Statistical Approaches in Learning Theory: boosting and ranking

Nicolas Vayatis

HABILITATION À DIRIGER DES RECHERCHES
DE
L'UNIVERSITÉ PIERRE-ET-MARIE-CURIE (PARIS 6)

HABILITATION À DIRIGER DES RECHERCHES

Spécialité : **Mathématiques**

présentée par

Nicolas VAYATIS

Approches statistiques en apprentissage : boosting et ranking

---

Rapporteurs :	M.	Peter	**BARTLETT**	UC Berkeley
	M.	Vladimir	**KOLTCHINSKII**	GeorgiaTech
	M.	Pascal	**MASSART**	Université Paris-Sud

Soutenue le **9 Décembre 2006** devant le jury composé de

M.	Lucien	**BIRGÉ**	Université Paris 6	Prsident
M.	Gábor	**LUGOSI**	Universitat Pompeu Fabra	Examinateur
M.	Vladimir	**KOLTCHINSKII**	Georgia Institute of Technology	Rapporteur
Mme	Dominique	**PICARD**	Université Paris 7	Examinateur
M.	Alexandre	**TSYBAKOV**	Université Paris 6	Examinateur

Remerciements

La recherche mathématique fait voyager notre esprit mais elle n'est pas pour autant une activité désincarnée. Le présent mémoire n'aurait pas pu exister sans la rencontre avec des collègues en tout point merveilleux.

Son histoire commence à Barcelone, il y a quelques années, et je me souviens encore très précisément du soir où nous avons formulé les principes du boosting régularisé avec Gábor Lugosi. Nous étions à grignoter quelques tapas sur Icaria avant d'aller au cinéma et c'est sur ce coin de table que notre aventure autour du boosting a démarré... Grâce à Gábor, j'ai appris que la recherche n'est pas une activité exclusivement à caractère monacal. Je le remercie infiniment de m'avoir fait partager son savoir, son énergie et son enthousiasme pendant ces années à Barcelone, et jusquà ce jour. Köszönöm Gábor !

L'histoire se poursuit à Paris. J'y suis accueilli par Sacha Tsybakov. Notre première discussion a lieu autour d'un repas auvergnat près de Chevaleret. Je m'en souviens comme si c'était hier. Dans mon imaginaire de grec errant, je réalisais que j'avais trouvé une nouvelle famille d'accueil. Je ne remercierai jamais assez Sacha pour son aide, ses encouragements et sa bienveillance à mon égard.

Dans cette famille d'accueil, j'ai aussi eu le plaisir de rencontrer Lucien Birgé. A travers nos discussions sur les choix pédagogiques, Lucien m'a révélé la noblesse de la statistique mathématique. C'est une image qui est à présent gravée dans mon esprit. Je souhaite exprimer à Lucien ma reconnaissance pour sa gentillesse et pour m'avoir toujours donné les meilleurs conseils.

Je remercie également Dominique Picard d'avoir accepté spontanément de participer au jury. La finalisation de ce mémoire doit beaucoup à l'énergie positive que Dominique sait transmettre aux autres.

L'histoire va se poursuivre à Atlanta. Vladimir Koltchinskii m'a proposé de l'y rejoindre pour quelques mois. Sa participation dans le jury est un très grand honneur pour moi. Je lui suis extrêmement reconnaissant d'avoir aussi accepté la tâche ingrate de rapporteur.

Il me tient à coeur d'exprimer mes remerciements à deux personnalités scientifiques que j'estime particulièrement: Peter Bartlett et Pascal Massart, qui ont également été

rapporteurs de ce travail. Même si leur influence fut plus distante, je leur dois beaucoup et je voudrais leur exprimer ma reconnaissance et ma sympathie.

Depuis quelques années, je collabore de façon soutenue avec Stéphan Clémençon. Nous avons la chance de partager, en plus de l'amitié, les mêmes intérêts scientifiques et une même vision des mathématiques appliquées. Je souhaite profiter de cette rubrique pour remercier Stéphan de sa confiance et de sa présence tout au long de ces années.

L'organisation de la recherche actuelle, au moins dans le domaine de la modélisation aléatoire, en a fait une activité fondamentalement collective. Or, le facteur humain est souvent perçu comme une source de dérèglements. Dans mes collaborations, je l'ai vécu comme une énergie motrice. Ce fut une chance pour moi de travailler avec Stéphan, Gábor, et Sacha, mais aussi avec Gilles Blanchard, Anatoli Iouditski, Arturo Kohatsu-Higa, et Sasha Nazin. Même la distance n'a que faiblement entamé l'intensité de nos échanges.

Je souhaite également remercier, les collègues à Barcelone ou à Paris, que j'ai côtoyés pendant ces années et avec qui j'ai beaucoup appris. Je pense surtout à : Jean-Yves Audibert, Patrice Bertail, Gérard Biau, Stéphane Boucheron, Olivier Bousquet, Cristina Butucea, Arnak Dalalyan, Theo Evgeniou, Stéphane Herbin, Gérard Kerkyacharian, Arturo Kohatsu-Higa, Guillaume Lecué, Florence Merlevède, Gilles Stoltz, Frederic 'Kic' Udina, Michel van de Velde, Jean-Philippe Vert. Je remercie au même titre, mais plus particulièrement: Philippe Rigollet qui m'a permis d'utiliser ses résultats de simulation sur l'algorithme de descente miroir, et Zhan Shi, avec qui j'ai l'immense plaisir de partager mon bureau au LPMA depuis mon arrivée.

J'en profite aussi pour exprimer ma gratitude envers les personnes qui constituent l'âme du LPMA: Maryvonne de Béru, Salima Chili, Nelly Lecquyer, Philippe Macé, Jacques Portès, et Josette Saman.

Je n'oublie pas mes fidèles compagnons de toujours, Matthieu, Shlomo, Yves, Laurent, Christian, Gilles, Yann, Martin, Panayis et Charles. Même si la vie adulte réduit la fréquence de nos rencontres, ils sont toujours là.

Enfin, je souhaite dédier ce travail à toute ma famille qui, malgré la distance, m'apporte toute son affection et son soutien, au petit Hercule qui doit conquérir sa force, et à Magdalena qui rend la vie plus belle.

Publications

[1] R. Azencott and N. Vayatis. Distribution-Dependent Vapnik-Chervonenkis Bounds. Proceedings of EuroColt 1999 (University of Dortmund), in LNCS Computational Learning Theory vol. 1572, pp. 230–240, Springer, 1999.

[2] R. Azencott and N. Vayatis. Refined Exponential Rates in Vapnik-Chervonenkis Inequalities. *Comptes Rendus de l'Académie des Sciences de Paris*, t.332, série I, 563–568, 2001.

[3] G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting methods. *Journal of Machine Learning Research*, 4:861–894, 2003.

[4] S. Clémençon, G. Lugosi, and N. Vayatis. From Ranking to Classification: a Statistical View. Proceedings of the 29th Annual Conference of the German Classification Society (GfKl 2005), University of Magdeburg, 2005.

[5] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and scoring using empirical risk minimization. Proceedings of COLT 2005, in LNCS Computational Learning Theory vol. 3559, pp. 1–15, Springer, 2005.

[6] S. Clémençon, G. Lugosi, and N. Vayatis. Discussion on the 2004 IMS Medallion Lecture "Local Rademacher complexities and oracle inequalities in risk minimization" by V. Koltchinskii. *Annals of Statistics*, 34(6), 2006. To appear.

[7] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. Submitted to the Annals of Statistics, 2006.

[8] S. Clémençon and N. Vayatis. On Ranking the Best Instances. Working paper, 2006.

[9] A. Juditsky, A. Nazin, A. Tsybakov, and N. Vayatis. Generalization Error Bounds for Aggregation by Mirror Descent with Averaging. Proceedings of NIPS 2005, in Advances in Neural Information Processing Systems 18 (eds. Y. Weiss and B. Schölkopf and J. Platt), 603–610, MIT Press, 2006.

[10] A. Juditsky, A. Nazin, A. Tsybakov, and N. Vayatis. Recursive Aggregation of Estimators via the Mirror Descent Algorithm with Averaging. *Problems of Information Transmission*, 41(4): 368–384, 2005.

[11] G. Lugosi and N. Vayatis. A Consistent Strategy for Boosting Algorithms. Proceedings of COLT 2002 (University of Sidney), in LNCS Computational Learning Theory vol. 2375, pp. 303–318, Springer, 2002.

[12] G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods (with discussion). *Annals of Statistics*, 32(1):30–55, 2004.

[13] G. Lugosi and N. Vayatis. Rejoinder "Three Papers on Boosting". *Annals of Statistics*, 32(1):124–127, 2004.

[14] N. Vayatis. *Inégalités de Vapnik-Chervonenkis et mesures de complexité*. Thèse de l'Ecole Polytechnique, 2000.

[15] N. Vayatis. The Role of Critical Sets in Vapnik-Chervonenkis Theory. Proceedings of COLT 2000, Stanford University, 2000.

[16] N. Vayatis. Exact Rates in Vapnik-Chervonenkis Bounds. *Annales de l'Institut Henri Poincaré (B) - Probabilités et Statistiques*, 39(1): 95–119, 2003.

Introduction

The field known as Statistical Learning Theory was born from the meeting of two communities: computer scientists involved in Machine Learning and mathematicians interested mainly in Non-Parametric Statistics.

The pioneering work of Vapnik and Chervonenkis [115, 116], and the books by Vapnik [112, 113, 114] later on, highly contributed to this event by establishing the connection between the generalization property of learning algorithms and empirical processes techniques. The so-called Structural Risk Minimization principle [112, 113] was also questioning nonparametric statistics from the viewpoint of model selection ([31], [24]). Indeed, a flow of new questions, connections, formulations was generated from the new approach. At the same time, an impressive amount of technical results such as concentration inequalities and empirical processes tools became available after the works of Talagrand [106], Van der Vaart and Wellner [110], Massart [87], and van de Geer [109].

Ten years (or so) after the milestones of Statistical Learning Theory were published (see the books by Vapnik [113, 114] and Devroye, Györfi, and Lugosi [52]) and the revolutionary algorithms known as Boosting and Support Vector Machines were introduced, an astonishing amount of theoretical breakthroughs has been reported (see [34] for a survey). Connections and differences between nonparametric statistics and learning theory are now clarified to a large extent and a new generation of researchers combining sensitivity to real-world applications and high-level education in theoretical statistics is growing. More recently, new developments are also attracting mathematicians involved in Information Theory, Approximation Theory and Game Theory. The evident bridge between Statistics and Optimization Theory is also being revisited.

In the present document, I will summarize a part of this evolution by describing some specific aspects of research in Statistical Learning Theory. My main concern will be to emphasize the statistical modelling part and the results of statistical nature in order to grasp the behavior of a learning algorithm (boosting) or to understand the main features of a learning problem (ranking).

The reference problem along the following pages and the starting point of this research is the simple binary classification problem. As a matter of fact, classical statistics have been considering this problem many years ago by the means of Discriminant Analysis. However, this setup presented a major drawback: it was unable to take into account complex and/or high-dimensional data. I briefly recall the *statistical model* and the *learning paradigm* for classification.

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. The *statistical model* for classification considers a random pair (X, Y) with unknown distribution P, where $X \in \mathcal{X}$ is an observation vector (being thought of high dimensionality) and $Y \in \{-1, +1\}$ is a binary label. The distribution P can be described by the pair (μ, η) where μ is the marginal distribution of X and η is the conditional distribution of Y given X that is to say $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$, $\forall x \in \mathcal{X}$. The setup is nonparametric and thus, no further assumptions on P are made. A *classifier* is a measurable function $g : \mathcal{X} \to \{-1, +1\}$ which makes a prediction $g(X)$ for each observation vector X. A "good" classifier should mostly predict the correct label Y of the observation X and thus, present a low *classification error* $L(g) = \mathbb{P}\{g(X) \neq Y\}$. The best of all classifiers is known to be the *Bayes rule* defined by $g^*(x) = 2\mathbb{I}_{\{\eta(x)>1/2\}} - 1$, $\forall x \in \mathcal{X}$, the Bayes error being denoted by $L^* = L(g^*)$. However, this ideal classifier g^* cannot be found in practice because it depends on the distribution P which is known only through a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of i.i.d. copies of (X, Y). The problem is then to build, on the basis of these data, a classifier $\hat{g}_n$ providing predictions as close as possible to those of g^*. We point out that, for any classifier g, we have the following expression of the *excess risk* $L(g) - L^* = \mathbb{E}(|2\eta(X) - 1|\, \mathbb{I}_{[g(X)\neq g^*(X)]})$. This expression reveals that the behavior of the function η around $1/2$ determines the difficulty of the classification problem.

The *learning paradigm* for classification which was formalized by Aizerman, Braverman and Rozonoer [19, 20] and further developed by Vapnik [112, 113] is based on the idea of Empirical Risk Minimization (ERM) which goes back to Le Cam [77]. Indeed, the simplest strategy to minimize $L(g)$ over a class $\mathcal{G}$ of candidate classifiers, given an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, is to minimize its empirical version, i.e. the empirical classification error $L_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{g(X_i)\neq Y_i\}}$. This strategy actually defines the ERM principle and delivers a classifier $\hat{g}_n = \arg\min_{g\in\mathcal{G}} L_n(g)$ which will hopefully mimic the target classifier g^*. The main contribution of Vapnik and Chervonenkis ([115, 116, 113, 114] was to provide a combinatorial characterization of the size of $\mathcal{G}$ that guarantees the success of the ERM principle. He also proposed a model selection principle known as the Structural Risk Minimization which extended the ERM and is based on the complexity concept of Vapnik-Chervonenkis (VC) dimension. Then, intensive efforts were made to improve upper bounds on the performance of classifiers based on ERM or SRM principles.

Despite these advances, Vapnik's programme had to face a major criticism: the practical implementation of the ERM principle is usually not feasible even for simple classes of classifiers. While classification theory was developing, new and highly performing classification algorithms known as Boosting ([98], [56], [58]) and Support Vector Machines (SVM, [33], [113], [45]) were proposed. At this point, the question was whether it was possible to discover underlying optimization principles accounting for these algorithms and to extend the classification theory developed so far in order to encompass them. This remark provided the first motivation of the work presented in the following pages, with a particular focus on BOOSTING.

Another limitation in the standard classification setup is the focus on a particular performance/error measure which is the classification error. Such a criterion is not necessarily relevant in applications. For instance, in information retrieval applications, for a given query, the instances (documents) can be labelled as "relevant" or "not relevant" for this

query. If we want to "learn" the relation between the descriptors of any document and the corresponding label in order to make predictions, we are facing a binary classification problem as before. But there is an important difference when visualizing the results: the query appeals to a list of relevant documents where the order in this list matters. Indeed, one expects to find the most relevant documents at the top of the list. Rather than learning the labels, the problem here is to learn the preferences between documents. This observation can be made formal and a specific error measure can be proposed for this RANKING problem.

In these two situations, the goal is to study principles (or algorithms) which perform optimization of special criteria. Once identified, these new optimization principles should be submitted to the same questions as in the standard setup of the ERM principle: find the optimal elements, explore universally consistent strategies, state nonasymptotic excess risk bounds, explore the conditions for having fast rates of convergence, prove oracle inequalities, and so on.

The present manuscript reports some contributions along these lines in the two contexts previously mentioned: (1) Convex Risk Minimization for classification using BOOSTING methods (Chapters 1 and 2), (2) Empirical and Convex Risk Minimization for the RANKING problem (Chapter 3).

Chapter 1

Boosting Methods - Theory and Algorithms

In the nineties, boosting algorithms became popular because of their simple heuristics, their tremendous efficiency on high dimensional data, but also because of the mystery surrounding their dynamics.

On the one hand, Vapnik-Chervonenkis theory was there to explain why learning algorithms could generalize properly on new data, on the other hand, boosting came out of a different approach and did not fit with the existing theoretical framework.

The one thing to avoid when using learning algorithms is overfitting but the question remained open for some time as far as boosting was concerned. Indeed, in most simulation studies, boosting was exhibiting high performance on test data but in some cases overfitting was observed [63]. The issue of consistency of boosting became of major interest in the learning community, the "most important question in Machine Learning" according to Leo Breiman.

Important steps in understanding why boosting worked so well were taken by Breiman [37] and Friedman, Hastie and Tibshirani [60]. Breiman showed the first consistency result but in the idealistic setup of an infinite data sample. Friedman, Hastie, and Tibshirani remarked that some boosting algorithms could be interpreted as stagewise fitting of additive logistic regression. At the same time, Mason, Bartlett, Baxter and Frean [84, 86, 85] considered optimization procedures similar to boosting with various cost functions and their work also has been influential in developing the theory for Convex Risk Minimization . We also refer to the work of Koltchinskii and Panchenko [76] in which they established margin bounds for combinations of classifiers. They improved on [99] and their use of empirical processes techniques in this problem triggered many of the further developments.

This chapter will focus on the basic framework allowing to understand boosting methods as statistical procedures. We provide simple theory for efficient algorithms and show how the subsequent reformulations can lead to new algorithms presenting interesting features.

First, we recall the fundamentals of boosting algorithms starting from the AdaBoost algorithm. Then, we describe the setup of convex risk minimization and provide the first

consistency results. The ideas introduced at this point turn out to be fruitful and made possible to develop a mirror descent algorithm for online classification.

1.1 Plain boosting algorithms

Boosting algorithms were introduced as iterative procedures of a deterministic nature which, given a set of data, would at each step t: (i) select a classifier g_t from a given class $\mathcal{G}$ of base classifiers, (ii) evaluate a real-valued weight w_t for this classifier, (iii) output the weighted majority vote $f_t = \sum_{i=1}^{t} w_i g_i$ of the selected classifiers up to this step.

Along with this simple aggregation principle, boosting is also characterized by the use of a probability distribution on the sample points. The idea is to start with the uniform distribution and to update it at each step according to some rule reinforcing the probability associated to misclassified points during the iterations.

We recall at this point the prototype boosting algorithms called AdaBoost introduced by Freund and Schapire ([56], [98], [58]) but first we need to introduce some notations. We denote by π_t the vector representing the discrete distribution on the sample points at step t and by

$$\epsilon_t(g) = \sum_{i=1}^{n} \pi_t(i) \mathbb{I}_{[g(X_i) \neq Y_i]}$$

the weighted empirical error of a base classifier g.

Algorithm 1 *(*ADABOOST*)*

- *Initialization: take* $\pi_1 = (1/n, \ldots, 1/n)$.

For $t = 1, \ldots, T$, *execute the following procedure:*

- *choose a base classifier* g_t *approximately minimizing* ϵ_t *over all* $g \in \mathcal{G}$
- *set* $\epsilon_t^* = \epsilon_t(g_t)$ *and adjust the weight of the classifier* g_t

$$w_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t^*}{\epsilon_t^*} \right)$$

- *update the distribution on the data* (X_i, Y_i)

$$\pi_{t+1}(i) \propto \pi_t(i) \exp(-w_t y_i g_t(x_i))$$

The final output of AdaBoost is $f_T = \sum_{t=1}^{T} w_t g_t$.

Numerous variants of AdaBoost have been proposed. We refer to the paper by Mason, Bartlett, Baxter and Frean [86] and the survey by Meir and Rätsch [91] for a sample of such algorithms. Particularly, one could wonder about the choices of the update rules for the distribution π_t and for the evaluation of the weights w_t, and whether these choices are optimal or not. In AdaBoost, they are intimately related to the exponential function but we will see shortly that there are other options.

Now that the algorithm is given, the pending issues in order to actually perform its implementation lie in the choice of a base class $\mathcal{G}$ of classifiers, an algorithm performing the extraction of a single base classifier at each step by minimizing the weighted empirical classification error, a stopping rule (number T of iterations). Relatively simple choices at this stage (e.g. take $\mathcal{G}$ as the class of decision trees and run AdaBoost for $T \simeq 100$ steps) already provide high performance on most of the benchmark classification data sets ([95]). However, it has been argued that resistance to overfitting is due to the fact that AdaBoost converges slowly and therefore, overfitting could occur if T was taken large enough, as confirmed by simulation experiments [63]. Indeed, it is easy to show that boosting overfits even for the simplest base class of decision stumps by considering noisy artificial data sets (see Figure 1.2 in Section 1.5). Machine Learning people have developed rules-of-thumb from their longtime experience on the subject in order to deal with these issues. We claim that applications-oriented statistical theory can provide a valuable guide in order to make the relevant choices and substantially improve existing algorithms.

One of the most interesting attempts to explain the success of boosting methods points out that they tend to maximize the *margin* of the correctly classified points. The margin of a real-valued function f on a training example $(X, Y) \in \mathcal{X} \times \{-1, +1\}$ is defined by $Yf(X)$. The margin can be interpreted as an indicator of confidence of the prediction based on f. The arguments developed in this direction were based on margin-based bounds for the probability of misclassification, see Schapire, Freund, Bartlett, and Lee [99], Koltchinskii and Panchenko [76]. However, as pointed out by Breiman [36], these bounds alone do not completely explain the efficiency of these methods (see also Freund and Schapire [59]). Boosting algorithms have also been found explicitly related to additive logistic regression by Friedman, Hastie, and Tibshirani [60] and Bühlmann and Yu [38]. This connection points out that boosting methods effectively minimize an empirical loss functional (different from the probability of misclassification). This property has also been pointed out in slightly different contexts by Breiman [36], Mason, Baxter, Bartlett, and Frean [86], and Collins, Schapire, and Singer [43]. Our approach builds on the interpretation by Friedman, Hastie and Tibshirani [60] according to which some boosting algorithms, like AdaBoost, are implementations of a gradient descent method to minimize a special risk criterion:

$$A_n(f) = \frac{1}{n} \sum_{i=1}^{n} \exp(-Y_i f(X_i)) \ ,$$

where f belongs to the linear span of the base class $\mathcal{G}$.

Soon after this observation was made, Breiman [37] proved the consistency of an algorithm minimizing the ideal convex risk $A(f) = \mathbb{E} \exp(-Yf(X))$. However, the consistency of a minimizer of A_n over the linear span of a class $\mathcal{G}$ of classifiers with respect to the classification error $L(f) = \mathbb{P}\{Y \cdot f(X) < 0\}$ was still an open problem at this stage.

1.2 Convex Risk Minimization

For practical optimization, convexity is a blessing. But, beyond computational considerations, studying statistical aspects cannot be avoided in order to assess the efficiency

of boosting methods based on the minimization of the functional A_n. In this section, we briefly explain why it makes sense to replace the classification error by a convex risk functional in the optimization procedure. We point out that there is nothing special about using the exponential as a cost function and we will replace it by a general cost function φ in the sequel.

We will make the following assumption on the cost function:

Assumption 2 $\varphi : \mathbb{R} \to \mathbb{R}_+$ *is strictly convex, differentiable, strictly increasing with* $\varphi(0) = 1$, $\lim_{x\to-\infty} \varphi(x) = 0$.

We introduce the following notation for the convex risk functional which will be called the φ-*risk*:

$$A(f) = \mathbb{E}\varphi(-Y \cdot f(X)) \ .$$

In this section, we study a population version of the results which amounts to knowing the distribution P. We recall that the goal in classification is to minimize the classification error $L(g) = \mathbb{P}\{Y \neq g(X)\}$ over a class $\mathcal{G}$ of classifiers. Boosting actually outputs a real-valued decision function f over a class $\mathcal{F}$ by minimizing a different criterion, the φ-risk. Denote by g_f the classifier based on the decision function f defined by $g_f(x) = +1$ if $f(x) > 0$ and -1 otherwise. We observe that $L(g_f) = \mathbb{P}\{Y \cdot f(X) < 0\} \leq A(f)$ as soon as $\varphi(x) \geq \mathbb{I}_{[x>0]}$ but we need to show why minimizing $A(f)$ implies minimizing $L(g_f)$.

1.2.1 Optimal elements

It is well known ([52]) that the optimal elements in classification are the Bayes classifier $g^*(x) = 2\mathbb{I}_{\{\eta(x)>1/2\}} - 1$ and the Bayes rule $L^* = L(g^*)$. The first thing to do is to relate the optimal elements for the φ-risk $A(f) = \mathbb{E}\varphi(-Y \cdot f(X))$ to g^* and L^*.

We introduce the function

$$f^*(x) = \arg\min_{\alpha\in\mathbb{R}} \{\eta(x)\varphi(-\alpha) + (1-\eta(x))\varphi(\alpha)\} \ .$$

The next proposition shows that f^* is well-defined for all x with $\eta(x) \in (0,1)$ but it can take infinite values when $\eta(X) \in \{0,1\}$.

Proposition 3 (Lugosi and Vayatis [12]) *Consider either one of the following two cases:*

- *If* $\eta(X) \notin \{0,1\}$ *almost surely then, there exists a unique measurable function* f^* *such that*
$$A(f^*) \leq A(f) \quad \textit{for all functions } f.$$
Then the classifier $2\mathbb{I}_{\{f^*(x)>0\}} - 1$ *is just the Bayes classifier* g^*.
- *If* $\eta(X) \in \{0,1\}$ *almost surely then, we have*
$$\inf_f A(f) = 0 \, .$$

Thus by extending the range of f^* with its infinite limits and by taking the sign of f^*, we obtain the Bayes classifier g^*. A wishful thinking is that near optimization of A yields nearly optimal classifiers and, indeed, this is true.

Lemma 4 (Lugosi and Vayatis [12]) *Let φ be a cost function satisfying Assumption 2. Let f_n be an arbitrary sequence of functions such that*

$$\lim_{n\to\infty} A(f_n) = A^* \ ,$$

where $A^* = \inf_f A(f)$. Then the classifier $g_{f_n}(x) = 2\mathbb{I}_{\{f_n(x)>0\}} - 1$ has a probability of error converging to L^*.

For usual choices of φ, the optimal elements of the φ-risk minimization problem are obtained through straightforward computations:

- exponential cost $\varphi(x) = \exp(x)$: optimal function $f^*(x) = \frac{1}{2}\ln\left(\frac{\eta(x)}{1-\eta(x)}\right)$, optimal risk $A^* = 2\mathbb{E}\sqrt{\eta(X)(1-\eta(X))}$
- logit cost $\varphi(x) = \log_2(1+\exp(x))$: optimal function $f^*(x) = \ln\left(\frac{\eta(x)}{1-\eta(x)}\right)$, optimal risk $A^* = \mathbb{E}\left(-\eta(X)\log_2\eta(X) - (1-\eta(X))\log_2(1-\eta(X))\right)$.

However, the hinge loss $\varphi(x) = (1+x)_+$ used in Support Vector Machines does not satisfy the assumption because it is not differentiable. Though a direct treatment of this case is simple, it is a challenging issue to find conditions which will cover as many costs as possible.

1.2.2 Zhang's lemma

While exploring the properties of convex risk minimization methods, we realized that there was a function playing a key role in characterizing the pointwise minimum of the φ-risk as a function of the posterior probability η. For a given cost function φ, define this function by

$$\forall \eta \in [0,1] \ , \quad H(\eta) = \inf_{\alpha\in\mathbb{R}} \left(\eta\varphi(-\alpha) + (1-\eta)\varphi(\alpha)\right) \ .$$

For our purpose, we established some simple properties of H under our Assumption 2.

Lemma 5 (Lugosi and Vayatis [12]) *Let φ be a cost function satisfying Assumption 2. Then the function $H(\eta)$ defined for $\eta \in [0,1]$, is concave, symmetric around $1/2$, and $H(0) = H(1) = 0$, $H(1/2) = 1$.*

About the same time, Zhang [119] formulated an alternative assumption on the cost function φ which directly involved this function H and allowed to remove the prerequisite of differentiability.

Assumption 6 (Zhang [119]) *Let φ be a convex, nonnegative, increasing cost function with $\varphi(x) \geq \mathbb{I}_{[x>0]}$ for all $x \in \mathbb{R}$, $\varphi(0) = 1$, $\lim_{x\to-\infty}\varphi(x) = 0$, and such that there exist constants $c > 0$ and $s \geq 1$ satisfying, for any $\eta \in [0,1]$,*

$$\left|\frac{1}{2} - \eta\right|^s \leq c^s(1 - H(\eta)) \ .$$

Under this assumption, Zhang obtained a stronger result than our Proposition 4 which will turn out to be important when deriving statistical results as rates of convergence of boosting methods.

Lemma 7 (Zhang [119], Lugosi and Vayatis [12]) *Under Assumption 6, for any estimator* f,

$$L(g_f) - L^* \leq 2c(A(f) - A^*)^{1/s} .$$

In the examples we mentioned above, we have

- exponential cost $\varphi(x) = \exp(x)$: $c = 1/\sqrt{2}$ and $s = 2$
- logit cost $\varphi(x) = \log_2(1 + \exp(x))$: $c = 1/\sqrt{2}$ and $s = 2$
- hinge loss $\varphi(x) = (1 + x)_+$: $c = 1/2$ and $s = 1$.

Eventually, a complete and final description of minimal assumptions on the cost function in order to guarantee "classification-calibration" was achieved by Bartlett, Jordan, and MacAuliffe [27]. In the case of convex cost functions, a necessary and sufficient condition for classification-calibration is differentiability at 0 with $\varphi'(0) > 0$.

Now that we have justified the use of convex risk functionals for classification purposes, we can turn to the statistical aspects of φ-risk minimization.

1.3 Consistency of Boosting with Regularization

In our view of boosting, we do not take into account the iterative nature of AdaBoost but we rather focus on what this algorithm actually does. Following the interpretation of [60], we will call a *boosting method* an estimation method constructing an estimator $\hat{f}_n$ by the the minimization of the empirical risk functional

$$A_n(f) = \frac{1}{n} \sum_{i=1}^{n} \varphi(-Y_i f(X_i)) .$$

over a class $\mathcal{F}$ expressed as the linear span of a base class $\mathcal{G}$ of classifiers:

$$\mathcal{F} = \left\{ f = \sum_{j=1}^{N} w_j g_j \, : \, N \in \mathbb{N}, \, w_1, \ldots, w_N \geq 0, \, g_1, \ldots, g_N \in \mathcal{G} \right\} .$$

By standard VC theory, it is well known that such a strategy might be vain if the class $\mathcal{F}$ is too large. We will assume that the base class $\mathcal{G}$ is a VC class, but still its convex hull might not be a VC class. Consider, for instance, the base class $\mathcal{G}$ including indicators of lower left orthants in $\mathbb{R}^2$. Its VC dimension is 2 but taking the sign of convex combinations $\sum_{k \geq 1} \mathbb{I}_{C_k}/2^k$ with $C_k \in \mathcal{G}$ and reindexing the C_k's, one can shatter any set of points of the line $x + y = 1$ (this example can be found in [54]). Therefore, there is no guarantee that the linear span $\mathcal{F}$ might be of reasonable size. This observation indicates that (1) VC

dimension might not be the right concept to capture the complexity of boosting methods, (2) the class $\mathcal{F}$ may be truly large and some *regularization* has to be introduced.

In this section, we provide a possible setup for designing universally consistent boosting methods in which the complexity issue can be taken care of.

1.3.1 Setup

In order to state the consistency results, we need some more notations.

Denote by $\|w\|_1$ the ℓ_1-norm of vector $w = (w_1, \ldots, w_N) \in \mathbb{R}^N$. We introduce $\mathcal{F}_\lambda$ the symmetric convex hull of the base class $\mathcal{G}$:

$$\mathcal{F}_\lambda = \left\{ f = \sum_{j=1}^{N} w_j g_j \,:\, N \in \mathbb{N},\, g_1, \ldots, g_N \in \mathcal{G}\,,\, \|w\|_1 \leq \lambda \right\} .$$

We note that the linear span of $\mathcal{G}$ may be represented as:

$$\mathcal{F} = \bigcup_{\lambda > 0} \mathcal{F}_\lambda \,.$$

The scale parameter λ will play the role of a smoothing parameter. Although scaling the estimator f has no effect on the corresponding decision function (performance is unchanged since $L(f) = L(\lambda f)$ for all $\lambda > 0$), the introduction of the parameter λ is indeed a decisive step to design consistent strategies in boosting. The complexity parameter λ reflects here the total variation of the candidate estimators in $\mathcal{F}_\lambda$. As the target estimator f^* can be wildly oscillating and even have unbounded total variation, large values of λ offer more flexibility of approximation at the price of making the estimation problem more difficult.

Now, introduce, for all $\lambda > 0$,

$$\varphi_\lambda(x) = \varphi(\lambda x) \,.$$

Denote the empirical and expected loss functional associated with the cost function φ_λ by A_n^λ and A^λ, that is,

$$A_n^\lambda(f) = \frac{1}{n} \sum_{i=1}^{n} \varphi_\lambda(-Y_i \cdot f(X_i)) \quad \text{and} \quad A^\lambda(f) = \mathbb{E}\varphi_\lambda(-Y \cdot f(X)) \,.$$

Note that on $[-1, 1]$ the function φ_λ is Lipschitz with constant $\lambda\varphi'(\lambda)$. If $\lambda = 1$, we simply write $A_n(f)$ and $A(f)$ instead of $A_n^\lambda(f)$ and $A_n^\lambda(f)$. Observe that

$$A_n^\lambda(f) = A_n(\lambda f) \qquad A^\lambda(f) = A(\lambda f) \,.$$

Hence, minimizing $A^\lambda(f)$ over $\mathcal{F}_1$ is equivalent to minimizing $A(f)$ over the scaled class $\mathcal{F}_\lambda$. It is worth noticing at this point that the original AdaBoost algorithm attempts to minimize the functional A in the linear span of $\mathcal{G}$. In contrast to this, here we consider a

family of optimization problems (minimizing various functionals A^λ) over the convex hull $\mathcal{F}_1$ of $\mathcal{G}$.

Now let $\hat{f}_n^\lambda$ denote a function in $\mathcal{F}_1$ which minimizes the empirical loss

$$A_n^\lambda(f) = \frac{1}{n}\sum_{i=1}^{n} \varphi_\lambda(-Y_i \cdot f(X_i))$$

over $f \in \mathcal{F}_1$.

1.3.2 Complexity control

In the present subsection, we briefly discuss the complexity issue in the case λ is fixed. First, applying the standard bias-variance decomposition to our problem, we have, for the empirical φ-risk estimator $\hat{f}_n^\lambda$ over the class $\mathcal{F}_\lambda$:

$$A(\hat{f}_n^\lambda) - A^* \leq 2 \sup_{f\in\mathcal{F}_\lambda} |A_n(f) - A(f)| + \inf_{f\in\mathcal{F}_\lambda} A(f) - A^*$$

The complexity issue concerns the trade-off between these two terms, estimation error and approximation error. The key here is to use λ as a smoothing parameter letting it grow to infinity in order to make the approximation error term small but in a controlled fashion not too spoil the estimation error. This being said, we still have to explain how to deal with the supremum over $\mathcal{F}_\lambda$ which can be a massive class of functions.

It turns out that it can be controlled in terms of the VC dimension V of the base class $\mathcal{G}$ according to the following lemma.

Lemma 8 (Koltchinskii and Panchenko [76], Lugosi and Vayatis [12]) *For any n and $\lambda > 0$,*

$$\mathbb{E}\sup_{f\in\mathcal{F}} \left|A^\lambda(f) - A_n^\lambda(f)\right| \leq 4\lambda\varphi'(\lambda)\sqrt{\frac{2V\ln(4n+2)}{n}}.$$

The key to this result lies in standard techniques from empirical processes such as symmetrization and a contraction principle from [78], but also the control of a now celebrated quantity known as the Rademacher average over the class $\mathcal{F}$ of functions. Indeed, it is now common knowledge in machine learning that the learning complexity is better captured by the Rademacher average of a class of functions rather than combinatorial or metric capacities (see [73], [25]). In our case, we can benefit of the linearity of the class $\mathcal{F}$ which reduces its complexity to that of $\mathcal{G}$.

An additional comment after this lemma concerns the choice of the cost function φ. The previous result provides a theoretical argument in favor of the logit function compared to the exponential function. In [12], we introduced the following function

$$\psi(x) = \begin{cases} \exp(x) & , \text{ if } x < 0 \\ \\ x + 1 & , \text{ if } x \geq 0 \end{cases}$$

which also satisfies $\psi'(\lambda) = 1$ as the logit function but mimics the exponential on correctly classified examples. We shall provide more insights on the choice of a cost function in Section 1.5.

1.3.3 Main consistency results

We are now in a position to formulate the main consistency results on regularized boosting methods. We first state a theorem on consistency of an idealized boosting procedure where the parameter λ is turned into a divergent deterministic sequence for which the speed of divergence is specified. Combining Lemma 8 with a simple concentration inequality, we can derive consistency in terms of the φ-risk under the following denseness assumption:

Assumption 9 *The distribution of* (X, Y) *and the class* $\mathcal{G}$ *are such that*

$$\textbf{(D)} \qquad \lim_{\lambda \to \infty} \inf_{f \in \mathcal{F}_\lambda} A(f) = A^* \,,$$

where $A^* = \inf_f A(f)$ *over all measurable functions* $f : \mathcal{X} \longrightarrow \mathbb{R}$.

We refer to Section 2.2 for more details on this assumption. Applying Zhang's Lemma, we eventually derive consistency in terms of convergence of the classification error towards the Bayes error. Note that we have universal consistency if we can exhibit a base class $\mathcal{G}$ such that for any distribution, Assumption 9 holds. We will see that there are many examples of such classes in Section 2.2.

Theorem 10 (Lugosi and Vayatis [12]) *Assume that Assumptions 6 and 9 hold. Assume also that* $\mathcal{G}$ *has a finite* VC *dimension. Let* $(\lambda_n)_{n \geq 1}$ *be a sequence of positive numbers satisfying*

$$\lambda_n \to \infty \quad \textit{and} \quad \lambda_n \varphi'(\lambda_n) \sqrt{\frac{\ln n}{n}} \to 0 \quad \textit{as } n \to \infty$$

and define the estimator $\hat{f}_n = \widehat{f}_n^{\lambda_n} \in \mathcal{F}_1$. *Then* $g_{\hat{f}_n}$ *is strongly Bayes-risk consistent, that is,*

$$\lim_{n \to \infty} L(g_{\hat{f}_n}) = L^* \quad \textit{almost surely.}$$

In the previous theorem, the sequence of estimators $\hat{f}_n = \widehat{f}_n^{\lambda_n}$ requires for each n to minimize the functional $A_n^{\lambda_n}$ over $\mathcal{F}$, for a predetermined sequence $(\lambda_n)_n$. Of course, it would be much more practical to handle the choice of λ on the basis of the sample. The following theorem shows that consistency remains true for a data-dependent regularized choice of the smoothing parameter λ:

Theorem 11 (Lugosi and Vayatis [12]) *Assume that Assumptions 6 and 9 hold. Assume also that the base class* $\mathcal{G}$ *has a finite VC dimension* V. *For any divergent sequence* $(\lambda_n)_{n \geq 1}$ *of positive numbers, let*

$$\hat{f}_n = \arg\min_{k \geq 1} \tilde{A}_n^{\lambda_k}(\widehat{f}_n^{\lambda_k}) \,,$$

where

$$\tilde{A}_n^{\lambda_k}(f) = A_n^{\lambda_k}(f) + 4\lambda_k \varphi'(\lambda_k) \sqrt{\frac{2V \ln(4n+2)}{n}} + \lambda_k \varphi'(\lambda_k) \sqrt{\frac{\ln(6n^2k^2/\pi^2)}{n}} \,,$$

and

$$\hat{f}_n^{\lambda_k} = \arg\min_{f \in \mathcal{F}} A_n^{\lambda_k}(f)\,.$$

Then $\hat{f}_n$ *is strongly Bayes-risk consistent, that is,*

$$\lim_{n\to\infty} L(\hat{f}_n) = L^* \quad \textit{almost surely.}$$

For more recent results on the consistency of related methods, including support vector machines, we refer to Bartlett, Jordan, and MacAuliffe [27], Mannor, Meir, and Zhang [82], Steinwart [105].

The take-home message of the consistency results above lies in the following practical procedure:

1. choose a class $\mathcal{G}$ of base elements (classifiers)
2. solve the optimization problem of minimizing the φ-risk over the λ-blown-up convex hull for various λ's
3. select the "best" value of the regularization parameter λ.

Apparently, the AdaBoost algorithm seems to take care of these issues all at once in a clever way. However, one may wonder about how suboptimal each of these choices can be. In the next section, we will focus on the optimization of the convex risk functional for fixed λ (Point 2).

1.4 Online Version: The Mirror Averaging Algorithm

A major computational limit in boosting is the extraction of each single base classifier out of a (weighted) empirical risk minimization step. The trade-off between representation capacity of the base class and computational constraints is generally in favor of the latter. A somewhat different approach is the *aggregation* framework ([71], [107], [96] and the references therein). There, it is assumed that we have a *finite* pool of base classifiers which may have been obtained by various estimation procedures (in particular, they could be the weak classifiers collected along one run of a boosting algorithm, or they could be several outputs of boosting on various subsamples, and so on). Aggregation is concerned with finding an optimal linear or convex combination of these base classifiers.

Another limitation (apart from computational issues) is the data generation process handled by boosting. Boosting algorithms belong to the family of *batch algorithms* where all the data have to be given at once before running them. But building sequential (or online) procedures is also of major interest when the data are delivered one-by-one (e.g. real-time applications). We refer to the book by Cesa-Bianchi and Lugosi [40] and the references therein for an up-to-date account on this topic.

In the sequel, we present a learning algorithm inspired by stochastic approximation theory. This algorithm is called the *mirror averaging algorithm*. It achieves efficient convex risk minimization over the λ-simplex and copes with the two issues mentioned above.

1.4.1 Setup and notations

Consider the setting of classification where the data are modelled by a pair (X, Y) with $X \in \mathcal{X}$ being an observation vector and $Y \in \{-1, +1\}$ a binary label. Boosting and SVM algorithms are related to the minimization of a functional

$$A(f) = \mathbb{E}\varphi(-Yf(X))$$

where φ is a convex nonnegative cost function (standard choices are exponential, logit or hinge loss) and f belongs to a given class of combined classifiers. The aggregation problem consists in finding the best linear combination of elements from a finite set of classifiers $\{h_1, \dots, h_M\}$ with $h_j : \mathcal{X} \to \{-1, +1\}$. Taking compact notations, it means that we search for f of the form $f = \theta^T H$ with H denoting the vector-valued function whose components are these base predictors:

$$H(x) = (h_1(x), \dots, h_M(x))^T \ ,$$

and assume θ belongs to a decision set Θ. Following the ideas developed above, we set, for $\lambda > 0$ and an integer $M \geq 2$:

$$\Theta_{M,\lambda} = \left\{ \theta = (\theta^{(1)}, \dots, \theta^{(M)})^T \in \mathbb{R}_+^M \ : \ \sum_{i=1}^M \theta^{(i)} = \lambda \right\} \ .$$

and take in the sequel $\Theta = \Theta_{M,\lambda}$. Hence, the problem boils down to minimize the function

$$A(\theta) = \mathbb{E}\varphi(-Y \cdot \theta^T H(X))$$

over Θ.

We can present our algorithm in a slightly more general setting by taking $Z = (X, Y)$, a random variable with values in a measurable space $(\mathcal{Z}, \mathcal{A})$, and $Q(Z, \theta) = \varphi(-Y\theta^T H(X))$. We can assume the loss function $Q : \Theta \times \mathcal{Z} \to \mathbb{R}_+$ is such that the random function $Q(\cdot, Z) : \Theta \to \mathbb{R}_+$ is convex for almost all Z. In what follows, we define the convex risk function $A : \Theta \to \mathbb{R}_+$ to be minimized as follows:

$$A(\theta) = \mathbb{E}\, Q(\theta, Z) \ .$$

Assume that a training sample is given in the form of a sequence $(Z_1, \dots, Z_{t-1})$, where each Z_i has the same distribution as Z. We assume for simplicity that the training sequence is i.i.d. though this assumption can be weakened.

We propose to minimize the convex target function A over the decision set Θ on the basis of the stochastic subgradients of Q:

$$u_i(\theta) = \nabla_\theta Q(\theta, Z_i) \, , \quad i = 1, 2, \dots ,$$

Note that the expectations $\mathbb{E}\, u_i(\cdot)$ belong to the subdifferential of $A(\cdot)$.

In the sequel, we will characterize the accuracy of an estimate $\widehat{\theta}_t$ of the minimizer of A by the excess risk:

$$\mathbb{E}\, A(\widehat{\theta}_t) - \min_{\theta \in \Theta} A(\theta)$$

where the expectation is taken over the sample $(Z_1, \ldots, Z_{t-1})$.

The mirror descent algorithm described below requires the choice of a so-called proxy function V which, for some $\alpha > 0$, is an α-strongly convex function with respect to the ℓ_1-norm defined on Θ, i.e.

$$V(sx + (1-s)y) \leq sV(x) + (1-s)V(y) - \frac{\alpha}{2}s(1-s)\|x-y\|_1^2$$

for all $x, y \in \Theta$ and any $s \in [0,1]$.

For any $\beta > 0$, we call β-*conjugate* of V the following convex transform:

$$\forall z \in \mathbb{R}^M, \quad W_\beta(z) = \sup_{\theta\in\Theta} \left\{ -z^T\theta - \beta V(\theta) \right\} .$$

Example 1 The entropic proxy function is defined by:

$$\forall \theta \in \Theta, \quad V(\theta) = \lambda \ln(M/\lambda) + \sum_{j=1}^{M} \theta^{(j)} \ln \theta^{(j)} , \tag{1.1}$$

which has its minimum at $\theta_0 = (\lambda/M, \ldots, \lambda/M)^T$. It is easy to check that this function is α-strongly convex with respect to the norm $\|\cdot\|_1$ with parameter $\alpha = 1/\lambda$. We can easily derive the corresponding β-conjugate:

$$W_\beta(z) = \lambda\,\beta\, \ln\left(\frac{1}{M} \sum_{k=1}^{M} e^{-z^{(k)}/\beta} \right), \quad \forall z \in \mathbb{R}^M,$$

which has a Lipschitz-continuous gradient with respect to the ℓ_1-norm in the dual space, namely:

$$\|\nabla W_\beta(z) - \nabla W_\beta(\tilde{z})\|_1 \leq \frac{\lambda}{\beta}\|z-\tilde{z}\|_\infty , \quad \forall z, \tilde{z} \in \mathbb{R}^M. \quad \blacksquare \tag{1.2}$$

In this presentation, we will focus on the particular algorithm based on the entropic proxy function but we mention that our results apply for a generic algorithmic scheme which takes advantage of the general properties of convex transforms (see [10] for details). The key property in the proof is the inequality (1.2).

1.4.2 Algorithm and main result

The mirror averaging algorithm is a stochastic gradient algorithm in the dual space. The idea of mirror descent algorithms in optimization theory goes back to Nemirovski and Yudin [93] and it has been proved to be a powerful idea both from a theoretical and a practical viewpoint (see e.g. [29], [30]). In particular, mirror descent algorithms should systematically be favored over direct gradient descent in high-dimensional problems.

The mirror descent procedure goes as follows. At each iteration i, a new data point (X_i, Y_i) is observed and there are two updates: one is the value ζ_i as the result of the stochastic gradient descent in the dual space, the other is the update of the parameter θ_i which is the "mirror image" of ζ_i. In order to tune the algorithm properly, we need two fixed positive sequences $(\gamma_i)_{i\geq 1}$ (stepsize) and $(\beta_i)_{i\geq 1}$ (temperature) such that $\beta_i \geq \beta_{i-1}$. The *mirror averaging algorithm* is as follows:

Algorithm 12 (JUDITSKY, NAZIN, TSYBAKOV, AND VAYATIS [10])

- *Fix the initial values* $\theta_0 \in \Theta$ *and* $\zeta_0 = 0 \in \mathbb{R}^M$.
- *For* $i = 1, \ldots, t-1$, *do*

$$\begin{aligned} \zeta_i &= \zeta_{i-1} + \gamma_i u_i(\theta_{i-1}) \ , \\ \theta_i &= -\nabla W_{\beta_i}(\zeta_i) \ . \end{aligned} \tag{1.3}$$

- *Output at iteration* t *the following convex combination:*

$$\widehat{\theta}_t = \frac{\sum_{i=1}^t \gamma_i \theta_{i-1}}{\sum_{j=1}^t \gamma_j} \ .$$

Given the observations of the stochastic subgradient $u_i(\theta)$, particular choices of the proxy function V, of the stepsize and temperature parameters, will determine the algorithm completely. We discuss these choices in greater detail in [10]. We focus here on the entropic proxy function and consider a nearly optimal choice for the stepsize and temperature parameters which is the following:

$$\gamma_i \equiv 1 \ , \quad \beta_i = \beta_0 \sqrt{i+1} \ , \quad \forall i \ , \quad \beta_0 > 0 \ .$$

We now provide some heuristics underlying this algorithm. Suppose that we want to minimize a convex function $\theta \mapsto A(\theta)$ over a convex set Θ. If $\theta_0, \ldots, \theta_{t-1}$ are the available search points at iteration t, we can provide the affine approximations φ_i of the function A defined, for $\theta \in \Theta$, by

$$\varphi_i(\theta) = A(\theta_{i-1}) + (\theta - \theta_{i-1})^T \nabla A(\theta_{i-1}), \qquad i = 1, \ldots, t \, .$$

Here $\theta \mapsto \nabla A(\theta)$ is a vector function belonging to the subdifferential of A. Taking a convex combination of the φ_i's, we obtain an averaged approximation of $A(\theta)$:

$$\bar{\varphi}_t(\theta) = \frac{\displaystyle\sum_{i=1}^t \gamma_i \left(A(\theta_{i-1}) + (\theta - \theta_{i-1})^T \nabla A(\theta_{i-1}) \right)}{\displaystyle\sum_{i=1}^t \gamma_i} \, .$$

At first glance, it would seem reasonable to choose as the next search point a vector $\theta_t \in \Theta$ minimizing the approximation $\bar{\varphi}_t$, i.e.,

$$\theta_t = \arg\min_{\theta \in \Theta} \bar{\phi}_t(\theta) = \arg\min_{\theta \in \Theta} \ \theta^T \left(\sum_{i=1}^t \gamma_i \nabla A(\theta_{i-1}) \right) .$$

However, this does not make any progress, because our approximations are "good" only in the vicinity of search points $\theta_0, \ldots, \theta_{t-1}$. Therefore, it is necessary to modify the criterion, for instance, by adding some penalty $B_t(\theta, \theta_{t-1})$ to the target function in order

to keep the next search point θ_t in the vicinity of previous one θ_{t-1}. Thus, one chooses the point

$$\theta_t = \arg\min_{\theta\in\Theta} \left[\theta^T \left(\sum_{i=1}^{t} \gamma_i \nabla A(\theta_{i-1}) \right) + B_t(\theta, \theta_{t-1}) \right] .$$

Our algorithm corresponds to a specific type of penalty $B_t(\theta, \theta_{t-1}) = \beta_t V(\theta)$, where V is the proxy function. Also note that the vector-function ∇A is not available. Therefore, we replace the non-observed gradients $\nabla A(\theta_{i-1})$ by the stochastic subgradients $u_i(\theta_{i-1})$. This yields a new definition of the t-th search point:

$$\theta_t = \arg\min_{\theta\in\Theta} \left[\theta^T \left(\sum_{i=1}^{t} \gamma_i u_i(\theta_{i-1}) \right) + \beta_t V(\theta) \right] = \arg\max_{\theta\in\Theta} \left[-\zeta_t^T \theta - \beta_t V(\theta) \right] ,$$

where

$$\zeta_t = \sum_{i=1}^{t} \gamma_i u_i(\theta_{i-1}) .$$

By an argument borrowed from convex analysis, we can show that the solution to the latter problem reads as $\theta_t = -\nabla W_{\beta_t}(\zeta_t)$.

We can now state our rate of convergence result.

Theorem 13 (Juditsky, Nazin, Tsybakov, and Vayatis [10]) *Assume that the loss function Q satisfies the following boundedness condition:*

$$\sup_{\theta\in\Theta} \mathbb{E} \|\nabla_\theta Q(\theta, Z)\|_\infty^2 \le L^2 < \infty .$$

Fix also $\beta_0 = L/\sqrt{\ln M}$. Then, for any integer $t \ge 1$, the excess risk of the mirror descent estimate $\widehat{\theta}_t$ with entropic proxy function satisfies the following bound:

$$\mathbb{E}\, A(\widehat{\theta}_t) - \min_{\theta\in\Theta} A(\theta) \le 2L\lambda (\ln M)^{1/2} \frac{\sqrt{t+1}}{t} .$$

Example 2 In the convex risk minimization setup for classification, we can take for instance φ to be nonincreasing. It is easy to see that $L = \varphi'(\lambda)$. ∎

The rate of convergence of order $\sqrt{\ln M}/\sqrt{t}$ is the expected one if no particular assumption is made on the distribution. Batch procedures based on minimization of the empirical convex risk functional present a similar rate. From the statistical point of view, there is no remarkable difference between batch and our mirror descent procedure. On the other hand, from the computational point of view, our procedure is quite comparable with the direct stochastic gradient descent. However, the mirror-descent algorithm presents two major advantages as compared both to batch and to direct stochastic gradient: (i) its behavior with respect to the cardinality of the base class is better than for direct stochastic gradient descent (of the order of $\sqrt{\ln M}$ in the Theorem, instead of M or $\sqrt{M}$ for direct stochastic gradient, see [120]); (ii) mirror-descent presents a higher efficiency especially in high-dimensional problems as its algorithmic complexity and memory requirements are of

strictly smaller order than for corresponding batch procedures (see [71] for a comparison). Moreover, by using the techniques of [71] and [107] it is not hard to prove minimax lower bound on the excess risk $\mathbb{E}\,A(\widehat{\theta}_t) - \min_{\theta\in\Theta_{M,\lambda}} A(\theta)$ having the order $(\ln M)^{1/2}/\sqrt{t}$ for $M \geq t^{1/2+\delta}$ with some $\delta > 0$. This indicates that the upper bound of the Theorem is rate optimal for such values of M.

We eventually mention that the mirror averaging algorithm is related to the exponentiated gradient descent algorithm proposed by Kivinen and Warmuth [72]. We also refer to the work of Cesa-Bianchi, Conconi, and Gentile [39] on the analysis of such algorithms.

1.5 Simulation results

We consider binary classification of artificial data and the weak learners are all decision stumps. We recall that a *decision stump* is a linear classifier whose separating hyperplane is orthogonal to one of the axes. We used synthetic multi-dimensional data from the "twonorm", "threenorm" and "ringnorm" generators (see Breiman [36]). These problems are expected to be of increasing difficulty for the class of convex combinations obtained from decision stumps. We considered relatively small sample sizes for the training set (n between 100 and 500).

1.5.1 Implementation of regularized boosting

We propose a first series of experiments in order to understand how the theoretical analysis presented in Section 1.3.3 can efficiently be converted into practical strategies. Indeed, the results presented above show that there are two elements governing the consistency of boosting methods: (i) the choice of the cost function φ, (ii) the tuning of the smoothing parameter λ. However, universal consistency (or particular non-consistency) can hardly be checked empirically. Therefore, we focus here on a rather qualitative analysis aiming at making clear that the performance of efficient model selection algorithms is highly sensitive to the tuning of the smoothing parameter λ depending on the noise level and on the difficulty of the classification problem.

We have implemented the following algorithms described in [86] (to which we refer for detailed description and convergence properties):

- **MarginBoost** - The algorithm MarginBoost implements a gradient descent in the *linear span* of the class $\mathcal{G}$ to minimize a criterion of the form $\frac{1}{n}\sum_{i=1}^n \varphi(-Y_i f(X_i))$, for $f = \sum_j w_j g_j$ with $g_j \in \mathcal{G}$. In this case, the parameter λ is interpreted as the sum of the unnormalized weights (their ℓ_1-norm). Note that the original AdaBoost algorithm is a particular case of MarginBoost with exponential cost function.

- **MarginBoost.L_1** - This algorithm implements a gradient descent in the *convex hull* of the class $\mathcal{G}$ to minimize $\frac{1}{n}\sum_{i=1}^n \varphi(-\lambda Y_i f(X_i))$, for $f = \sum_j w_j g_j$ with $\sum_j w_j = 1$.

In the experiments, we track the generalization error and the optimal value of the cost functional as functions of the smoothing parameter λ, for fixed samples. More precisely, for each λ, the combined classifier $\widehat{f}_n^{\lambda}$ is constructed by the MarginBoost.L_1 algorithm after

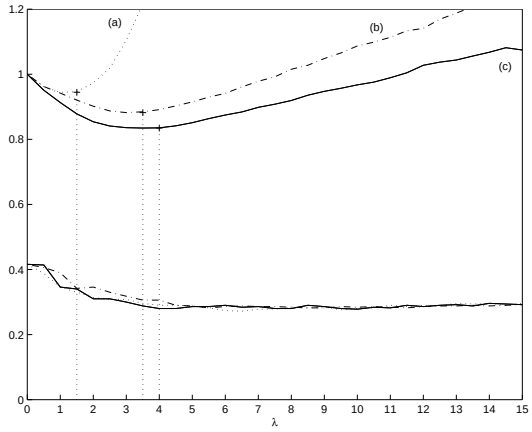


Figure 1.1: Threenorm. $\eta = 0.1$. $n = 100$. $m = 500$. Plots of the cost $A^{\lambda}(\widehat{f}_n^{\lambda})$ (upper curves) and test error (lower curves) for various cost functions (a) exp, (b) logit, (c) ψ.

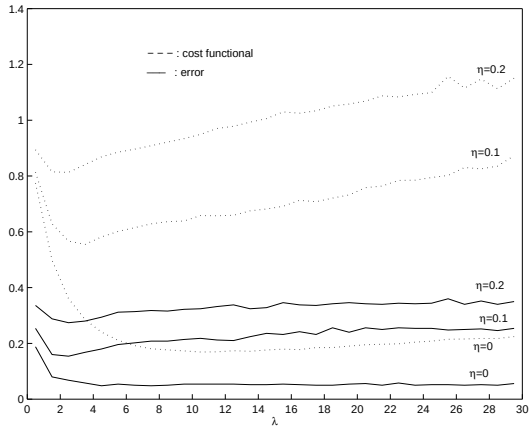


Figure 1.2: Twonorm. Cost $\varphi = \psi$. $n = 100$. $m = 500$. Plots of $A^{\lambda}(\widehat{f}_n^{\lambda})$ (dotted lines) and of the test error (solid lines) for levels of noise $\eta = 0, 0.1, 0.2$.

300 iterations, on the basis of training samples of size n. We then estimate the expected cost $A^{\lambda}(\widehat{f}_n^{\lambda})$ and the generalization error $L(\widehat{f}_n^{\lambda})$ on a test set of size m. Moreover, we have added a uniform label noise (probability of flipping the label) denoted by η.

We have focused on the following topics:

- the influence of the choice of the cost function (see Figure 1.1)

  For a fixed sample, the choice of a particular cost function appears to have a notable impact on the generalization error performed by the corresponding boosting algorithm. In the long run though, all of these choices lead to a consistent method if the label noise is small. In the sequel we report only experiments with cost function $\varphi = \psi$ (see Section 1.3.2) which seems to behave slightly better on this particular range of sample sizes.

- Sensitivity to the level of label noise (see Figure 1.2)

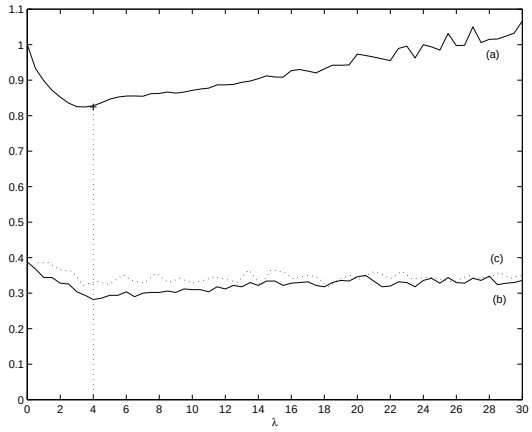


Figure 1.3: Threenorm. Cost $\varphi = \psi$. $\eta = 0.1$. $n = 100$. $m = 500$. Plot of (a) $A^{\lambda}(\widehat{f}_n^{\lambda})$ with MarginBoost.L_1 for various λ's, (b) test error with MarginBoost.L_1 for various λ's, (c) test error with one run of MarginBoost (unnormalized weights) where the test error is plotted as a function of the sum of the weights of the combined classifier denoted by λ_T.

In these experiments, the observations vectors X_i are fixed and the labels Y_i are exposed to a constant level of noise η. The algorithms are run for different levels of label noise. The overfitting phenomenon can be observed even for small values of λ. The general effect is that the increase of the level of label noise η results in a decrease of the optimal λ. Moreover, the fact that the minimizer of the cost functional tracks so well the optimal classifier needs to be mentioned.

- Comparison with AdaBoost (see Figure 1.3)

  We think that these experiments provide some interesting insights on how the original AdaBoost algorithm works. Indeed, we can give a comparison by representing AdaBoost performance as a function of the norm of the weights in the combined classifier (instead of the number of iterations). Note that here, in order to make fair comparisons, we implemented AdaBoost using MarginBoost with cost function $\varphi = \psi$. This algorithm constructs iteratively a combined classifier associated to the estimator $f_T = \sum_{t=1}^{T} w_t g_t$ with $g_t \in \mathcal{G}$ (step T) and w_t are positive weights (no normalization). Therefore, at each step T, MarginBoost outputs some element f_T from the class $\mathcal{F}_{\lambda_T}$ where $\lambda_T = \sum_{t=1}^{T} w_t$. In Figure 1.3, we keep track of the test error of MarginBoost along the iterations with respect to λ_T. On this simple example, it turns out that AdaBoost constructs very quickly a classifier with the "optimal" complexity but that the intrinsic discretization of the method (at least in its original version) does not allow it to approximate the optimal generalization error too well.

1.5.2 The mirror averaging algorithm

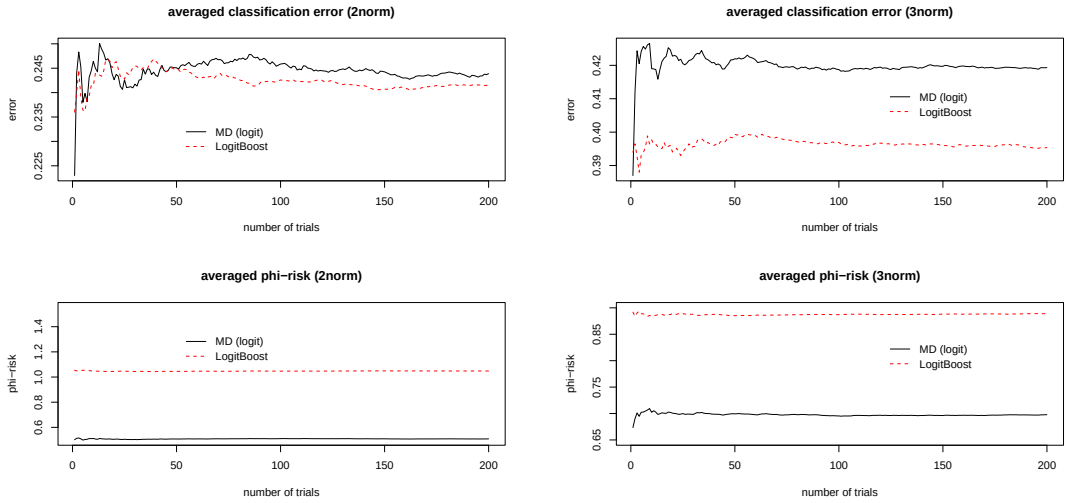
The second series of simulation experiments aims at assessing the performance of the mirror averaging algorithm proposed in Section 1.4. The observation vectors X_i are drawn

in a high-dimensional space ($\mathbb{R}^{200}$), the sample sizes are $n = 100$ for the training set and $m = 1000$ for the test set. Each experiment is repeated for 200 different training samples and the various errors are computed by taking the average over these distinct samples. The cardinality of base classifiers involved in the aggregation is $M = 8000$ and the ℓ_1-norm of the coefficient vector is $\lambda = 25$.

The cost function used in these experiments is the logit function and the results are compared to those of a LogitBoost algorithm [50] with 100 iterations. Given that the mirror descent algorithm outputs en element of $\mathcal{F}_\lambda$ with $\lambda = 25$, the LogitBoost output is weighted accordingly in order to compare comparable objects.

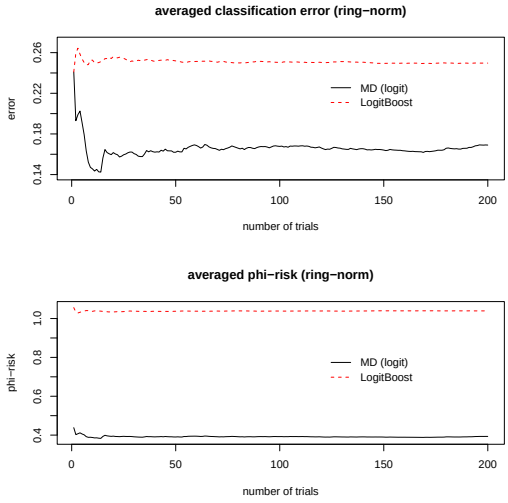
We present here two types of plots: one is the average of the test error (φ-risk or classification error) as a function of the trials, and the other is the estimated density (using kernel smoothing) of these test errors.

The results (Figures 1.4 and 1.5) show that the Mirror Averaging Algorithm always perform better than the LogitBoost Algorithm as a φ-risk minimizer, however the comparison in terms of classification error is ambivalent. The algorithms are comparable on the easy problem (twonorm), LogitBoost is more efficient on the intermediate problem (threenorm), while Mirror Averaging wins in the difficult problem (ringnorm). Figure 1.5 also provides some information about the robustness of each method with respect to the two notions of error.



(a) Twonorm

(b) Threenorm



(c) Ringnorm

Figure 1.4: Averaged errors in terms of the φ-risk and the classification error. [Simulations performed by Philippe Rigollet]

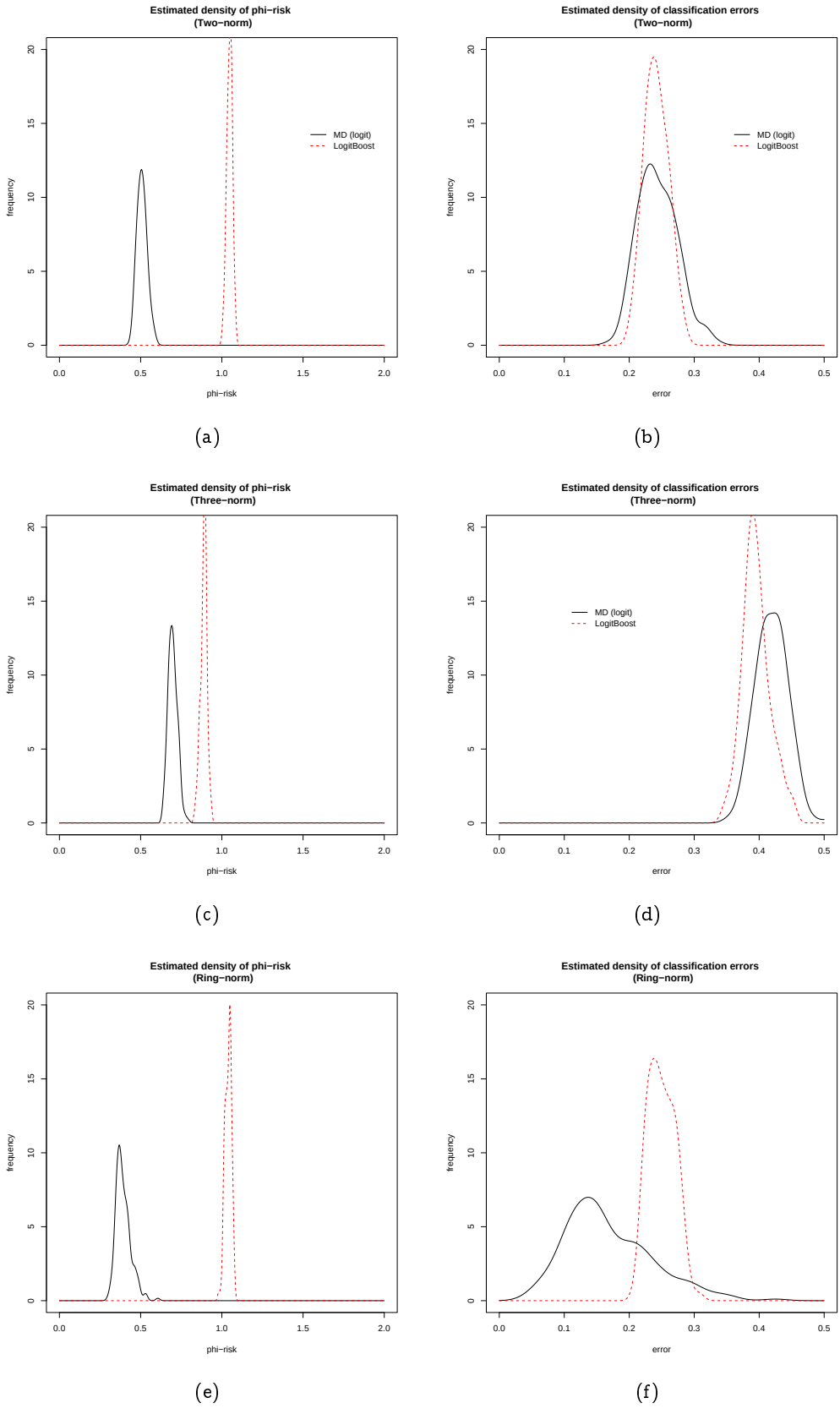


Figure 1.5: Spread of empirical results for the φ-risk (left column) and for the classification error (right column). [Simulations performed by Philippe Rigollet]

Chapter 2

Refinements for Boosting Theory

In the previous chapter, we saw how boosting can be interpreted as an instance of a penalized M-estimation procedure. In the field of nonparametric statistics, such procedures have been studied for a long time and they have been a field for intensive applications of empirical processes techniques (see for instance the books by van der Vaart and Wellner [110] and van de Geer [109] and the references therein). However, it is noteworthy that the problems emerging from learning theory present some specific features:

1. **Finite sample.** Beyond consistency issues, learning theory is mainly concerned with formulating non-asymptotic statements on the performance of a learning method, such as (exact) confidence intervals.

2. **Approximation properties.** Learning algorithms are particularly invoked in high-dimensional problems and thus, computational tractability governs the choise of base elements. Decision stumps and decision trees are common dictionaries used in boosting but Approximation Theory in this framework is still at an early stage.

3. **Complexity.** Another specificity is the centrality of complexity control in learning while this is a mere technical assumption in nonparametric statistics. Identifying adequate (and practical!) complexity measures in learning problems is still a challenging issue.

This chapter partly explores these topics. We first present an oracle inequality for regularized boosting methods. Then, we provide some preliminary results and comments about the approximation issue and show which complexity parameter governs the behavior of boosting with decision stumps.

2.1 Oracle inequalities and fast rates for boosting methods

We now turn to the study of excess risk bounds in order to better understand the properties of boosting. In the previous chapter, we established the importance of regularization in designing consistent boosting methods. The idea of converting boosting algorithms into penalized M-estimators raises the question of the form of the penalty. In order to derive model selection results for regularized boosting procedures, we have followed

the methodology described by Massart [88]. We apply a general model selection theorem from [32] and our challenge was to check the underlying fundamental assumptions: mainly, variance control and local complexity control.

An essential condition is the one of variance control and we have formulated a sufficient condition on the cost function to guarantee it. For any twice differentiable cost function φ, we define the following quantity:

$$L_\varphi = 0 \vee \max_{x \in \mathbb{R}} \left(\frac{2(\varphi'(x) + \varphi'(-x))}{\frac{\varphi''}{\varphi'}(x) + \frac{\varphi''}{\varphi'}(-x)} - (\varphi(x) + \varphi(-x)) \right) .$$

Lemma 14 (Blanchard, Lugosi, and Vayatis [3]) *Assume $\varphi : \mathbb{R} \to \mathbb{R}_+$ is a twice differentiable, strictly increasing and strictly convex function. If $L_\varphi < \infty$, then for any function $f \in \mathcal{F}_\lambda$, we have*

$$\mathbb{E}[(\varphi(-Y \cdot f(X)) - \varphi(-Y \cdot f^*(X)))^2] \leq (\varphi(\lambda) + \varphi(-\lambda) + L_\varphi)\mathbb{E}[\varphi(-Y \cdot f(X)) - \varphi(-Y \cdot f^*(X))] .$$

Another important condition, is the control of the local modulus of continuity of the empirical process indexed by the loss class. After symmetrization, it is sufficient to provide an upper bound of the local Rademacher average [26] which can be done using a result by Mendelson [92] under a polynomial behavior of the metric entropy. This can be guaranteed by Theorem 2.6.9 from [110]. Solving a fixed point equation for this upper bound to determine the right scaling in the localization, we finally derive the right order for the penalty term in designing a regularized boosting procedure.

Theorem 15 (Blanchard, Lugosi, and Vayatis [3]) *Assume that the cost function φ is twice differentiable, satisfies Assumption 2 and is such that the constant L_φ is finite. Define*

$$R(\lambda, n) = (V+2)^{\frac{V+2}{V+1}} ((L_\varphi + 2)\varphi(\lambda))^{\frac{1}{V+1}} (\lambda \varphi'(\lambda))^{\frac{V}{V+1}} n^{-\frac{1}{2}\frac{V+2}{V+1}} ,$$

$$b(\lambda) = (L_\varphi + 2)\varphi(\lambda) ,$$

and let $(\lambda_k)_{k \in \mathbb{N}}$ be an increasing sequence in $(1, +\infty)$ such that $\sum_{k \in \mathbb{N}} \lambda_k^{-\alpha} \leq 1$ for some $\alpha > 0$. Then there exist positive constants c_1, c_2 such that if $\mathrm{pen} : \mathbb{R}^+ \to \mathbb{R}^+$ satisfies

$$\forall \lambda > 0, \quad \mathrm{pen}(\lambda) \geq c_1 R(\lambda, n) + \frac{c_2 b(\lambda)(\alpha \log(\lambda) + \xi + \log(2))}{n}$$

for some positive number ξ, then, with probability at least $1 - \exp(-\xi)$, the penalized estimator $\hat{f}_n$ defined by

$$\hat{f}_n = \arg\min_{k \geq 1} \{A_n(\hat{f}_n^{\lambda_k}) + \mathrm{pen}(\lambda_k)\}$$

satisfies

$$A(\hat{f}_n) - A(f^*) \leq \inf_{k \geq 1} \left\{ 2 \inf_{f \in \mathcal{F}_{\lambda_k}} (A(f) - A(f^*)) + \mathrm{pen}(\lambda_k) \right\} .$$

For a fixed model, the same rate was achieved in the case of a single model in Bartlett, Jordan, and McAuliffe [27]. The general model selection theorems by Massart have also led to oracle inequalities for penalized procedures inspired by Support Vector Machines (see Blanchard, Bousquet, and Massart [32]). This type of result is now better understood thanks to the impressive paper by Koltchinskii [74] (see also [34] for a detailed account on model selection techniques in classification).

In the case when the distribution of the (X, Y) happens to be such that the approximation error $\inf_{f \in \mathcal{F}_{\lambda_k}} A(f) - A^*$ vanishes for some value of λ, the above theorem implies the following immediate corollary for the rate of convergence of $A(\hat{f}_n)$ to A^*.

Corollary 16 (Blanchard, Lugosi, and Vayatis [3]) *Assume that the distribution of (X, Y) is such that there exists a $\lambda_0 > 0$ such that $\inf_{f \in \mathcal{F}_{\lambda_0}} A(f) = A(f^*)$. Under the conditions of Theorem 15, if the penalty is chosen to be*

$$\mathrm{pen}(\lambda) = c_1 R(\lambda, n) + \frac{c_2 b(\lambda)(\alpha \log(\lambda) + 2 \log n + \log 2)}{n}$$

then for every n, with probability at least $1 - 1/n^2$,

$$A(\hat{f}_n) - A(f^*) \leq C n^{-\frac{1}{2}\left(\frac{V+2}{V+1}\right)}$$

where the constant C depends on the distribution, on the class $\mathcal{F}$, and on the cost function φ.

Note that the penalty function does *not* depend on λ_0 above, so that the procedure is truly adaptive. We can then apply Zhang's lemma to derive a bound on the excess risk in terms of classification error $L(\hat{f}_n) - L^*$.

We now turn to improvements beyond the use of Zhang's inequality which can be obtained under additional assumptions. Indeed, Mammen and Tsybakov [81, 108] pointed out that under certain *low-noise assumptions* (also known as *margin conditions*) on the distribution much faster rates of convergence for the ERM principle may be achieved. The original form of these assumptions is the following: for some $\alpha \in [0, 1]$, there exists a constant $B > 0$ such that for any $t \geq 0$, we have

$$\mathbb{P}\{|2\eta(X) - 1| \leq t\} \leq B\, t^{\frac{\alpha}{1-\alpha}} \ .$$

An equivalent form, with the same α, is the following: there exists a constant $\beta > 0$ such that for any real-valued measurable function f,

$$\mathbb{P}\{g_f(X) \neq g^*(X)\} \leq \beta\, (L(f) - L^*)^{\alpha} \ . \tag{2.1}$$

Notice that all distributions satisfy this condition with $\alpha = 0$ and $\beta = 1$, while larger values of α place more restriction on the distribution. Intuitively, a large value of α means that the probability that $\eta(X)$ is close to $1/2$ is small. In the extreme case of $\alpha = 1$ it is easy to see that $\eta(X)$ stays bounded away from $1/2$ with probability one. For the treatment of the extremely-low-noise or zero-noise cases, we refer to the work of Massart and Nédélec [89].

We make use of the next lemma which uses the the margin condition and is adapted to the convex risk minimization setup .

Lemma 17 (BARTLETT, JORDAN, AND MCAULIFFE [27]) *Let φ be a cost function satisfying the conditions of Lemma 7 and assume that condition (2.1) holds for some $\alpha \in [0,1]$ and $\beta > 0$. Then*

$$L(f) - L^* \leq \left(\frac{2^s c}{\beta^{1-s}} (A(f) - A(f^*)) \right)^{1/(s - s\alpha + \alpha)} .$$

For the exponential and the logit cost functions, we have $s = 2$ and in that case, as α moves from zero to one, the exponent $1/(s - s\alpha + \alpha)$ changes from $1/2$ to 1. Thus, large values of α significantly improve the rates of convergence of $L(f)$ to L^*. We formulate a theorem summarizing our findings in terms of excess risk for the classification error.

Corollary 18 (BLANCHARD, LUGOSI, AND VAYATIS [3]) *Let φ be either the exponential or the logit cost function and consider the penalized estimate $\hat{f}_n$ of Corollary 16. Assume that the distribution of (X,Y) is such that there exists a $\lambda > 0$ such that $\inf_{f \in \mathcal{F}_\lambda} A(f) = A(f^*)$. Then for every n, with probability at least $1 - 1/n^2$, the probability of error $L(\hat{f}_n)$ of the associated classifier satisfies*

$$L(\hat{f}_n) - L^* \leq C n^{-\frac{1}{4}\left(\frac{V+2}{V+1}\right)}$$

where the constant C depends on the distribution, on the class $\mathcal{F}$, and on the cost function φ. Also, with probability one,

$$\lim_{n \to \infty} \left(L(\hat{f}_n) - L^* \right) n^{\frac{1}{4}\left(\frac{V+2}{V+1}\right)} = 0 .$$

If, in addition, condition (2.1) holds for some $\alpha \in [0,1]$ and $\beta > 0$, then with probability at least $1 - 1/n^2$,

$$L(\hat{f}_n) - L^* \leq C n^{-\frac{1}{2(2-\alpha)}\left(\frac{V+2}{V+1}\right)} .$$

The remarkable fact about this corollary is that the obtained rate of convergence is independent of the dimension of the space in which the observations take their values. The rates depend on the VC dimension of the base class which may be related to the dimension of the input space. However, this dependence is mild and even if V is very large, the rates are always faster than $n^{-1/(2(2-\alpha))}$. The dependence on the dimension is mostly reflected in the value of the constant C. Recall from Theorem 15 that the value of C is determined by the smallest value of λ for which $\inf_{f \in \mathcal{F}_\lambda} A(f) = A^*$ and its dependence on λ is determined by the cost function φ. For complex distributions, high-dimensional input spaces, and simple base classes, this constant will be very large. The main message of Corollary 18 is that, as a function of the sample size n, the probability of error converges at a fast rate, independently of the dimension. To understand the meaning of this result, we need to study the main condition on the distribution, that is, that the minimizer f^* of the expected cost falls in the closure of $\mathcal{F}_\lambda$ for some finite value of λ.

2.2 Approximation Properties of Boosting Decision Rules

An important feature related to the inputs of a boosting method is the choice of the base class. The underlying trade-off is not fully understood despite the growing activity from

the field of Approximation Theory in connection with Machine Learning (a special issue of Constructive Approximation on Mathematical Learning Theory is scheduled for spring 2007, we also refer to the recent work of De Vore, Kerkyacharian, Picard, and Temlyakov [51], and Cohen, Dahmen, and De Vore [41]). There are two levels in this trade-off:

- theoretical level: classical bias-variance dilemma. The difficulty comes here from the fact that standard complexity measures involved in the estimation part, like VC dimension, do not account for the approximation capacity of the method (see the note by Koltchinskii, Lugosi, and Mendelson [75]).
- practical level: computational constraints vs. representation capacity. Indeed, the algorithmic complexity of boosting algorithm is linear in the complexity of extracting a single weak classifier. There is a hidden Empirical Risk Minimization step in each boosting method and solving it has to be kept simple. That explains that practitioners prefer to use simple base classes such as decision stumps or short decision trees. The question is which distributions can efficiently be learned with these simple classes.

In the sequel, we provide some qualitative results to describe consistent base classes and explore the representation capacity of boosting decision stumps.

2.2.1 Examples of consistent base classes

The question of representation capacity for boosting (or kernel)-type methods is whether the set $C^* = \{x \in \mathcal{X} \ : \ \eta(x) > 1/2\}$ can arbitrarily be approached by sets of the form $C_f = \{x \in \mathcal{X} \ : \ f(x) > 0\}$ where possible f's belong to an increasing family of balls in a Hilbert or Banach space of functions.

The first simple result leads to universally consistent boosting methods. We recall that the assumption on the class $\mathcal{G}$ is given by $\lim_{\lambda\to\infty} \inf_{f\in\mathcal{F}_\lambda} A(f) = A^*$ (see also the argument of Breiman in [37] using a simpler completeness condition).

Lemma 19 (Lugosi and Vayatis [12]) *Let the class $\mathcal{G}$ be such that its convex hull contains all the indicators of elements of $\mathcal{B}_0$, a subalgebra of the Borel σ-algebra of $\mathbb{R}^d$, denoted by $\mathcal{B}(\mathbb{R}^d)$, with $\mathcal{B}_0$ generating $\mathcal{B}(\mathbb{R}^d)$. Then,*

$$\lim_{\lambda\to\infty} \inf_{f\in\mathcal{F}_\lambda} A(f) = A^* \, ,$$

where $A^* = \inf A(f)$ over all measurable functions $f : \mathbb{R}^d \longrightarrow \mathbb{R}$.

More generally, a straightforward modification of this Lemma shows that whenever $\mathcal{F} = \bigcup_{\lambda>0} \mathcal{F}_\lambda$ is dense in $L_1(\mu)$ with μ being the marginal distribution of the random variable X, then it is true that $\inf_{f\in\mathcal{F}} A(f) = A(f^*)$.

A few simple choices of base classes $\mathcal{G}$ over $\mathcal{X} = \mathbb{R}^d$ satisfying this richness property are: the class of all linear classifiers (that is, functions of the form $g(x) = 2\mathbb{I}_{[a\cdot x\le b]} - 1$, $a \in \mathbb{R}^d$, $b \in \mathbb{R}$), the class of all closed hyperrectangles, the class of all closed balls and

their complements, the class of binary decision tree classifiers using axis parallel cuts with $d+1$ terminal nodes.

Clearly, the list of possibilities is endless, and these few examples are just some of the most natural choices. All five examples are such that $\bigcup_{\lambda>0}\mathcal{F}_\lambda$ is dense in $L_1(\mu)$ for any probability distribution μ. Indeed, in the cases of hyper-rectangles and balls, this statement is obvious. For the class of linear classifiers, this follows from denseness results of neural networks, see [47], [68]. For the case of trees, see [37]. We also refer to the general statement given as a universal approximation theorem by [119] and which shows that, for the classical choices of the cost function φ, we have, for any distribution, $\inf_{f\in\bigcup_{\lambda>0}\mathcal{F}_\lambda} A(f) = A^*$ as soon as $\bigcup_{\lambda>0}\mathcal{F}_\lambda$ is dense in the space of continuous functions under the supremum norm.

We point out that the rates of convergence established in [3] depend primarily on the VC dimension of the base class. The VC dimension equals $V = d+1$ in the case of linear classifiers, $V = 2d+1$ for the class of hyperrectangles, $V = d+2$ for the class of balls, and by $V = d\log_2(2d)$ for the class of binary decision trees [52]. Clearly, the lower the VC dimension is, the faster the rate (estimation is easier). In order to account for the difficulty of controlling the approximation error term, we refer to the surprising note by Koltchinskii, Lugosi and Mendelson [75] which establishes the existence of such a rich class with VC dimension equal to one.

Now the most interesting problem is to determine the class of distributions for which we have $\inf_{f\in\mathcal{F}_\lambda} A(f) = A^*$ for some finite value of λ. In all the above-mentioned special cases this class is quite large, giving rise to a remarkably rich class of distributions for which dimension-independent rates of convergence hold. The characterization of these classes of distributions is far from being well understood. In the case of linear classes the problem is closely related to the approximation properties of neural networks. We merely refer to [22], [23], [48], [62], [80], [90], [94], [102] for related results. Most of these references provide quantitative results relating the approximation error to the smoothness of the target function. However, there are very few attempts to characterize the functions that can actually be reconstructed with given dictionaries.

2.2.2 Boosting using decision stumps

We now turn to a special case for which the approximation properties will be fully described. We will also provide the particular rates of convergence achieved by boosting methods. In what follows, we consider the base class $\mathcal{G}$ of decision stumps. We recall that a decision stump is a linear classifier whose cuts are parallel to the axes.

One-dimensional case

We first consider the simple one-dimensional case when $\mathcal{X} = [0,1]$ and when the base class contains all classifiers g of the form $g(x) = \mathbb{I}_{[x\geq t]} - \mathbb{I}_{[x<t]}$ and of the form $g(x) = \mathbb{I}_{[x<t]} - \mathbb{I}_{[x\geq t]}$ where $t \in [0,1]$ can take any value. Clearly, the VC dimension of $\mathcal{G}$ is $V = 2$. We describe the class of distributions such that there exists a λ such that $\inf_{f\in\mathcal{F}_\lambda} A(f) = A^*$. The next lemma states a simple sufficient condition.

Lemma 20 (Blanchard, Lugosi, and Vayatis [3]) *Assume that the cost function and the distribution of* (X, Y) *are such that the function* f^* *is of bounded variation. If* $|\cdot|_{BV}$ *denotes the total variation norm, define* $|f^*|_{BV,0,1} = \frac{1}{2}(f^*(0) + f^*(1) + |f^*|_{BV})$. *Then* $\inf_{f \in \mathcal{F}_\lambda} A(f) = A(f^*)$ *whenever* $\lambda \geq |f^*|_{BV,0,1}$.

Thus, the fast rates of convergence can be guaranteed whenever f^* is everywhere finite and has a bounded variation. Recall that for the exponential cost function $f^* = (1/2)\log(\eta/(1-\eta))$ and for the logit cost function $f^* = \log(\eta/(1-\eta))$. In both cases, it is easy to see that f^* has a bounded variation if and only if η is bounded away from zero and one and has a bounded variation.

The situation when η is bounded away from zero and one may seem to be quite unnatural at first sight. Indeed, values of η close to zero and one mean that the distribution has little noise and should make the classification problem easier. However, regularized boosting methods suffer when faced with a low-noise distribution since very large values of λ are required to drive the approximation error $\inf_{f \in \mathcal{F}_\lambda} A(f) - A^*$ close to zero.

Corrupting the data

If η can be arbitrarily close to 0 and 1, then f^* takes arbitrarily large positive or negative values and thus cannot be in any $\mathcal{F}_\lambda$ (since functions in this set take values in $[-\lambda, \lambda]$). In order to avoid such a behavior, one may artificially add some random noise to the data. Indeed, if, for example, we define the random variable Y' such that it equals Y with probability $3/4$ and $-Y$ with probability $1/4$, then the function $\eta'(x) = \mathbb{P}[Y' = 1|X = x] = 1/4 + \eta(x)/2$ takes its values in the interval $[1/4, 3/4]$ (a similar transformation was also proposed by Yang [117], [118]). More importantly, the Bayes classifier g' for the distribution (X, Y') coincides with the Bayes classifier g^* of the original problem. We denote the probability of error of g under the distribution of (X, Y') by $L'(g)$ and the corresponding Bayes error by L'^*. If we recall from [52] that for any classifier g,

$$L(g) - L^* = \mathbb{E}\big(|2\eta(X) - 1|\, \mathbb{I}_{[g(X) \neq g^*(X)]}\big)\,,$$

then we see that for any classifier g,

$$L(g) - L^* = 2(L'(g) - L'^*)\,. \tag{2.2}$$

This means that if one can design a classifier which performs well for the "noisy" problem (X, Y'), then the same classifier will also work well for the original problem (X, Y). Thus, in order to enlarge the class of distributions for which the fast rates of convergence holds, one may artificially corrupt the data by a random noise, replacing each label Y_i by a noisy version Y_i' as described above. Then the distribution of the noisy data is such that $\eta'(x)$ is bounded away from zero.

Of course, by corrupting the data deliberately with noise, one looses information, but it is a curious property of the regularized boosting methods studied here that the rate of convergence may be speeded up considerably for some distributions. Indeed, this fact was already pointed out by Yang in establishing general minimax rates of convergence in various settings (see [117], [118]). We recover here the optimal minimax rate from [117] with a different method.

Corollary 21 (Blanchard, Lugosi, and Vayatis [3]) *Let $X \in [0,1]$. Let φ be either the exponential or the logit cost function and consider the penalized estimate $\widehat{f}_n$ based on decision stumps, calculated based on the noise-corrupted data set described above. If $\eta(x)$ has a bounded variation, then for every n, with probability at least $1 - 1/n^2$, the probability of error $L(\widehat{f}_n)$ of the associated classifier satisfies*

$$L(\widehat{f}_n) - L^* \leq Cn^{-\frac{1}{3}}$$

where the constant C depends only on $|\eta|_{BV}$. If, in addition, condition (2.1) holds for some $\alpha \in [0,1]$ and $\beta > 0$, then

$$L(\widehat{f}_n) - L^* \leq Cn^{-\frac{2}{3(2-\alpha)}} \ .$$

High-dimensional case

Further, we investigate the case when $\mathcal{X} = [0,1]^d$ and the base class $\mathcal{G}$ contains all "decision stumps", that is, all classifiers of the form $g^+_{i,t}(x) = \mathbb{I}_{[x^{(i)} \geq t]} - \mathbb{I}_{[x^{(i)} < t]}$ and $g^-_{i,t}(x) = \mathbb{I}_{[x^{(i)} < t]} - \mathbb{I}_{[x^{(i)} \geq t]}$, $t \in [0,1]$, $i = 1, \ldots, d$, where $x^{(i)}$ denotes the i-th coordinate of x.

It is easy to see, as in Lemma 20, that the closure of $\mathcal{F}_\lambda$ with respect to the supremum norm contains all functions f of the form

$$f(x) = f_1(x^{(1)}) + \cdots + f_d(x^{(d)})$$

where the functions $f_i : [0,1] \to \mathbb{R}$ are such that $|f_1|_{BV,0,1} + \cdots + |f_d|_{BV,0,1} \leq \lambda$. Therefore, if f^* has the above form, we have $\inf_{f \in \mathcal{F}_\lambda} A(f) = A(f^*)$.

Recalling the form of the function f^* optimizing the cost in the case of the exponential or logit cost function, we can understand that boosting using decision stumps is especially well fitted to the so-called additive logistic model in which η is assumed to be such that $\log(\eta/(1-\eta))$ is an additive function (i.e., it can be written as a sum of univariate functions of the components of x), see Hastie and Tibshirani [66]. The fact that boosting is intimately connected with additive logistic models of classification has already been pointed out by Friedman, Hastie, and Tibshirani [60]. The next result shows that indeed, when η permits an additive logistic representation then the rate of convergence of the regularized boosting classifier is fast and has a very mild dependence on the distribution.

Corollary 22 (Blanchard, Lugosi, and Vayatis [3]) *Let $X \in [0,1]^d$ with $d \geq 2$. Let φ be either the exponential or the logit cost function and consider the penalized estimate $\widehat{f}_n$ of Corollary 16 based on decision stumps. Let $V_2 = 3$, $V_3 = 4, V_4 = 5$, and for $d \geq 5$, $V_d = \lfloor 2\log_2(2d) \rfloor$. If there exist functions $f_1, \ldots, f_n : [0,1] \to \mathbb{R}$ of bounded variation such that $\log \frac{\eta(x)}{1-\eta(x)} = \sum_{i=1}^d f_i(x^{(i)})$ then for every n, with probability at least $1 - 1/n^2$, the probability of error $L(\widehat{f}_n)$ of the associated classifier satisfies*

$$L(\widehat{f}_n) - L^* \leq Cn^{-\frac{1}{4}\left(\frac{V_d+2}{V_d+1}\right)}$$

where the constant C *depends on* $\sum_{i=1}^{d} |f_i|_{BV,0,1}$. *If, in addition, condition (2.1) holds for some* $\alpha \in [0,1]$ *and* $\beta > 0$, *then*

$$L(\hat{f}_n) - L^* \leq C n^{-\frac{1}{2(2-\alpha)}\left(\frac{V_d+2}{V_d+1}\right)} .$$

Under the assumption of the additive logistic model, the rate of convergence is of the order of $n^{-\frac{1}{2(2-\alpha)}\left(\frac{V_d+2}{V_d+1}\right)}$ where V_d depends on d in a logarithmic fashion. Even for large values of d, the rate is always faster than $n^{-1/2(2-\alpha)}$. It is also useful to examine the dependence of the constant C on the dimension. A quick look at Theorem 15 reveals that C in the first inequality of Corollary 22 may be bounded by a universal constant times $\sqrt{V_d \varphi(\lambda)^{1/V_d} \lambda \varphi'(\lambda)}$ where λ is the smallest number such that $\inf_{f \in \mathcal{F}_\lambda} A(f) = A^*$. Thus, we may take $\lambda = \sum_{i=1}^{d} |f_i|_{BV,0,1}$. Since $V_d = \lfloor 2\log_2(2d) \rfloor$, the dependence on the dimension is primarily determined by the growth of the cost function φ. Here there is a significant difference between the behavior of the exponential and the logistic cost functions in high dimensions. For the purpose of comparison, it is reasonable to consider distributions such that $\lambda = \sum_{i=1}^{d} |f_i|_{BV,0,1}$ is bounded by a linear function of d. In that case the constant C depends on d as $O(\sqrt{de^d \log d})$ in the case of the exponential cost function, but only as $O(\sqrt{d \log d})$ in the case of the logistic cost function (using directly Theorem 15 instead of the upper bound mentioned above). In summary, regularized boosting using the logit cost function and decision stumps has a remarkably good behavior under the additive logistic model in high dimensional problems, as stated in the next corollary.

Corollary 23 (Blanchard, Lugosi, and Vayatis [3]) *Let* $X \in [0,1]^d$ *with* $d \geq 2$. *Let* φ *be the logit cost function and consider the penalized estimate* $\hat{f}_n$ *of Corollary 16 based on decision stumps. Let* B *be a positive constant. If there exist functions* $f_1, \ldots, f_n : [0,1] \to \mathbb{R}$ *with* $\lambda = \sum_{i=1}^{d} |f_i|_{BV,0,1} \leq Bd$ *such that* $\log \frac{\eta(x)}{1-\eta(x)} = \sum_{i=1}^{d} f_i(x^{(i)})$ *then for every* n, *with probability at least* $1 - 1/n^2$, *the probability of error* $L(\hat{f}_n)$ *of the associated classifier satisfies*

$$L(\hat{f}_n) - L^* \leq C\sqrt{d \log d}\; n^{-\frac{1}{4}\left(\frac{V_d+2}{V_d+1}\right)}$$

where C *is a universal constant and* V_d *is as in Corollary 22. If, in addition, condition (2.1) holds for some* $\alpha \in [0,1]$ *and* $\beta > 0$, *then*

$$L(\hat{f}_n) - L^* \leq C(d \log d)^{\frac{1}{2-\alpha}}\; n^{-\frac{1}{2(2-\alpha)}\left(\frac{V_d+2}{V_d+1}\right)} .$$

Remark 1 Just like in the one-dimensional case, the conditions of Corollary 22 require that η be bounded away from zero and one. To relax this assumption, one may try to add a random noise to the data, just like in the one-dimensional case. However, this may not work in the higher-dimensional problem because even if f^* is an additive function, it may not have this property any longer after the noise is added. ▮

In [3], we provide further results on approximation properties of sets C_f obtained by taking the sign of linear combinations of decision stumps. However, for boosting to be successful it is not enough that the Bayes classifier g^* can be written in such a form. It may happen that even though g^* is in the class of classifiers induced by functions in $\mathcal{F}_\lambda$, the classifier corresponding to $\bar{f}_\lambda$ minimizing the cost $A(f)$ in $\mathcal{F}_\lambda$ is very different. We also propose an example in which *for any* $\lambda > 0$ there exists an $f \in \mathcal{F}_\lambda$ such that $g_f = g^*$.

Chapter 3

The Ranking Problem

Our approach of the ranking problem was mainly motivated by two applications: information retrieval and credit risk screening. In both cases, the problem is to provide a ranked list of a set of instances instead of simply classifying them. The idea is that the most relevant (or the most reliable) instance should arrive at the top of the list. The design and the theoretical analysis of ranking methods is considered today as a burning issue in Machine Learning ([42], [67], [57]). We have used the experience acquired in the classification setup in order to schedule an in-depth investigation of the statistical aspects of this ranking problem. From the viewpoint of statistics, rank data and rank statistics have been thoroughly studied (see e.g. [83], [64]) however the learning view (closer in spirit to M-estimation) requires different concepts and tools.

In our approach, the important feature of the ranking problem is that natural estimates of the ranking risk involve U-statistics. Therefore, the methodology is based on the theory of U-processes, and the key tools involve maximal and concentration inequalities, symmetrization tricks, and a contraction principle for U-processes. In this chapter, we formulate the problem and provide consistency results for certain nonparametric ranking methods. We also formulate a novel tail inequality for degenerate U-processes and, based on the latter result, show that fast rates of convergence may be achieved for empirical risk minimizers under suitable noise conditions.

We also point out that under certain conditions, finding a good ranking rule amounts to constructing a scoring function s. An important special case is the bipartite ranking problem in which the available instances are labelled by binary labels ("good" or "bad"). In this case the ranking criterion is closely related to the so-called AUC criterion. In the last part of the chapter, we investigate the question of localizing the ranking problem in order to focus on the best instances.

3.1 Ranking as classification of pairs of observations

Let (X, Y) be a pair of random variables taking values in $\mathcal{X} \times \mathbb{R}$ where $\mathcal{X}$ is a measurable space. The random object X models some observation and Y its real-valued label.

The purpose of ranking is to rank the instances in $\mathcal{X}$ on the basis of the information provided by their label. A natural way to do this is to find a *scoring function* $s : \mathcal{X} \to \mathbb{R}$

which will induce an order on $\mathcal{X}$ and reflect the order of the corresponding labels. We consider three examples where we describe the optimal scoring function.

Example 3 - Noise-free regression model. In this model $Y = m(X)$ for some (unknown) function $m : \mathcal{X} \to \mathbb{R}$. Here obviously there is a perfect ranking and the optimal scoring function is $s^* = m$, or any strictly increasing transformation of it. ▮

Example 4 - Regression model with noise. Now we turn to the general regression model with heteroscedastic errors in which $Y = m(X) + \sigma(X)\epsilon$ for some (unknown) functions $m : \mathcal{X} \to \mathbb{R}$ and $\sigma : \mathcal{X} \to \mathbb{R}$, where ϵ is a symmetric random variable, independent of X. We have again $s^* = m$, or any strictly increasing transformation of it. ▮

Example 5 - Classification model. Consider here the standard binary classification model. In this case, the situation is ambiguous since it is not clear why ranking differs from classifying: how to order a set of binary labels? However, this is a common situation in applications such as credit scoring. Financial institutions classify their clients in two categories - "good" or "bad" - but they actually need to rank them according to their socio-economic attributes (contained in the X vector) by the means of a scoring function. Since they have a limited amount of money, they want to find the debtors who are *the most likely* to pay their loan back. Ideally, they would like to recover the function $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$, $\forall x \in \mathcal{X}$, or any strictly increasing transformation of it. ▮

The scoring problem considered in the previous examples can be stated as follows: find a scoring function $\hat{s}_n$ on the basis of empirical data $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ which will approximate the ranking on the instances of $\mathcal{X}$ achieved by the regression function $\mathbb{E}(Y \mid X = x)$. Hence, the solution to this problem is the equivalence class of increasing transforms of the regression function. We point out that this problem should be easier than regression so we seek an appropriate strategy devoted to the ranking/scoring problem.

A simple idea in order to assess the performance of a scoring function s in ranking the instances is to count how many pairs of data have their order inverted by s compared to their respective labels. Let (X', Y') denote a pair of random variables identically distributed with (X, Y), and independent of it. We think about X being "better" than X' if $Y > Y'$. Hence s will commit an error every time we have a pair $\big((X, Y), (X', Y')\big)$ such that the signs of $Y - Y'$ and $s(X) - s(X')$ are different. We can then introduce the *ranking risk* of a scoring function:

$$L(s) = \mathbb{P}\{(Y - Y')(s(X) - s(X')) < 0\} ,$$

which is to be minimized over a given class of scoring functions.

We point out that this criterion suffers from a serious drawback since it weights all pairs uniformly independently of how bad the inversion in the ranking is (inverting, say, the third instance with the fifth will imply the same loss as inverting the first instance with the last one). In spite of this observation, we have chosen to consider the ranking risk above in order to initiate our investigations on this topic and we will see it presents

some interesting features. It is also worth mentioning that, under the classification model, this criterion can be related to standard performance measures used in applications such as the Receiving Operator Characteristic (ROC) ([111], [55]) and the Area Under an ROC Curve (AUC) criterion ([65], [44]).

In order to provide a formal study, we adopt a more general setup. Since the goal is to rank X and X' such that the probability that the better ranked of them has a smaller label is as small as possible, we consider a function $r : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$ which will be called a *ranking rule*. If $r(x, x') = 1$ then the rule ranks X higher than X'.Denote by $Z = \frac{Y-Y'}{2}$ the ranking label of the pair of observations (X, X'). The performance of a ranking rule is measured by the *ranking risk*

$$L(r) = \mathbb{P}\{Z \cdot r(X, X') < 0\} ,$$

that is, the probability that r ranks two randomly drawn instances incorrectly. In this formulation, the ranking problem is equivalent to a binary classification problem in which the sign of the random variable Z is to be guessed based upon the pair of observations (X, X'). Introduce the notation

$$\rho_+(X, X') = \mathbb{P}\{Z > 0 \mid X, X'\}$$
$$\rho_-(X, X') = \mathbb{P}\{Z < 0 \mid X, X'\}.$$

Now it is easy to derive the ranking rule r^* with minimal risk over all possible ranking rules.

Proposition 24 (Clémençon, Lugosi, and Vayatis [7]) *Define*

$$r^*(x, x') = 2\mathbb{I}_{[\rho_+(x,x') \geq \rho_-(x,x')]} - 1$$

and denote $L^* = L(r^*) = \mathbb{E}\{\min(\rho_+(X, X'), \rho_-(X, X'))\}$. *Then for any ranking rule* r,

$$L^* \leq L(r) .$$

In the previous examples, the ranking problem formulated here may be reduced to finding an appropriate *scoring function*. These are the cases when the joint distribution of X and Y is such that there exists a function $s^* : \mathcal{X} \to \mathbb{R}$ such that

$$r^*(x, x') = 1 \quad \text{if and only if} \quad s^*(x) \geq s^*(x') .$$

A function s^* satisfying the assumption is called an *optimal scoring function*. Obviously, any strictly increasing transformation of an optimal scoring function is also an optimal scoring function. Below we describe the important special case of the bipartite ranking problem which has been considered in the machine learning literature ([57], [18]).

Example 6 (the bipartite ranking problem.) In the bipartite ranking problem, the label Y is binary and takes values in $\{-1, 1\}$. Writing $\eta(x) = \mathbb{P}\{Y = 1 | X = x\}$, it is not hard to see that the Bayes ranking risk equals

$$L^* = \mathrm{Var}\left(\frac{Y+1}{2}\right) - \frac{1}{2}\mathbb{E}\left|\eta(X) - \eta(X')\right| .$$

In particular,

$$L^* \leq \mathrm{Var}\left(\frac{Y+1}{2}\right) \leq 1/4$$

where the equality $L^* = \mathrm{Var}\left(\frac{Y+1}{2}\right)$ holds when X and Y are independent and the maximum is attained when $\eta \equiv 1/2$. Observe that the difficulty of the bipartite ranking problem depends on the concentration properties of the distribution of $\eta(X) = \mathbb{P}(Y = 1 \mid X)$ through the quantity $\mathbb{E}(|\eta(X) - \eta(X')|)$ which is a classical measure of concentration, known as *Gini's mean difference*. It is clear from the form of the Bayes ranking rule that the optimal ranking rule is given by a scoring function s^* where s^* is any strictly increasing transformation of η. Then one may restrict the search to ranking rules defined by scoring functions s, that is, ranking rules of form $r(x, x') = 2\mathbb{I}_{[s(x) \geq s(x')]} - 1$. Writing $L(s) \stackrel{\text{def}}{=} L(r)$, one has

$$L(s) - L^* = \mathbb{E}\left(|\eta(X') - \eta(X)|\, \mathbb{I}_{[(s(X)-s(X'))(\eta(X)-\eta(X'))<0]}\right) . \quad \blacksquare$$

Just like in the standard setting of binary classification of single observations, this optimal rule r^* cannot be known unless the underlying distribution is specified. However a reasonable goal could be to investigate the construction of ranking rules of low risk based on training data. We assume that n independent, identically distributed copies of (X, Y), are available: $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Given a ranking rule r, one may use the training data to estimate its risk $L(r) = \mathbb{P}\{Z \cdot r(X, X') < 0\}$. The perhaps most natural estimate is the *U-statistic*

$$L_n(r) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}_{[Z_{i,j} \cdot r(X_i, X_j) < 0]}.$$

In the sequel, we consider minimizers of the empirical estimate $L_n(r)$ over a class $\mathcal{R}$ of ranking rules and study the performance of such empirically selected ranking rules.

3.2 Representations of U-statistics

Here we recall some basic facts about U-statistics. Consider the i.i.d. random variables $X, X_1, ..., X_n$ and denote by

$$U_n = \frac{1}{n(n-1)} \sum_{i \neq j} q(X_i, X_j)$$

a U-statistic of order 2 where the kernel q is a symmetric real-valued function. The U-statistic U_n is said *degenerate* if its kernel q satisfies $\mathbb{E}\left(q(x, X)\right) = 0$, $\forall x$.

There are two basic representations of U-statistics which we recall next (see Serfling [101] for more details).

Average of 'sums-of-i.i.d.' blocks

This representation is the key for obtaining 'first-order' results for non-degenerate U-statistics. The U-statistic U_n can be expressed as

$$U_n = \frac{1}{n!} \sum_{\pi} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} q\left(X_{\pi(i)}, X_{\pi(\lfloor n/2 \rfloor + i)}\right)$$

where the sum is taken over all permutations π of $\{1, \ldots, n\}$. The idea underlying this representation is to reduce the analysis to the case of sums of i.i.d. random variables.

Hoeffding's decomposition

Another way to interpret a U-statistics is as an orthogonal expansion known as Hoeffding's decomposition.

Assuming that $q(X_1, X_2)$ is square integrable, $U_n - \mathbb{E}U_n$ may be decomposed as a sum T_n of i.i.d. random variables plus a *degenerate* U-statistic W_n. In order to write this decomposition, consider the following function of one variable

$$h(X_i) = \mathbb{E}(q(X_i, X) \mid X_i) - \mathbb{E}U_n \,,$$

and the function of two variables

$$\widehat{h}(X_i, X_j) = q(X_i, X_j) - \mathbb{E}U_n - h(X_i) - h(X_j).$$

Then we have the orthogonal expansion

$$U_n = \mathbb{E}U_n + 2T_n + W_n \,,$$

where

$$T_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i),$$
$$W_n = \frac{1}{n(n-1)} \sum_{i \neq j} \widehat{h}(X_i, X_j) \,.$$

The statistic W_n is a degenerate U-statistic. Its variance is of the order $1/n^2$. Thus, T_n is the leading term in this orthogonal decomposition.

3.3 First-order analysis

3.3.1 Empirical risk minimization

Based on the empirical estimate $L_n(r)$ of the risk $L(r)$ of a ranking rule defined above, one may consider choosing a ranking rule by minimizing the empirical risk over a class $\mathcal{R}$ of ranking rules $r : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$. Define the empirical risk minimizer, over $\mathcal{R}$, by

$$\hat{r}_n = \arg\min_{r \in \mathcal{R}} L_n(r) \,.$$

In a "first-order" approach, we may study the performance $L(\hat{r}_n) = \mathbb{P}\{Z \cdot \hat{r}_n(X, X') < 0 \mid D_n\}$ of the empirical risk minimizer by the standard bound (see, e.g., [52])

$$L(\hat{r}_n) - \inf_{r \in \mathcal{R}} L(r) \leq 2 \sup_{r \in \mathcal{R}} |L_n(r) - L(r)| \,. \tag{3.1}$$

This inequality points out that bounding the performance of an empirical minimizer of the ranking risk boils down to investigating the properties of U-*processes*, that is, suprema

of U-statistics indexed by a class of ranking rules. For a detailed and modern account of U-process theory we refer to the book of de la Peña and Giné [49]. In a first-order approach we basically reduce the problem to the study of ordinary empirical processes. Indeed, using the first representation of a U-statistic, Chernoff's bounding method, and the concentration property of Rademacher averages, one can derive the following result.

Proposition 25 (Clémençon, Lugosi, and Vayatis [7]) *Let $\delta > 0$. With probability at least $1 - \delta$,*

$$L(\hat{r}_n) - \inf_{r \in \mathcal{R}} L(r) \leq 4\mathbb{E}\hat{r}_n + 4\sqrt{\frac{\ln(1/\delta)}{n-1}} \ .$$

where

$$\hat{r}_n = \sup_{r \in \mathcal{R}} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \epsilon_i \mathbb{I}_{[Z_{i,\lfloor n/2 \rfloor + i} r(X_i, X_{\lfloor n/2 \rfloor + i}) < 0]} \right|$$

where $\epsilon_1, ..., \epsilon_n$ are i.i.d. Rademacher random variables (i.e., random symmetric sign variables).

The expected value of the Rademacher average $\hat{r}_n$ may now be bounded by standard methods, see, e.g., Lugosi [79], Boucheron, Bousquet, and Lugosi [34]. For example, if the class $\mathcal{R}$ of indicator functions has finite VC dimension V, then

$$\mathbb{E}\hat{r}_n \leq c\sqrt{\frac{V}{n}}$$

for a universal constant c.

The proposition above is, in a certain sense, not improvable. However, it is well known from the theory of statistical learning and empirical risk minimization for classification that the bound (3.1) is often quite loose. In classification problems the looseness of such a "first-order" approach is due to the fact that the variance of the estimators of the risk is ignored and bounded uniformly by a constant. Therefore, the main interest in considering U-statistics precisely consists in the fact that they have minimal variance among all unbiased estimators. We will take advantage of the reduced-variance property of U-statistics for the ranking problem in our analysis of fast rates of convergence (Section 3.4).

Observe that the bound of Proposition 25 remains true for an empirical risk minimizer that, instead of using estimates based on U-statistics, estimates the risk of a ranking rule by splitting the data set into two halves and estimates $L(r)$ by

$$L'_n(r) = \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{I}_{[Z_{i,\lfloor n/2 \rfloor + i} \cdot r(X_i, X_{\lfloor n/2 \rfloor + i}) < 0]} \ .$$

Hence, in the previous study one loses the advantage of using U-statistics. In Section 3.4 it is shown that under certain, not uncommon, circumstances, significantly smaller risk bounds are achievable. There it will have an essential importance to use sharp exponential bounds for U-processes, involving their reduced variance.

3.3.2 Convex risk minimization

Several successful algorithms for classification, including various versions of *boosting* and *support vector machines* are based on replacing the loss function by a convex function and minimizing the corresponding empirical convex risk functionals over a certain class of functions (typically over a ball in an appropriately chosen Hilbert or Banach space of functions). This approach has important computational advantages, as the minimization of the empirical convex functional is often computationally feasible by gradient descent algorithms (see Chapter 1).

The purpose of this section is to extend the principle of convex risk minimization to the ranking problem. Our analysis also provides a theoretical framework for the analysis of some successful ranking algorithms such as the RankBoost algorithm of Freund, Iyer, Schapire, and Singer [57]. In what follows we adapt the arguments of Lugosi and Vayatis [12] to the ranking problem.

The basic idea is to consider ranking rules induced by real-valued functions, that is, ranking rules of the form $r(x,x') = 2\mathbb{I}_{[f(x,x')>0]} - 1$, where $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is some measurable real-valued function. With a slight abuse of notation, we will denote by

$$L(f) = \mathbb{P}\{\operatorname{sgn}(Z) \cdot f(X,X') < 0\} = L(r)$$

the risk of the ranking rule induced by f. (Here $\operatorname{sgn}(x) = 1$ if $x > 0$, $\operatorname{sgn}(x) = -1$ if $x < 0$, and $\operatorname{sgn}(x) = 0$ if $x = 0$.) Let $\varphi : \mathbb{R} \to [0,\infty)$ be a convex *cost function* satisfying $\varphi(0) = 1$ and $\varphi(x) \geq \mathbb{I}_{[x \geq 0]}$. Typical choices of φ include the exponential cost function $\varphi(x) = e^x$, the "logit" function $\varphi(x) = \log_2(1 + e^x)$, or the "hinge loss" $\varphi(x) = (1+x)_+$. Define the *ranking φ-risk functional* associated to the cost function φ by

$$A(f) = \mathbb{E}\varphi(-\operatorname{sgn}(Z) \cdot f(X,X')) \ .$$

Obviously, $L(f) \leq A(f)$. We denote by $A^* = \inf_f A(f)$ the "optimal" value of the cost functional where the infimum is taken over all measurable functions $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

The most natural estimate of the cost functional $A(f)$, based on the training data D_n, is the *empirical ranking φ-risk functional* defined by the U-statistic

$$A_n(f) = \frac{1}{n(n-1)} \sum_{i \neq j} \varphi(-\operatorname{sgn}(Z_{i,j}) \cdot f(X_i, X_j)) \ .$$

The ranking rules based on *convex risk minimization* minimize, over a set $\mathcal{F}$ of real-valued functions $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the empirical ranking φ-risk functional A_n, that is, we choose $\hat{f}_n = \arg\min_{f \in \mathcal{F}} A_n(f)$ and assign the corresponding ranking rule $\hat{r}_n(x,x') = 2\mathbb{I}_{[\hat{f}_n(x,x')>0]} - 1$. Here we assume implicitly that the minimum exists.

By minimizing convex risk functionals, one hopes to make the excess convex risk $A(\hat{f}_n) - A^*$ small. This is meaningful for ranking if one can relate the excess convex risk to the excess ranking risk $L(\hat{f}_n) - L^*$. This may be done quite generally thanks to a result of Bartlett, Jordan, and McAuliffe [27] (see Chapter 2). In short, to analyze the excess ranking risk $L(f) - L^*$ for convex risk minimization, it suffices to bound the excess convex

risk. This may be done by decomposing it into "estimation" and "approximation" errors as follows:

$$A(\hat{f}_n) - A^*(f) \leq \left(A(\hat{f}_n) - \inf_{f \in \mathcal{F}} A(f) \right) + \left(\inf_{f \in \mathcal{F}} A(f) - A^* \right) .$$

Clearly, we may (loosely) bound the excess convex risk over the class $\mathcal{F}$ as

$$A(\hat{f}_n) - \inf_{f \in \mathcal{F}} A(f) \leq 2 \sup_{f \in \mathcal{F}} |A_n(f) - A(f)| ,$$

and we can then state the analogue of Proposition 25 for convex φ-risk functionals which is also based on U-statistics' first representation.

Proposition 26 (Clémençon, Lugosi, and Vayatis [7]) *Let $\hat{f}_n$ be the ranking rule minimizing the empirical convex risk functional $A_n(f)$ over a class $\mathcal{F}$ of functions uniformly bounded by $-B$ and B. Then, with probability at least $1 - \delta$,*

$$A(\hat{f}_n) - \inf_{f \in \mathcal{F}} A(f) \leq 8B\varphi'(B)\hat{r}_n(\mathcal{F}) + \sqrt{\frac{2B^2 \log(1/\delta)}{n}}$$

where $\hat{r}_n$ denotes the Rademacher average

$$\hat{r}_n(\mathcal{F}) = \mathbb{E} \sup_{f \in \mathcal{F}} \left(\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \epsilon_i \cdot f(X_i, X_{\lfloor n/2 \rfloor + i}) \right) ,$$

and $\epsilon_1, ..., \epsilon_n$ are i.i.d. Rademacher random variables.

It remains to provide examples where the Rademacher average describing the learning complexity can be controlled. Indeed, many interesting bounds are available for the Rademacher average of various classes of functions.

Example 7 - Boosting-type ranking.

In analogy to boosting-type classification problems, one may consider a class $\mathcal{F}_B$ of functions defined by

$$\mathcal{F}_B = \left\{ f(x, x') = \sum_{j=1}^{N} w_j g_j(x, x') \, : \, N \in \mathbb{N}, \, , \sum_{j=1}^{N} |w_j| = B, \, g_j \in \mathcal{R} \right\}$$

where $\mathcal{R}$ is a class of ranking rules as defined in Section 3.3.1. In this case it is easy to see that

$$\hat{r}_n(\mathcal{F}_B) \leq B\hat{r}_n(\mathcal{R}) \leq cB \sqrt{\frac{V}{n}}$$

where V is the VC dimension of the "base" class $\mathcal{R}$.

Summarizing, we have shown that a ranking rule based on the empirical minimization $A_n(f)$ over a class of ranking functions $\mathcal{F}_B$ of the form defined above, the excess ranking risk satisfies, with probability at least $1 - \delta$,

$$A(\hat{f}_n) - A^* \leq 8c \; B^2\varphi'(B) \sqrt{\frac{V}{n}} + \sqrt{\frac{2B^2 \log(1/\delta)}{n}} + \left(\inf_{f \in \mathcal{F}_B} A(f) - A^* \right) .$$

This inequality may be used to derive the consistency of such ranking rules. For example, the following corollary is immediate.

Corollary 27 (Clémençon, Lugosi, and Vayatis [7]) *Let $\mathcal{R}$ be a class of ranking rules of finite* VC *dimension V such that the associated class of functions $\mathcal{F}_B$ is rich in the sense that*

$$\lim_{B\to\infty} \inf_{f\in\mathcal{F}_B} A(f) = A^*$$

for all distributions of (X,Y). Then if $\hat{f}_n$ is defined as the empirical minimizer of $A_n(f)$ over $\mathcal{F}_{B_n}$ where the sequence B_n satisfies $B_n \to \infty$ and $B_n^2\varphi'(B_n)/\sqrt{n} \to 0$, then

$$\lim_{n\to\infty} L(\hat{f}_n) = L^* \quad \textit{almost surely.}$$

Classes $\mathcal{R}$ satisfying the conditions of the corollary exist, we refer to Section 2.2 for more details. ▌

Example 8 - SVM ranking.

Proposition 26 can also be used for establishing performance bounds for kernel methods such as support vector machines. A prototypical kernel-based ranking method may be defined as follows. To lighten notation, we write $\mathcal{W} = \mathcal{X} \times \mathcal{X}$.

Let $k : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ be a symmetric positive definite function, that is,

$$\sum_{i,j=1}^{n} \alpha_i\alpha_j k(w_i, w_j) \geq 0\,,$$

for all choices of n, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $w_1, \ldots, w_n \in \mathcal{W}$.

A kernel-type ranking algorithm may be defined as one that performs minimization of the empirical convex risk $A_n(f)$ (typically based on the hinge loss $\varphi(x) = (1+x)_+$) over the class $\mathcal{F}_B$ of functions defined by a ball of the associated reproducing kernel Hilbert space of the form (where $w = (x, x')$)

$$\mathcal{F}_B = \left\{ w \mapsto \sum_{j=1}^{N} c_j k(w_j, w) : \; N \in \mathbb{N},\; \sum_{i,j=1}^{N} c_i c_j k(w_i, w_j) \leq B^2, w_1, \ldots, w_N \in \mathcal{W} \right\}.$$

In this case we have

$$\hat{r}_n(\mathcal{F}_B) \leq \frac{2B}{n}\mathbb{E}\sqrt{\sum_{i=1}^{\lfloor n/2 \rfloor} k((X_i, X_{\lfloor n/2 \rfloor + i}), (X_i, X_{\lfloor n/2 \rfloor + i}))}\,,$$

see, for example, Boucheron, Bousquet, and Lugosi [34]. Once again, universal consistency of such kernel-based ranking rules may be derived in a straightforward way if the approximation error $\inf_{f\in\mathcal{F}_B} A(f) - A^*$ can be guaranteed to go to zero as $B \to \infty$. For the approximation properties of such kernel classes we refer the reader to Cucker and Smale [46], Scovel and Steinwart [100], Smale and Zhou [103], Steinwart [104]. ▌

3.4 Fast rates of convergence

The results provided in the previous section do not reveal any particular feature of the ranking problem except for the complexity measures involved. Indeed, by initially splitting the sample and considering independent pairs of observations, one would obtain the same rates of convergence up to constants. The superiority of U-statistics-based estimators upon averages of i.i.d. random variables can be established when variance is involved in the analysis. The two following observations will lead to the main outcome of our study:

1. The structure of the U-statistic is better described by Hoeffding's decomposition rather than the average of 'sums of i.i.d.' blocks representation. In particular, Hoeffding's decomposition makes explicit the variance term of the statistic.

2. The analysis of fast rates of convergence (faster than $n^{-1/2}$) relies heavily on a variance control condition which limits the range of possible distributions for which these rates can be achieved. It is well known (see, e.g., §5.2 in the survey [34] and the references therein) that tighter bounds for the excess risk in the context of binary classification may be obtained if one can control the variance of the excess risk by its expected value. In classification this can be guaranteed under certain "low-noise" conditions that have already been discussed in Chapter 2.

Next we examine the possibilities of obtaining such improved performance bounds for empirical ranking risk minimization.

Set first

$$q_r((x,y),(x',y')) = \mathbb{I}_{[(y-y')\cdot r(x,x')<0]} - \mathbb{I}_{[(y-y')\cdot r^*(x,x')<0]}$$

and consider the following estimate of the *excess risk* $\Lambda(r) = L(r) - L^* = \mathbb{E}q_r((X,Y),(X',Y'))$:

$$\Lambda_n(r) = \frac{1}{n(n-1)} \sum_{i \neq j} q_r((X_i,Y_i),(X_j,Y_j)),$$

which is a U-statistic of degree 2 with symmetric kernel q_r. Clearly, the minimizer $\hat{r}_n$ of the empirical ranking risk $L_n(r)$ over $\mathcal{R}$ also minimizes the empirical excess risk $\Lambda_n(r)$. To study this minimizer, consider the Hoeffding decomposition of $\Lambda_n(r)$:

$$\Lambda_n(r) - \Lambda(r) = 2T_n(r) + W_n(r)\,,$$

where

$$T_n(r) = \frac{1}{n} \sum_{i=1}^{n} h_r(X_i,Y_i)$$

is a sum of i.i.d. random variables with

$$h_r(x,y) = \mathbb{E}q_r((x,y),(X',Y')) - \Lambda(r)$$

and

$$W_n(r) = \frac{1}{n(n-1)} \sum_{i \neq j} \widehat{h}_r((X_i,Y_i),(X_j,Y_j))$$

is a *degenerate* U-statistic with symmetric kernel

$$\widehat{h}_r((x,y),(x',y')) = q_r((x,y),(x',y')) - \Lambda(r) - h_r(x,y) - h_r(x',y') .$$

In the analysis we show that the contribution of the degenerate part $W_n(r)$ of the U-statistic is negligible compared to that of $T_n(r)$. This means that minimization of Λ_n is approximately equivalent to minimizing $T_n(r)$. But since $T_n(r)$ is an average of i.i.d. random variables, this can be studied by known techniques worked out for empirical risk minimization.

It is well known from the theory of empirical risk minimization (see Tsybakov [108], Bartlett and Mendelson [28], Koltchinskii [74], Massart [88]), that, in order to improve the rates of convergence (such as the bound $O(\sqrt{V/n})$ obtained for VC classes in Section 3.3.1), it is necessary to impose some conditions on the joint distribution of (X,Y). In our case the key assumption takes the following form:

Assumption 28 *There exist constants* $c > 0$ *and* $\alpha \in [0,1]$ *such that for all* $r \in \mathcal{R}$,

$$\mathrm{Var}(h_r(X,Y)) \le c\,\Lambda(r)^\alpha .$$

The improved rates of convergence will depend on the value of α. Interestingly, this assumption is satisfied for a surprisingly large family of distributions (see [7] and the example below), guaranteeing improved rates of convergence. For $\alpha = 0$ the assumption is always satisfied and the corresponding performance bound does not yield any improvement over those of Section 3.3.1. However, in many natural examples Assumption 28 is satisfied with values of α close to one, providing significant improvements in the rates of convergence.

The main tool for handling the degenerate part is a new general moment inequality for U-processes that may be interesting on its own right.

Theorem 29 (CLÉMENÇON, LUGOSI, AND VAYATIS [7]) *Let* $X, X_1, \ldots, X_n$ *be i.i.d. random variables and let* $\mathcal{F}$ *be a class of kernels. Consider a degenerate* U*-process* Z *of order 2 indexed by* $\mathcal{F}$,

$$Z = \sup_{f\in\mathcal{F}} \left| \sum_{i,j} f(X_i, X_j) \right|$$

where $\mathbb{E}f(X,x) = 0$, $\forall x, f$. *Assume also* $f(x,x) = 0$, $\forall x$ *and* $\sup_{f\in\mathcal{F}} \|f\|_\infty = F$. *Let* $\epsilon_1, \ldots, \epsilon_n$ *be i.i.d. Rademacher random variables and introduce the random variables*

$$Z_\epsilon = \sup_{f\in\mathcal{F}} \left| \sum_{i,j} \epsilon_i \epsilon_j f(X_i, X_j) \right| ,$$

$$U_\epsilon = \sup_{f\in\mathcal{F}} \sup_{\alpha:\|\alpha\|_2\le 1} \sum_{i,j} \epsilon_i \alpha_j f(X_i, X_j) ,$$

$$M = \sup_{f\in\mathcal{F}} \max_{k=1\ldots n} \left| \sum_{i=1}^n \epsilon_i f(X_i, X_k) \right| .$$

Then there exists a universal constant $C > 0$ *such that for all* n *and* $q \geq 2$,

$$(\mathbb{E}Z^q)^{1/q} \leq C\left(\mathbb{E}Z_\epsilon + q^{1/2}\mathbb{E}U_\epsilon + q(\mathbb{E}M + Fn) + q^{3/2}Fn^{1/2} + q^2F\right).$$

Also, there exists a universal constant C *such that for all* n *and* $t > 0$,

$$\mathbb{P}\{Z > C\mathbb{E}Z_\epsilon + t\} \leq \exp\left(-\frac{1}{C}\min\left(\left(\frac{t}{\mathbb{E}U_\epsilon}\right)^2, \frac{t}{\mathbb{E}M + Fn}, \left(\frac{t}{F\sqrt{n}}\right)^{2/3}, \sqrt{\frac{t}{F}}\right)\right).$$

Remark 2 This result is based on moment inequalities obtained for empirical processes and Rademacher chaoses in Bousquet, Boucheron, Lugosi, and Massart [35] and generalizes an inequality due to Arcones and Giné [21]. We also refer to the corresponding results obtained for U-statistics by Adamczak [17], Giné, Latala, and Zinn [61], and Houdré and Reynaud-Bouret [69]. ▮

In order to state the main result of this section, we need to introduce some quantities related to the class $\mathcal{R}$. Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. Rademacher random variables independent of the (X_i, Y_i)'s. Let

$$\begin{aligned}
Z_\epsilon &= \sup_{r\in\mathcal{R}}\left|\sum_{i,j}\epsilon_i\epsilon_j\widehat{h}_r((X_i,Y_i),(X_j,Y_j))\right|, \\
U_\epsilon &= \sup_{r\in\mathcal{R}}\sup_{\alpha:\|\alpha\|_2\leq 1}\sum_{i,j}\epsilon_i\alpha_j\widehat{h}_r((X_i,Y_i),(X_j,Y_j)), \\
M &= \sup_{r\in\mathcal{R}}\max_{k=1,\ldots,n}\left|\sum_{i=1}^n\epsilon_i\widehat{h}_r((X_i,Y_i),(X_k,Y_k))\right|.
\end{aligned}$$

Introduce the "loss function"

$$\ell(r,(x,y)) = 2\mathbb{E}\mathbb{I}_{[(y-Y)\cdot r(x,X)<0]} - L(r)$$

and define

$$\nu_n(r) = \frac{1}{n}\sum_{i=1}^n \ell(r,(X_i,Y_i)) - L(r).$$

(Observe that $\nu_n(r)$ has zero mean.) Also, define the pseudo-distance

$$d(r,r') = \left(\mathbb{E}\left(\mathbb{E}[\mathbb{I}_{[r(X,X')\neq r'(X,X')]}|X]\right)^2\right)^{1/2}.$$

Let $\varphi : [0,\infty) \to [0,\infty)$ be a nondecreasing function such that $\varphi(x)/x$ is nonincreasing and $\varphi(1) \geq 1$ such that for all $r \in \mathcal{R}$,

$$\sqrt{n}\mathbb{E}\sup_{r'\in\mathbb{R}, d(r,r')\leq\sigma}|\nu_n(r) - \nu_n(r')| \leq \varphi(\sigma).$$

Theorem 30 (Clémençon, Lugosi, and Vayatis [7]) *Consider a minimizer $\hat{r}_n$ of the empirical ranking risk $L_n(r)$ over a class $\mathcal{R}$ of ranking rules and assume Assumption 28. Then there exists a universal constant C such that, with probability at least $1-\delta$, the ranking risk of $\hat{r}_n$ satisfies*

$$\begin{aligned} L(\hat{r}_n) - L^* \quad \leq \quad & 2\left(\inf_{r\in\mathcal{R}} L(r) - L^*\right) \\ & + C\left(\frac{\mathbb{E}Z_\epsilon}{n^2} + \frac{\mathbb{E}U_\epsilon\sqrt{\log(1/\delta)}}{n^2} + \frac{\mathbb{E}M\log(1/\delta)}{n^2} + \frac{\log(1/\delta)}{n} + \rho^2\log(1/\delta)\right) \end{aligned}$$

where $\rho > 0$ is the unique solution of the equation

$$\sqrt{n}\rho^2 = \varphi(\rho^\alpha) \ .$$

The proof relies on Hoeffding's decomposition, on Theorem 29, and applies Theorem 8.3 of Massart [88]. Theorem 30 provides a performance bound in terms of expected values of certain Rademacher chaoses indexed by $\mathcal{R}$ and local properties of an ordinary empirical process. These quantities have been thoroughly studied and well understood, and may be easily bounded in many interesting cases. Below we will work out an example when $\mathcal{R}$ is a VC class of indicator functions.

Observe that the only condition for the distribution is that the variance of h_r can be bounded in terms of $\Lambda(r)$. In the example below, we show that Assumption 28 can be satisfied with $\alpha > 0$ under mild conditions on the distribution.

Example 9 - The bipartite ranking problem. We derive a simple sufficient condition for achieving fast rates of convergence for the bipartite ranking problem. Recall that here it suffices to consider ranking rules of the form $r(x,x') = 2\mathbb{I}_{[s(x)\geq s(x')]} - 1$ where s is a scoring function.

Noise assumption. *There exist constants $c > 0$ and $\alpha \in [0,1]$ such that for all $x \in \mathcal{X}$,*

$$\mathbb{E}_{X'}(|\eta(x) - \eta(X')|^{-\alpha}) \leq c \, . \tag{3.2}$$

Under this noise assumption, the variance control condition is satisfied. For $\alpha < 1$, there is a wide range of distributions fulfilling this property. For instance, we can consider that $\eta(x) = \mathbb{P}\{Y = 1|X = x\}$ is such that the random variable $\eta(X)$ has an absolutely continuous distribution on $[0,1]$ with a density bounded by B. Indeed, it requires that the distribution of $\eta(X)$ is sufficiently spread out, for example it cannot have atoms or infinite peaks in its density. Under such a condition a rate of convergence of the order of $n^{-1+\epsilon}$ is achievable for any $\epsilon > 0$. Note that we crucially used the reduced variance of the U-statistic $L(\hat{r}_n)$ to derive fast rates from the rather weak condition (3.2). Applying a similar reasoning for the variance of $q_s((X,Y),(X',Y'))$ (which would be the case if one considered a risk estimate based on independent pairs by splitting the training data into two halves, see Section 3.3.1), would have led to the condition:

$$|\eta(x) - \eta(x')| \geq c, \tag{3.3}$$

for some constant c, and $x \neq x'$. This condition is clearly too restrictive since it is satisfied only when $\eta(X)$ has a discrete distribution. ▮

In [5], we also derive similar conditions in the case of regression data. We mention that in the case of a noiseless regression model, the noise assumption is automatically satisfied with $c = 1$ and $\alpha = 1$.

In order to illustrate Theorem 30 and provide a simpler statement, we now consider the case when $\mathcal{R}$ is a VC class, that is, it has a finite VC dimension V. We obtain a simple bound which reveals that the value of α in Assumption 28 determines the magnitude of the last term which, in turn, dominates the right-hand side (apart from the approximation error term).

Corollary 31 (Clémençon, Lugosi, and Vayatis [7]) *Consider the minimizer $\hat{r}_n$ of the empirical ranking risk $L_n(r)$ over a class $\mathcal{R}$ of ranking rules of finite VC dimension V and assume Assumption 28. Then there exists a universal constant C such that, with probability at least $1 - \delta$, the ranking risk of $\hat{r}_n$ satisfies*

$$L(\hat{r}_n) - L^* \leq 2\left(\inf_{r \in \mathcal{R}} L(r) - L^*\right) + C\left(\frac{V \log(n/\delta)}{n}\right)^{1/(2-\alpha)}$$

Based on the bounds presented here, one may design penalized empirical minimizers of the ranking risk that select the class $\mathcal{R}$ from a collection of classes achieving this objective. The techniques presented in Massart [88] and Koltchinskii [74] may be used in a relatively straightforward manner to derive such "oracle inequalities" for penalized empirical risk minimization in the present framework.

3.5 Ranking the best instances

We now discuss the choice of risk functionals for the ranking problem in order to include priors on the desired ranking. We consider the classification model with binary-valued labels $Y \in \{-1, +1\}$ where we search for a real-valued scoring function s as close as possible to the equivalence class of functions obtained as increasing transforms of $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$.

Previously, we have been considering a version of the ranking problem where the ranking risk of a scoring function s is given by:

$$L(s) = \mathbb{P}\{(Y - Y')(s(X) - s(X')) < 0\} .$$

Interestingly, it can be proved that minimizing the ranking risk $L(s)$ is equivalent to maximizing the well-known AUC criterion. This is trivial once we write down the probabilistic interpretation of the AUC ([8]):

$$\text{AUC}(s) = \mathbb{P}\left\{s(X) > s(X') \mid Y = 1, Y' = -1\right\} = 1 - \frac{1}{2p(1-p)} L(s) .$$

However, in some applications such as information retrieval or credit risk screening, the AUC criterion is of limited interest. Such a criterion weights any discordant pair of observations uniformly while the challenge there lies in ranking the "best" - according to the optimal scoring function - instances. We point out that this problem is of huge interest in the applications aforementioned but has not been addressed yet from a theoretical point of view. Empirical criteria have been proposed in the machine learning community (e.g. the Discounted Cumulative Gain [70] which can be assimilated to a linear rank statistics, or the "p-norm push" approach of Rudin [97]) but, as far as we know, optimality issues have not been explored. Indeed, the following discussion will reveal that the very idea of a criterion for local ranking derived from the AUC presents some unexpected features.

We state the *local ranking* problem as follows:

> Let $u \in (0,1)$ be fixed, the problem is to rank the instances of the input space $\mathcal{X}$ in order to have a proportion u of the best instances ranked as accurately as possible.

We face here a two-fold problem: find the best instances and rank them simultaneously. Here, there is no natural criterion at hand, so we will rather formulate what kind of scoring functions would be optimal for such a problem (except the regression function η itself).

First the set of best instances can be described by the following set:

$$C_u^* = \{x \in \mathcal{X} \mid \eta(x) > Q(\eta, 1-u)\}$$

where $Q(\eta, 1-u)$ is the $(1-u)$-quantile of the random variable $\eta(X)$ (meaning that $Q(\eta,u) = F_\eta^{-1}(1-u)$ with F_η^{-1} being the generalized inverse of $F_\eta(z) = \mathbb{P}\{\eta(X) \leq z\}$). It is noteworthy that the set C_u^* is invariant by strictly increasing transformations of η.

Hence, the optimal elements for the local ranking problem belong to the equivalence class (functions defined up to the composition with a nondecreasing transformation) defined by the scoring function s^*:

$$s^*(x) = \begin{cases} = & \eta(x) & \text{if } x \in C_u^* \\ \\ < & \inf\limits_{C_u^*} \eta & \text{if } x \notin C_u^* \end{cases}$$

In [8], we have formulated the following criterion which extends (and localizes) the AUC criterion:

$$AUC_u(s) = \mathbb{P}\left\{s(X) > s(X') \ , \ s(X) \geq Q(s, 1-u) \ \mid \ Y = 1, Y' = -1\right\} \ ,$$

where $Q(s, 1-u) = F_s^{-1}(1-u)$ and $F_s(z) = \mathbb{P}\{s(X) \leq z\}$. This criterion obviously boils down to the standard AUC criterion for $u = 1$.

The following theorem states that the scoring function s^* maximizes this criterion and that $AUC_u(s)$ may be decomposed as a sum of a 'power' term $\beta(s,u)$ and a local ranking risk term. Before stating the result, we set the following notations:

- the candidate set for the set of best instances for each scoring function

$$C_{s,u} = \{x \in \mathcal{X} \mid s(x) > Q(s, 1-u)\}$$

- the false positive rate at level $1-u$

$$\alpha(s,u) = \mathbb{P}\{s(X) \geq Q(s,1-u) \mid Y = -1\}$$

- the true positive rate at level $1-u$

$$\beta(s,u) = \mathbb{P}\{s(X) \geq Q(s,1-u) \mid Y = +1\}$$

- the local ranking risk on a measurable set $C \subset \mathcal{X}$

$$L(s,C) = \mathbb{P}\left\{(s(X)-s(X'))(Y-Y') < 0 \ , \ (X,X') \in C^2\right\} \ .$$

Theorem 32 (Clémençon and Vayatis [8]) *Let $u \in (0,1)$. We have*

$$\forall s, \quad AUC_u(s) \leq AUC_u(s^*) \ .$$

Moreover:

$$AUC_u(s) = \int_0^{\alpha(s,u)} \beta(s,v)dv + \beta(s,u)(1-\alpha(s,u))$$

$$= \beta(s,u) - \frac{1}{2p(1-p)} L(s, C_{s,u}) \ .$$

Remark 3 Note that naive strategies (see the partial AUC approach in [53]) which consist in optimizing a truncated AUC criterion as $\int_0^{\alpha(s,u)} \beta(s,v)dv$ do not lead to the desired result. Indeed, as the scoring function s comes closer to η, the power $\beta(s,\alpha)$ increases but the integration domain shrinks and it is not clear whether η is a maximizer.

This result highlights the fact that divide-and-conquer strategies (first find the best instances, and then rank them according to the local AUC criterion) will fail in solving the local ranking problem.

In a joint work with Clémençon [8], we first study the problem of finding the best instances by reducing it to a special classification problem. We prove consistency of ERM procedures for the setup of classification with mass-constraints and study fast rates of convergence. We discuss further the statistical aspects of the local ranking problem which raises new challenges for statistical learning theory.

References

[17] R. Adamczak. Moment inequalities for U-statistics. Technical report, Institute of Mathematics of the Polish Academy of Sciences, 2005.

[18] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425, 2005.

[19] M. Aizerman, E. Braverman, and L. Rozonoer. *Method of Potential Functions in the Theory of Learning Machines.* Nauka, Moscow [in Russian], 1970.

[20] M. Aizerman, E. Braverman, and L. Rozonoer. Extrapolative problems in automatic control and the method of potential functions. *American Math. Society Translations*, 87:281–303, 1970.

[21] M. A. Arcones and E. Giné. U-processes indexed by Vapnik-Cervonenkis classes of functions with applications to asymptotics and bootstrap of u-statistics with estimated parameters. *Stochastic Processes and their Applications*, 52:17–38, 1994.

[22] A.R. Barron. Neural net approximation. In *Yale Workshop on Adaptive and Learning Systems*, 1992.

[23] A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39:930–944, 1993.

[24] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields* 113:301–415, 1999.

[25] P.L. Bartlett, S. Boucheron, and G. Lugosi. Model Selection and Error Estimation. *Machine Learning*, 45:85–113, 2001.

[26] P.L. Bartlett, O. Bousquet and S. Mendelson. Localized Rademacher Complexities. *Proceedings of the 15th annual conference on Computational Learning Theory*, 44-58, 2002.

[27] P.L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138-156, 2006.

[28] P.L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135, 2006.

[29] A. Ben-Tal, and A.S. Nemirovski. The conjugate barrier mirror descent method for non-smooth convex optimization. MINERVA Optimization Center Report, Technion Institute of Technology. Available at http://iew3.technion.ac.il/Labs/Opt/opt/Pap/CP_MD.pdf, 1999.

[30] A. Ben-Tal, T. Margalit and A.S. Nemirovski. The ordered subsets mirror descent optimization method and its use for positron emission tomography reconstruction problem. *SIAM J. on Optimization*, 12:79–108, 2001.

[31] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4:329–375, 1998.

[32] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of Support Vector Machines. *Submitted*, 2004.

[33] B. Boser, I. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. Proceedings of COLT 1992, 144–152, 1992.

[34] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM. Probability and Statistics*, 9:323–375, 2005.

[35] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *The Annals Probability*, 33:514–560, 2005.

[36] L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26:801–849, 1998.

[37] L. Breiman. Some infinite theory for predictor ensembles. Technical Report 577, Statistics Department, UC Berkeley, 2000.

[38] P. Bühlmann and B. Yu. Boosting with the l_2-loss: Regression and classification. *Journal of the American Statistical Association*, 98:324–340, 2003.

[39] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of online learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

[40] N. Cesa-Bianchi and G. Lugosi. *Prediction, Leanring, and Games.* Cambridge University Press, 2006.

[41] A. Cohen, W. Dahmen, and R. De Vore. Approximation and Learning by Greedy Algorithms. Preprint, 2006.

[42] M. Collins, R.E. Schapire, and Y. Singer. Learning to order things. *Journal of Artificial Intelligence Research*, 10:243–270, 1999.

[43] M. Collins, R.E. Schapire, and Y. Singer. Logistic regression, AdaBoost and Bregman distances. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, 2000.

[44] C. Cortes and M. Mohri. AUC Optimization vs. Error Rate Minimization. In *Advances in Neural Information Processing Systems*, eds Thrun, S. Saul, L. Schölkopf, B., MIT Press, 2004.

[45] C. Cortes and V.N. Vapnik. Support Vector Networks. *Machine Learning*, 20(3):273-297, 1995.

[46] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

[47] G. Cybenko. Approximations by superpositions of sigmoidal functions. *Math. Control, Signals, Systems*, 2:303–314, 1989.

[48] C. Darken, M. Donahue, L. Gurvits, and E. Sontag. Rates of Convex Approximation in Non-Hilbert Spaces. *Constructive Approximation*, 13(2):187-220, 1997.

[49] V.H. de la Peña and E. Giné. *Decoupling: from Dependence to Independence.* Springer, New York, 1999.

[50] M. Dettling and P. Bühlmann. Boosting for Tumor Classification with Gene Expression Data. *Bioinformatics*, 2003. (R-code available at `http://stat.ethz.ch/~dettling/boosting.html`).

[51] R. De Vore, G. Kerkyacharian, D. Picard, and V. Temlyakov. On Mathematical Methods for Supervised Learning. *J. FoCM*, 6:3–58, 2006.

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

[53] L.E. Dodd and M.S. Pepe. Partial AUC Estimation and Regression. Biometrics, 59(3):614–623, 2003.

[54] R.M. Dudley. *Uniform Central Limit Theorems.* Cambridge University Press, 1999.

[55] J.P. Egan. *Signal Detection Theory and ROC Analysis.* Academic Press, 1975.

[56] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121:256–285, 1995.

[57] Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4(6):933–969, 2004.

[58] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

[59] Y. Freund and R.E. Schapire. Discussion of the paper "additive logistic regression: a statistical view of boosting" by J. Friedman, T. Hastie and R. Tibshirani. *The Annals of Statistics*, 38(2):391–393, 2000.

[60] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion). *The Annals of Statistics*, 28:307–337, 2000.

[61] E. Giné, R. Latała, and J. Zinn. Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II—Progress in Probability*, pages 13–38. Birkhauser, 2000.

[62] F. Girosi and G. Anzelloti. Rates of convergence for radial basis functions and neural networks. *Artificial Neural Networks for Speech and Vision*, p.169-176, Chapman and Hall, 1993.

[63] A. Grove and D. Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98), 1998.

[64] J. Hajek and Z. Sidak *Theory of Rank Tests.* Academic Press, 1967.

[65] J.A. Hanley and J. McNeil. The meaning and use of the area under a ROC curve. Radiology, 143: 29-36, 1982.

[66] T. Hastie and R.J. Tibshirani. *Generalized Additive Models.* Chapman and Hall, London, U. K., 1990.

[67] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In A. Smola, P.L. Bartlett, B.Schölkopf, and D.Schuurmans (eds.), *Advances in Large Margin Classifiers*, The MIT Press, 115–132, 2000.

[68] K. Hornik, M. Stinchcombe, and H. White. Multi-layer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[69] C. Houdré and P. Reynaud-Bouret. Exponential Inequalities, with constants, for U-statistics of order two. Stochastic Inequalities and Applications - Progress in Probability, Birkhauser, 2003.

[70] K. Järvelin and J. Kekäläinen IR evaluation methods for retrieving highly relevant documents. In Proceedings of SIGRI 2000, p.41–48, 2000.

[71] Juditsky, A., Nemirovski, A.: Functional aggregation for nonparametric estimation. The Annals of Statistics, Vol.28(3) (2000) 681–712.

[72] J. Kivinen and M.K. Warmuth. Additive versus exponentiated gradient updates for linear prediction. *Information and Computation*, 132(1):1–64, 1997.

[73] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47:1902-1914, 2001.

[74] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6), 2006.

[75] V. Koltchinskii, G. Lugosi and S. Mendelson. A note on the richness of convex hulls of VC classes. Electronic Communications in Probability, 8:1–3, 2003.

[76] V. Koltchinskii and D. Panchenko. Empirical margin distribution and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30:1–50, 2002.

[77] L. Le Cam. On some asymptotic properties of maximum likelihood estimated and related Bayes estimates. *Univ. Calif. Publ. Statist.*, 3:27-98, 1953.

[78] M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer-Verlag, New York, 1991.

[79] G. Lugosi. Pattern classification and learning theory. In L. Györfi, editor, *Principles of Nonparametric Learning*, pages 5–62. Springer, Wien, 2002.

[80] V. Maiorov, R. Meir, and J. Ratsaby. On the Approximation of Functional Classes Equipped with a Uniform Measure using Ridge Functions. *Jour. of Approximation Theory*, 99:95-111, 1999.

[81] E. Mammen and A. Tsybakov. Smooth Discriminant Analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[82] S. Mannor, R. Meir, and T. Zhang. The consistency of greedy algorithms for classification. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*, 2002.

[83] J.I. Marden. *Analyzing and Modeling Rank Data.* Chapman & Hall, New York, 1995.

[84] L. Mason, P.L. Bartlett, and J. Baxter. Direct Optimization of Margins Improves Generalization in Combined Classifiers. Proceedings of NIPS 1998, 288–294, 1998.

[85] L. Mason, P.L. Bartlett, and J. Baxter. Improved generalization through explicit optimization of margins. *Machine Learning*, 38(3):243–255, 2000.

[86] L. Mason, J. Baxter, P.L. Bartlett, and M. Frean. Functional gradient techniques for combining hypotheses. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 221–247. MIT Press, Cambridge, MA, 1999.

[87] P. Massart. Some applications of concentration inequalities in statistics. *Annales de la Faculté des Sciences de Toulouse, Mathématiques*, 9(2):245-303, 2000.

[88] P. Massart. *Concentration inequalities and model selection.* Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics. Springer, 2006.

[89] P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34, 2006.

[90] R. Meir and V. Maiorov. On the Optimality of neural network approximation using incremental algorithms. *IEEE Trans. Neural Network*, 11(2):323-337, 2000.

[91] R. Meir and G. Rätsch. An introduction to boosting and leveraging. In S. Mendelson and A. Smola, editors, *Advanced Lectures on Machine Learning*, LNCS, pages 119-184. Springer, 2003.

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

[93] A.S. Nemirovski and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

[94] A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8;143-196, 1999.

[95] G. Rätsch, T. Onoda, and K.-R. Mller. Soft Margins for AdaBoost. *Machine Learning*, 42(3), 287–320, 2001.

[96] P. Rigollet. *Inégalités d'oracle, agrégation et adaptation*. Thèse de l'Université Pierre-et-Marie-Curie, 2006.

[97] C. Rudin Ranking with a P-Norm Push. Proceedings of COLT 2006, 2006.

[98] R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

[99] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26:1651–1686, 1998.

[100] S. Scovel and I. Steinwart. Fast rates for support vector machines using Gaussian kernels *The Annals of Statistics*, 2007. To appear.

[101] R.J. Serfling. Approximation theorems of mathematical statistics. John Wiley & Sons, 1980.

[102] E. Sontag. Feedback Stabilization Using Two-Hidden-Layer Nets. *IEEE Trans. Neural Networks*, 3:981-990, 1992.

[103] S. Smale and D.-X. Zhou. Estimating the approximation error in learning theory. Analysis and Applications, 1, pp. 17-41. Support Vector Machine Soft Margin Classifiers, 2003.

[104] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[105] I. Steinwart. Consistency of Support Vector Machines and other Regularized Kernel Machines. *IEEE Transactions on Information Theory*, 51:128–142, 2005.

[106] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996.

[107] A. Tsybakov. Optimal Rates of Aggregation. Proceedings of COLT'03, LNCS, Vol. 2777, Springer, 303–313, 2003.

[108] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32:135–166, 2004.

[109] S. van de Geer. *Empirical Processes in M-Estimation* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[110] A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes.* Springer-Verlag, New York, 1996.

[111] H. van Trees. *Detection, Estimation, and Modulation Theory: Part I.* John Wiley & Sons, 1968.

[112] V.N. Vapnik. *Estimation of Dependencies Based on Empirical Data.* Springer, 1982.

[113] V.N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.

[114] V.N. Vapnik. *Statistical Learning Theory.* Wiley-Interscience, 1998.

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

[116] V.N. Vapnik and A.Ya. Chervonenkis. Necessary and sufficient conditions for the uniform convergence of the means to their expectations. *Theory of Probability and their Applications*, 26(3):532–555, 1981.

[117] Y. Yang. Minimax Nonparametric Classification-Part I: Rates of Convergence. *IEEE Transaction on Information Theory*, vol. 45, pp. 2271-2284, 1999.

[118] Y. Yang. Minimax Nonparametric Classification-Part II: Model Selection for Adaptation. *IEEE Transaction on Information Theory*, vol. 45, pp. 2285-2292, 1999.

[119] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization (with discussion). *The Annals of Statistics*, 32:56–85, 2004.

[120] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In Proceedings of ICML'04, 2004.