in Risk Minimization, *Annals of Statistics,* 34, 6, 2593–2656.

[17] Koltchinskii, V. (2008) Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Lecture Notes. *Ecole d'ete de Probabilités de Saint-Flour 2008.*

[18] Koltchinskii, V. and Panchenko, D. (2000) Rademacher processes and bounding the risk of function learning. In: Giné, E., Mason, D. and Wellner, J. (Eds) *High Dimensional Probability II,*443-459, Birkhäuser, Boston.

[19] Massart, P. (2007) Concentration Inequalities and Model Selection. *Ecole d'ete de Probabilités de Saint-Flour 2003,* Lecture Notes in Mathematics, Springer.

[20] Massart, P. and Nedelec, E. (2006) Risk bounds for statistical learning, *Annals of Statistics,* 34, 5, 2326–2366.

[21] Tsybakov, A. (2004) Optimal aggregation in statistical learning, *Annals of Statistics,* 32, 135–166.

[22] van der Vaart, A. and Wellner, J. (1996) Weak Convergence and Empirical Processes, Springer.