Convergence and Consistency of Regularized Boosting With Weakly Dependent Observations

Aurélie C. Lozano, Sanjeev R. Kulkarni, *Fellow, IEEE*, and Robert E. Schapire

***Abstract*— This paper studies the statistical convergence and consistency of regularized boosting methods, where the samples need not be independent and identically distributed but can come from stationary weakly dependent sequences. Consistency is proven for the composite classifiers that result from a regularization achieved by restricting the 1-norm of the base classifiers' weights. The less restrictive nature of sampling considered here is manifested in the consistency result through a generalized condition on the growth of the regularization parameter. The weaker the sample dependence, the faster the regularization parameter is allowed to grow with increasing sample size. A consistency result is also provided for data-dependent choices of the regularization parameter.**

***Index Terms*— Bayes-risk consistency, beta-mixing, boosting, classification, dependent data, empirical processes, memory, non-i.i.d, regularization, penalized model selection.**

I. Introduction

A SIGNIFICANT advance in machine learning for classification has been the development of boosting algorithms [26]. Simply put, a boosting algorithm is an iterative procedure that combines weak prediction rules to produce a composite classifier, the idea being that one can obtain very precise prediction rules by combining rough ones. Each weak rule is obtained by running a weak learning algorithm under a different distribution over the training examples. Starting with the uniform distribution, the sample distribution is updated by reinforcing the weight of the misclassified examples upon which the weak learner will be forced to concentrate. The weak rules are combined using a weighted majority vote. The concept of boosting was originally introduced by Kearns & Valiant [15], followed by the first boosting algorithms of Schapire [25] and Freund [8], and later the establishment of AdaBoost [9] as the first practical boosting algorithm.

Manuscript received April 22, 2007; revised August 22, 2012; accepted August 23, 2013. Date of publication October 30, 2013; date of current version December 20, 2013. This work was supported in part by the National Science Foundation under Grants CCR-0312413, CCR-0325463, and IIS-0325500, in part by the Army Research Office under Contract DAAD19-00-1-0466, and in part by the Center for Science of Information, an NSF Science and Technology Center, under Grant agreement CCF-0939370.

A. C. Lozano is with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: aclozano@us.ibm.com).

S. R. Kulkarni is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: kulkarni@princeton.edu).

R. E. Schapire is with the Department of Computer Science, Princeton University, Princeton, NJ 08540 USA (e-mail: schapire@cs.princeton.edu).

Communicated by A. Krzyżak, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2013.2287726

Friedman et al. [10] have shown that AdaBoost performs stage-wise fitting of additive models under the exponential loss function and effectively minimizes an empirical loss function that differs from the probability of incorrect prediction. From this perspective, boosting methods in general can be seen as performing a greedy stage-wise minimization of various loss functions empirically. The question of whether boosting achieves Bayes-consistency then arises, since minimizing an empirical loss function does not necessarily imply minimizing the generalization error. When run a very long time, the original boosting methods such as AdaBoost, though resistant to overfitting, are not immune to it (see [10] and [13]). In the limit of infinite sample size, running these algorithms for an infinite number of rounds can also lead to a prediction error larger than the Bayes error. Consequently, one approach for the study of consistency is to modify the original boosting algorithms by imposing some constraints on the weights of the composite classifier to avoid overfitting. In these regularized versions of boosting, the 1-norm of the weights of the base classifiers is restricted to a fixed value. The minimization of the loss function is performed over the restricted class. This approach is proposed by Lugosi & Vayatis [19], and Mannor et al. [20]. Note that Zhang [32] adopts a similar approach, where a kernel formulation is considered whose penalty function is the square norm of the composite classifier. An alternative approach is to analyze how the prediction error L (Boosting (m, t)) for sample size m evolves with the number of rounds t to determine whether boosting is process consistent, i.e. whether $\lim_{m\to\infty} \inf_{t\in\{1,2,3,\ldots\}} L$ (Boosting $(m, t)) = L^*$, where L^* is the Bayes error. This strategy suggests stopping the boosting procedure early to achieve consistency, and has been adopted by Bartlett & Traskin [2], Jiang [14], Bühlman [5], and Zhang & Yu [33].

All the aforementioned studies concern i.i.d. processes, upon which much of the work on classical learning theory has been based. However, the postulate of independence is often invoked, rather than established from data observations. Consequently, results based on the i.i.d. assumption do not necessarily apply well to intermediary situations such as ones involving weakly dependent data. In order to obtain more general and widely applicable results, we examine in this paper the convergence and consistency of regularized boosting algorithms from samples that are no longer i.i.d. but come from stationary weakly dependent sequences. A practical motivation for our study of non i.i.d. sampling is that in many learning applications (e.g., in finance, speech recognition,

medical diagnosis, and data mining applications) observations are intrinsically temporal and hence often weakly dependent. Ignoring this dependency could seriously undermine the performance of the learning process (for instance, information related to the time-dependent ordering of samples would be lost). Recognition of this issue has led to several studies of non i.i.d. sampling (see [11], [12], [21], [23], [24], [28], [30]).

To cope with weak dependence we apply mixing theory which, through its definition of mixing coefficients, offers a powerful approach to extend results for the traditional i.i.d. observations to the case of weakly dependent or mixing sequences. We consider the β-mixing coefficients, whose mathematical definition is deferred to Sec. II-A. Intuitively, they provide a measure of how fast the dependence between observations diminishes as the distance between them increases. For instance, in the context of time series, the β-mixing coefficients characterize how weakly the "future" observations depend on the "past" ones. If certain conditions on the mixing coefficients are satisfied to reflect a sufficiently fast decline in the dependence between observations as their distance grows, we can establish counterparts to results for i.i.d. random processes. A comprehensive review of mixing theory results is provided in [7].

Our principal finding is that consistency of regularized boosting methods can be established in the case of non-i.i.d. samples coming from empirical sequences of stationary β-mixing sequences. Among the conditions that guarantee consistency, the mixing nature of sampling appears only through a generalization of the one on the growth of the regularization parameter stated for the i.i.d. case [19]. The weaker the sample dependence, the faster the regularization parameter is allowed to grow as the sample size increases. In our main consistency result, the regularization parameters are chosen before data observation. In addition, we provide a consistency result for data-dependent choices of the regularization parameters.

The rest of this paper is organized as follows. Section II provides an overview of mixing sequences and formulates the setup of the classification problem for training data coming from stationary β-mixing sequences. After a review of regularized boosting methods, the main results are stated and discussed in Section III. The proofs are given in the appendix.

II. Background and Setup

A. Mixing Sequences

For completeness, we first recall the definitions of strict stationarity and sigma-fields generated by random variables. A sequence of random variables $(W_1, W_2, \ldots)$ is strictly stationary if for every integers k and n, $(W_{1+k}, W_{2+k}, \ldots, W_{n+k})$ has the same joint distribution as $(W_1, W_2, \ldots, W_n)$. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and let $(E, \mathcal{E})$ be a measurable space. Consider a random variable W taking values in $(E, \mathcal{E})$. W is measurable with respect to $\mathcal{H}$ and $\mathcal{E}$. That is for all $A \in \mathcal{E}$, $W^{-1}A = \{\omega \in \Omega : W(\omega) \in A\}$ belongs to $\mathcal{H}$, i.e., $W^{-1}A$ is an event. Then, the sigma-field generated by W is the collection of events $\sigma W = \left\{W^{-1}A : A \in \mathcal{E}\right\}$. It can be shown that it is the set of all random variables of the form $f \circ W$, where f is a measurable function.

We now give an overview of mixing sequences. Let $\underline{W} = (W_i)_{i \geq 1}$ be a strictly stationary sequence of random variables, each having the same distribution P on $\mathcal{D} \subset R^d$. Let $\sigma_1^l = \sigma(W_1, W_2, \ldots, W_l)$ be the σ-field generated by $W_1, \ldots, W_l$. Similarly, define σ_{l+k}^{∞} as $\sigma_{l+k}^{\infty} = \sigma(W_{l+k}, W_{l+k+1}, \ldots,)$. The following mixing coefficients characterize how close a sequence $\underline{W}$ is to being independent.

Definition 1: For any sequence $\underline{W}$, the β-mixing coefficient is defined by

$$\beta_W(n) = \sup_k \mathbb{E} \sup \left\{ |\mathbb{P}\left(A|\sigma_1^k\right) - \mathbb{P}(A)| : A \in \sigma_{k+n}^{\infty} \right\},$$

where the expectation is taken w.r.t. σ_1^k.

Note that for a random variable W, $\mathbb{P}(A|\sigma_W)$ is the conditional probability of A given W which is usually simplified to $\mathbb{P}(A|W)$. In particular $\mathbb{P}(A|\sigma_1^k)$ is the conditional probability of A given $W_1, \ldots, W_k$.

Definition 2: A sequence $\underline{W}$ is called β-mixing if $\lim_{n\to\infty} \beta_W(n) = 0$. Further, it is algebraically β-mixing if there is a positive constant r_β such that $\beta_W(n) = O\left(n^{-r_\beta}\right)$.

In this paper, we will assume that the sequences we consider are algebraically β-mixing. This property implies that the dependence between observations decreases fast enough as the distance between them increases. The choice of β-mixing appears appropriate given previous results that showed "uniform convergence of empirical means uniformly in probability" and "probably approximately correct" properties to be preserved for β-mixing inputs [28]. Some examples of β-mixing sequences that fit naturally in a learning scenario are certain Markov processes and hidden Markov models [28]. For instance it is shown in [28] (Section 3.5) that a large class of Markov chains generated by noise inputs with finite variance is β-mixing. Moreover, if a Markov chain has a given mixing property, then it is shared by the corresponding hidden Markov model (see [28], Theorem 3.12). In practice, if the mixing properties are unknown, they need to be estimated. Although it is difficult to find them in general, there exist simple methods to determine the mixing rates for various classes of random processes (e.g., Gaussian, Markov, ARMA, ARCH, GARCH). Hence the assumption of a known mixing rate is reasonable and has been adopted by many studies (see [11], [12], [21], [23], [24], [30]).

B. Classification With Stationary β-Mixing Training Data

In the standard binary classification problem, the training data consist of a sequence $S_n = \langle (X_1, Y_1), \ldots, (X_n, Y_n) \rangle$, where X_k belongs to some measurable space $\mathcal{X}$, and Y_k is in $\{-1, 1\}$. Using S_n, a classifier $h_n : \mathcal{X} \to \{-1, 1\}$ is built to predict the label Y of an unlabeled observation X. Traditionally, the samples are assumed to be i.i.d. and to our knowledge, this assumption is made by all previous studies on boosting consistency (see [5], [14], [19], [20], [32], [33]). In this paper, we no longer require the sampling to be i.i.d. but instead allow it to be stationary β-mixing. More precisely, let $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{-1, +1\}$. Let $W_i = (X_i, Y_i)$.

We suppose that $\underline{W} = (W_i)_{i\geq 1}$ is a strictly stationary sequence of random variables, each having the same distribution P on $\mathcal{D}$ and that $\underline{W}$ is β-mixing (see Definition 2). This setup is in line with [12].

When studying a learning algorithm, we would like to evaluate the performance of the resulting classifier on a new observation. This is accomplished by looking at the generalization error, which is the probability of misclassifying a new example. We assume that the test sample (X, Y) is such that (X, Y) is independent of S_n but with the same marginal. The generalization error of a classifier h_n built using the training data S_n given by

$$L(h_n) = \mathbb{P}\{h_n(X) \neq Y | S_n\}$$

is compared to the minimum possible probability of error: the Bayes error. The Bayes error satisfies

$$L^* = \inf_h L(h) = \mathbb{E}\{\min(\eta(X), 1 - \eta(X))\},$$

with $\eta(X) = \mathbb{P}(Y = 1|X)$, and the corresponding Bayes classifier is

$$h^*(X) = \mathbf{I}_{[\eta(X)>1/2]} - \mathbf{I}_{[\eta(X\leq 1/2)]}.$$

We recall the definition of consistency.

Definition 3 ([6], Definition 6.1): A classification rule is consistent for a certain distribution P if $\mathbb{E}L(h_n) = \mathbb{P}\{h_n(X) \neq Y\} \to L^*$ as $n \to \infty$. It is strongly Bayes-risk consistent if $\lim_{n\to\infty} L(h_n) = L^*$ almost surely.

III. Statistical Convergence and Consistency Results

A. Regularized Boosting

We adopt the framework of [19] which we now recall. Let $\mathcal{H}$ denote the class of base classifiers $h : \mathcal{X} \to \{-1, 1\}$, which usually consists of simple rules (for instance decision stumps). This class is required to have finite VC-dimension. Call $\mathcal{F}$, the class of functions $f : \mathcal{X} \to [-1, 1]$ obtained as convex combinations of the classifiers in $\mathcal{H}$:

$$\mathcal{F} = \Big\{ f(X) = \sum_{j=1}^{t} \alpha_j h_j(X) : t \in \mathbb{N}, \alpha_1, \ldots, \alpha_t \geq 0, \sum_{j=1}^{t} \alpha_j = 1, h_1, \ldots, h_t \in \mathcal{H} \Big\}. \quad (1)$$

Each $f_n \in \mathcal{F}$ defines a classifier $h_{f_n} = \operatorname{sign}(f_n)$ and for simplicity the generalization error $L(h_{f_n})$ is denoted by $L(f_n)$. Then the training error is denoted by

$$L_n(f_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{[h_{f_n}(X_i)\neq Y_i]}.$$

Define $Z(f) = -f(X)Y$ and $Z_i(f) = -f(X_i)Y_i$. Instead of minimizing the indicator of misclassification ($\mathbf{I}_{[-f(X)Y\geq 0]}$), boosting methods are shown to effectively minimize a smooth convex cost function of $Z(f)$. For instance, AdaBoost is based on the exponential function. In addition to the exponential loss, we consider here all cost functions that satisfy the following. Define a positive, differentiable, strictly increasing, and strictly convex function $\phi : \mathbb{R} \to \mathbb{R}^+$ and assume that $\phi(0) = 1$ and that $\lim_{x\to-\infty} \phi(x) = 0$. The corresponding cost function and empirical cost function are respectively

$$C(f) = \mathbb{E}\phi(Z(f)) \text{ and } C_n(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(Z_i(f)).$$

Note that $L(f) \leq C(f)$, since $\mathbf{I}_{[x\geq 0]} \leq \phi(x)$.

The iterative aspect of boosting methods is ignored to consider only their performing an (approximate) minimization of the empirical cost function or, as we shall see, a series of cost functions. To avoid overfitting, the following regularization procedure is developed for the choice of the cost functions. Define ϕ_λ such that $\forall\lambda > 0\ \phi_\lambda(x) = \phi(\lambda x)$. The corresponding empirical and expected cost functions become

$$C_n^\lambda(f) = \frac{1}{n}\sum_{i=1}^{n} \phi_\lambda(Z_i(f)) \text{ and } C^\lambda(f) = \mathbb{E}\phi_\lambda(Z(f)).$$

Then, for sample size n and regularization parameter λ, the regularized boosting methods considered here output the composite classifier

$$h_n = \operatorname{sign}\left(\hat{f}_n^\lambda\right),$$

where $\hat{f}_n^\lambda$ approximately minimizes $C_n^\lambda(f)$ over $f \in \mathcal{F}$, i.e., $\hat{f}_n^\lambda$ is such that

$$C_n^\lambda\left(\hat{f}_n^\lambda\right) \leq \inf_{f\in\mathcal{C}} C_n^\lambda(f) + \epsilon_n,$$

with $\epsilon_n \to 0$ as $n \to \infty$. The minimization of a series of cost functions C^λ over the convex hull of $\mathcal{H}$ is then analyzed.

B. Statistical Convergence

The nature of the sampling intervenes in the following two lemmas that relate the empirical cost $C_n^\lambda(f)$ and true cost $C^\lambda(f)$.

Lemma 1: Suppose that for any n, the training data $(X_1, Y_1), \ldots (X_n, Y_n)$ comes from a stationary algebraically β-mixing sequence with β-mixing coefficients $\beta(m)$ satisfying $\beta(m) = O\left(m^{-r_\beta}\right)$, $m \in \mathbb{N}$ and r_β a positive constant. Then for any $\lambda > 0$ and $b \in [0, 1)$,

$$\mathbb{E}\sup_{f\in\mathcal{F}} |C^\lambda(f) - C_n^\lambda(f)| \leq 4\lambda\phi'(\lambda)\frac{c_1}{n^{(1-b)/2}} + 2\phi(\lambda)\left(\frac{1}{n^{b(1+r_\beta)-1}} + \frac{2}{n^{1-b}}\right) \quad (2)$$

for some constant c_1.

Lemma 2: Let the training data be as in Lemma 1. For any $b \in [0, 1)$, $\lambda > 0$, and $\alpha \in (0, 1 - b)$, let $\epsilon_n = 3(2c_1 + n^{\alpha/2})\lambda\phi'(\lambda)/n^{(1-b)/2}$, where c_1 is as in Lemma 1. Then

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |C^\lambda(f) - C_n^\lambda(f)| > \epsilon_n\right) \leq \exp(-4c_2 n^\alpha) + O(n^{1-b(r_\beta+1)}) \quad (3)$$

for some constant c_2.

The constants c_1 and c_2 in the above lemmas are given in the proofs of Lemma 1 (Appendix B) and Lemma 2 (Appendix C) respectively.

Note that having i.i.d. observations translates into r_β being infinite. Hence by setting $r_\beta \to \infty$, $b = 0$, and $\alpha \in (0,1)$ in Lemma 1 and Lemma 2 we obtain convergence bounds for the case where the observations are independent. Thus, in Lemma 2 for instance, the term $O(n^{1-b(r_\beta+1)})$ can be intuitively viewed as a penalty we pay when considering β-mixing sequences.

C. Main Consistency Result

The following summarizes the assumptions that are made to prove consistency.

Assumptions:

A1- Properties of the sample sequence: The samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ *are assumed to come from a stationary algebraically* β*-mixing sequence with* β*-mixing coefficients* $\beta_{X,Y}(n) = O\left(n^{-r_\beta}\right)$, r_β *being a positive constant.*

A2- Properties of the cost function: ϕ *is assumed to be differentiable, strictly convex, strictly increasing and such that* $\phi(0) = 1$ *and* $\lim_{x\to-\infty}\phi(x) = 0$.

A3- Properties of the base hypothesis space: $\mathcal{H}$ *has finite VC dimension. The distribution of* (X, Y) *and the class* $\mathcal{H}$ *are such that* $\lim_{\lambda\to\infty}\inf_{f\in\lambda\mathcal{F}} C(f) = C^*$, *where* $\lambda\mathcal{F} = \{\lambda f : f \in \mathcal{F}\}$ *and* $C^* = \inf C(f)$ *over all measurable functions* $f : \mathcal{X} \to \mathbb{R}$.

A4- Properties of the smoothing parameter: We assume that $\lambda_1, \lambda_2, \ldots$ *is a sequence of positive numbers satisfying* $\lambda_n \to \infty$ *as* $n \to \infty$, *and that there exists a constant* c *such that* $c \in \left(\frac{1}{1+r_\beta}, 1\right)$ *if* $r_\beta \le 1$ *and* $c \in \left(\frac{2}{1+r_\beta}, 1\right)$ *if* $r_\beta > 1$, *and such that* $\lambda_n\phi'(\lambda_n)/n^{(1-c)/2} \to 0$ *as* $n \to \infty$.

The following main result states that under Assumptions A1-A4, regularized boosting for stationary β-mixing sequences is Bayes-risk consistent if the mixing rate $r_\beta \le 1$, and strongly Bayes-risk consistent if $r_\beta > 1$.

Theorem 1 (Consistency of regularized boosting for stationary $\boldsymbol{\beta}$*-mixing sequences): Denote by* $\hat{f}_n^\lambda$ *the function in* $\mathcal{F}$ *resulting from boosting with regularization parameter* λ *and sample size* n, *so that* $\hat{f}_n^\lambda$ *approximately minimizes* $C_n^\lambda(f)$. *That is,* $\hat{f}_n^\lambda$ *is such that*

$$C_n^\lambda(\hat{f}_n^\lambda) \le \inf_{f\in\mathcal{F}} C_n^\lambda(f) + \epsilon_n = \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \phi_\lambda(Z_i(f)) + \epsilon_n,$$

with $\epsilon_n \to 0$ *as* $n \to \infty$.

Let $f_n = \hat{f}_n^{\lambda_n} \in \mathcal{F}$ *and* $h_{f_n} = sign(f_n)$. *Suppose Assumptions A1-A4 are satisfied. If* $r_\beta \le 1$ *then* $\lim_{n\to\infty} L\left(h_{f_n}\right) = L^*$ *inprobability and* h_{f_n} *is Bayes-risk consistent. If* $r_\beta > 1$ *then* $\lim_{n\to\infty} L\left(h_{f_n}\right) = L^*$ *almost surely and* h_{f_n} *is strongly Bayes-risk consistent.*

D. Consistency for Data Dependent Regularization

Notice that in Section III-C, the minimization of the cost functions $C_n^{\lambda_n}$ is carried out for a sequence of smoothing parameters $\lambda_1, \lambda_2 \ldots$ that is chosen in advance. We show that consistency can also be established if the smoothing parameter is chosen in a data-dependent manner. Specifically, for any integer n and regularization parameter λ, consider the following penalized cost function

$$\tilde{C}_n^\lambda(f) = C_n^\lambda(f) + \frac{3(2c_1 + n^{\alpha/2})\lambda\phi'(\lambda)}{n^{(1-b)/2}}, \tag{4}$$

where α and c_1 are as in Lemma 2, and for any $b \in \left(\frac{1}{1+r_\beta}, 1\right)$ if $r_\beta \le 1$ or for any $b \in \left(\frac{2}{1+r_\beta}, 1\right)$ if $r_\beta > 1$. Define the set K_n by

$$K_n = \left\{1, \ldots, \lfloor n^{b'} \rfloor\right\}, \tag{5}$$

where $b' \in \left(0, b(1+r_\beta) - 1\right)$ if $r_\beta \le 1$ or $b' \in \left(0, b(1+r_\beta) - 2\right)$ if $r_\beta > 1$. For any sequence of positive numbers $\lambda_1, \lambda_2, \ldots$, with $\lambda_n \to \infty$, let

$$k^* = \arg\min_{k\in K_n} \tilde{C}_n^{\lambda_k}\left(\hat{f}_n^{\lambda_k}\right),$$

where

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

i.e., $\hat{f}_n^{\lambda_k}$ results from regularized boosting with sample size n and regularization parameter λ_k. Then, for the choice of the regularization parameter λ_{k^*}, we obtain the following consistency result.

Theorem 2 (Consistency of data-dependent regularized boosting for stationary $\boldsymbol{\beta}$*-mixing sequences): For any integer* n *and regularization parameter* λ, *consider the penalized cost function* $\tilde{C}_n^\lambda(f)$ *given by (4). Suppose that Assumptions A1-A3 hold. For any sequence of positive numbers* $\lambda_1, \lambda_2, \ldots$, *with* $\lambda_n \to \infty$, *let*

$$k^* = \arg\min_{k\in K_n} \tilde{C}_n^{\lambda_k}\left(\hat{f}_n^{\lambda_k}\right),$$

where K_n *is given by (5) and*

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

Further let

$$f_n = \hat{f}_n^{\lambda_{k^*}} \quad \text{and} \quad h_{f_n} = sign(f_n).$$

If $r_\beta \le 1$, *then* $\lim_{n\to\infty} L\left(h_{f_n}\right) = L^*$ *inprobability and* h_{f_n} *is Bayes-risk consistent. If* $r_\beta > 1$, *then* $\lim_{n\to\infty} L\left(h_{f_n}\right) = L^*$ *almost surely and* h_{f_n} *is strongly Bayes-risk consistent.*

E. Discussion

We first comment on various assumptions. Cost functions satisfying Assumption A2 include the exponential function, which is used by AdaBoost, and the logit function $\log_2(1 + e^x)$. This assumption can be relaxed to consider functions that are "classification-calibrated" as characterized by Bartlett et al. [1].

The second part of Assumption A3 guarantees that the class of basis functions is rich enough so that the measurable function on R^d minimizing $C(f)$ can be approximated by functions in $\lambda\mathcal{F}$ as $\lambda \to \infty$. It is well known that the class of indicators of all rectangles, and the class of indictors of hyperplanes defined by half spaces satisfy this denseness assumption. It was shown in [4] that the class of ± 1 trees with a number of terminal nodes strictly greater than d satisfies the

required condition. Another example for which the assumption holds is the class of basis functions

$$\mathcal{H} = \left\{ \sigma\left(\alpha_0 + \alpha^T X\right) : \alpha \in R^d, \alpha_0 \in R \right\},$$

where the activation function σ is monotone, bounded and continuous (e.g., the sigmoid $\sigma(v) = 1/(1+e^{-v})$) (see [18]). This class generates two-level neural-networks in R^d.

We now examine the impact of the weakly dependent sampling. In Assumption A4, the nature of the sampling leads to a generalization of the condition on the growth of $\lambda_n \phi'(\lambda_n)$ present in the i.i.d. setting [19]. More precisely, the nature of the sampling manifests through the parameter c, which is limited by r_β. The weaker the dependence (i.e. the larger r_β), the faster λ_n is allowed to grow. The assumption that r_β be known is stringent but necessary (for instance this assumption is widely made in the field of time series analysis). However, if r_β is unknown, it can be determined for various classes of processes as mentioned in Section II-A. A future extension of this paper would be to provide adaptivity with respect to r_β. Note that Theorem 2 does not require Assumption A4 to hold as the sequence of parameters $\lambda_1, \lambda_2, \ldots$ can be any divergent sequence. The nature of sampling is reflected however in the penalized cost function $\tilde{C}_n^{\lambda_k}(\cdot)$ in which the constant b is required to belong to $\left(\frac{1}{1+r_\beta}, 1\right)$ if $r_\beta \leq 1$ or to $\left(\frac{2}{1+r_\beta}, 1\right)$ if $r_\beta > 1$. Also notice that the integer k^* is restricted to belong to $\left\{1, \ldots, n^{b'}\right\}$, where $b' \in \left(0, b(1+r_\beta)\right)$ if $r_\beta \leq 1$ or $b' \in \left(0, b(2+r_\beta)\right)$ if $r_\beta > 1$, as opposed to the i.i.d. case where k^* can be any positive integer. This is due to the penalty term $O\left(n^{1-b(r_\beta+1)}\right)$ in Lemma 2, which results from the fact that we are considering β-mixing sequences. Details are provided in Appendix E.

In Theorem 1 and Theorem 2, if the dependence of the mixing sequence is weak enough ($r_\beta > 1$) we obtain strong Bayes consistency results, while if the dependence is stronger ($r_\beta \leq 1$) we arrive only at Bayes consistency results. This is a consequence of the penalty term $O\left(n^{1-b(r_\beta+1)}\right)$ in Lemma 2.

To conclude this section we comment on the applicability of our analysis to various regularization strategies in boosting. In this paper we focus on boosting procedures where regularization is performed by restricting the 1-norm of the base classifiers. Our analysis, however, is also applicable to regularization via stopping rules as in Bartlett & Traskin [2]. Specifically the key function classes to consider in this case are the class of t combinations of classifiers:

$$\mathcal{F}^t = \left\{ f(X) = \sum_{j=1}^{t} \alpha_j h_j(X), \alpha_i \in \mathbb{R}, h_1, \ldots, h_t \in \mathcal{H} \right\}.$$

and its"truncated" version

$$\pi_\xi \circ \mathcal{F}^t = \left\{ \tilde{f}(X) = \pi_\xi(f(X)), f \in \mathcal{F}^t, \right.$$
$$\left. \pi_\xi(x) = \text{sign(x)} \min(|x|, \xi) \right\}. \quad (6)$$

For the latter class (6), the uniform convergence results of Bartlett & Traskin [2] can be generalized to mixing sequences in a straightforward manner, by following our proof techniques of Lemma 1 and Lemma 2. Indeed the *i.i.d.* analysis in Bartlett & Traskin [2] involves the application of symmetrization, "contraction principle," maximum inequality and Mc Diarmid's Bounded Difference inequality, in a similar fashion as for the 1-norm regularized version considered here. Counterparts of our statistical convergence and consistency results can then be obtained, in which stopping points t_n are key parameters (instead of regularization paramters λ_n) whose growth with the sample size n are also affected by the nature of sampling.

IV. Conclusion

In this paper, we considered the consistency of regularized boosting methods, where the samples are weakly dependent. Under the assumption that the sequences are stationary algebraically β-mixing, we have proven the Bayes consistency for the case where the regularization parameters are chosen before data observation. In the consistency result, the mixing nature of sampling leads to the generalization of the condition on the growth of the regularization parameter stated for the i.i.d. setting. The weaker the sample dependence, the faster the regularization parameter can grow as the sample size increases. In addition, we have also shown the Bayes consistency for data dependent choice of the regularization parameters.

Our results demonstrate that the consistency of regularized boosting methods is preserved under the much less restrictive assumption of β-mixing. Further, our study is of wide applicability as many processes encountered in learning scenarios are β-mixing.

Note that both regularization strategies assume knowledge of the mixing rate. If unknown, it can be determined for various classes of random processes. A direction for future work would consist, however, in developing a strategy that provides adaptivity with respect to the mixing rate.

Appendix

A. Preparation to the Proofs: the Blocking Technique

The key issue resides in upper bounding

$$\sup_{f \in \mathcal{F}} \left| C_n^\lambda(f) - C^\lambda(f) \right|$$
$$= \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \phi\left(-\lambda f(X_i) Y_i\right) - \mathbb{E}\phi\left(-\lambda f(X_1) Y_1\right) \right|, \quad (7)$$

where $\mathcal{F}$ is given by (1). Let $W = (X, Y)$, $W_i = (X_i, Y_i)$. Define the function $g_{\lambda,f}$ by

$$g_{\lambda,f}(W) = g_{\lambda,f}(X, Y) = \phi\left(-\lambda f(X) Y\right)$$

and the class $\mathcal{G}_\lambda$ by $\mathcal{G}_\lambda = \left\{ g_{\lambda,f} : f \in \mathcal{F} \right\}$. Then (7) can be rewritten as

$$\sup_{f \in \mathcal{F}} \left| C_n^\lambda(f) - C^\lambda(f) \right| = \sup_{g_{\lambda,f} \in \mathcal{G}_\lambda} \left| n^{-1} \sum_{i=1}^{n} g_{\lambda,f}(W_i) - \mathbb{E} g_{\lambda,f}(W_1) \right|.$$

Note that the class $\mathcal{G}_\lambda$ is uniformly bounded by $\phi(\lambda)$. Also, if $\mathcal{H}$ is a class of measurable functions, then $\mathcal{G}_\lambda$ is also a class of measurable functions, by measurability of $\mathcal{F}$.

As the W_i's are not i.i.d, we propose to use the blocking technique developed by [29], [30] to construct i.i.d blocks

of observations which are close in distribution to the original sequence $W_1, \ldots, W_n$. This enables us to work on the sequence of independent blocks instead of the original sequence. We use the same notation as in [30]. The protocol is the following. Let (b_n, μ_n) be a pair of integers, such that

$$(n - 2b_n) \leq 2b_n\mu_n \leq n. \tag{8}$$

For example, a valid choice is $\mu_n = \left[\frac{1}{2}n^{\frac{s}{1+s}}\right]$ and $b_n = \left[n^{\frac{1}{1+s}}\right]$, where $0 < s < r_\beta$. Divide the segment $W_1 = (X_1, Y_1), \ldots, W_n = (X_n, Y_n)$ of the mixing sequence into $2\mu_n$ blocks of size b_n, followed by a remaining block (of size at most $2b_n$). Consider the odd blocks only. If their size b_n is large enough, the dependence between them is weak, since two odd blocks are separated by an even block of the same size b_n. Therefore, the odd blocks can be approximated by a sequence of independent blocks with the same within-block structure. The same holds if we consider the even blocks. Let $\left(\tilde{W}_1, \ldots, \tilde{W}_{b_n}\right), \left(\tilde{W}_{b_n+1}, \ldots, \tilde{W}_{2b_n}\right), \ldots, \left(\tilde{W}_{(2\mu_n-1)b_n}, \ldots, \tilde{W}_{2\mu_n b_n}\right)$ be independent blocks such that $\left(\tilde{W}_{jb_n+1}, \ldots, \tilde{W}_{(j+1)b_n}\right)$ and $\left(W_{jb_n+1}, \ldots, W_{(j+1)b_n}\right)$ have the same distribution, for $j = 0, \ldots, \mu_n - 1$.

For $j = 1, \ldots, 2\mu_n$, and any $g \in \mathcal{G}_\lambda$, define

$$\tilde{\xi}_{j,g} := \sum_{i=(j-1)b_n+1}^{jb_n} g\left(\tilde{W}_i\right) - b_n \mathbb{E} g\left(\tilde{W}_1\right),$$

$$\xi_{j,g} := \sum_{i=(j-1)b_n+1}^{jb_n} g\left(W_i\right) - b_n \mathbb{E} g\left(W_1\right).$$

Let $\mathcal{O}_{\mu_n} = \{1, 3, \ldots, 2\mu_n - 1\}$ and $\mathcal{E}_{\mu_n} = \{2, 4, \ldots, 2\mu_n\}$. Define $Z_{i,j}(f)$ as

$$Z_{i,j}(f) := -f\left(\tilde{X}_{(2j-2)b_n+i}\right) \cdot \tilde{Y}_{(2j-2)b_n+i},$$

where $\tilde{X}_k$ and $\tilde{Y}_k$ are respectively the first and second coordinate of the vector $\tilde{W}_k$. These correspond to the $Z_k(f) = -f(X_k) Y_k$ for k in the odd blocks $1, ..., b_n, 2b_n + 1, ..., 3b_n, ...$

B. Proof of Lemma 1

The first of the following steps allows us to work with independent blocks instead of the original β-mixing sequence.

Step 1: Working with Independent Blocks. We show that

$$\mathbb{E} \sup_{g \in \mathcal{G}_\lambda} \left| \frac{1}{n} \sum_{i=1}^{n} g\left(W_i\right) - \mathbb{E} g\left(W_1\right) \right| \leq 2\mathbb{E} \sup_{g \in \mathcal{G}_\lambda} \left| \frac{1}{n} \sum_{j \in \mathcal{O}_{\mu_n}} \tilde{\xi}_{j,g} \right| + \phi(\lambda) \left(\mu_n \beta_W(b_n) + \frac{2b_n}{n} \right). \tag{9}$$

Proof: Without loss of generality, assume that $\mathbb{E} g(W_1) = \mathbb{E} g\left(\tilde{W}_1\right) = 0$. Then,

$$\mathbb{E} \sup_g \left| \frac{1}{n} \sum_{i=1}^{n} g\left(W_i\right) \right| = \mathbb{E} \sup_g \left| \frac{1}{n} \left(\sum_{\mathcal{O}_{\mu_n}} \xi_{j,g} + \sum_{\mathcal{E}_{\mu_n}} \xi_{j,g} + R \right) \right|,$$

where R is the remainder term consisting of a sum of at most $2b_n$ terms. Noting that $\forall g \in \mathcal{G}_\lambda$, $|g| \leq \phi(\lambda)$, it follows that

$$\mathbb{E} \sup_g \left| \frac{1}{n} \sum_{i=1}^{n} g\left(W_i\right) \right| \leq \mathbb{E} \left(\sup_g \left| \frac{1}{n} \sum_{\mathcal{O}_{\mu_n}} \xi_{j,g} \right| \right) + \mathbb{E} \left(\sup_g \left| \frac{1}{n} \sum_{\mathcal{E}_{\mu_n}} \xi_{j,g} \right| \right) + \frac{\phi(\lambda)(2b_n)}{n}. \tag{10}$$

We use the following intermediary lemma.

Lemma 3 ([31], Lemma 4.1): Denote by $\mathbf{Q}$ *the distribution of the odd blocks*

$$\left(W_1, \ldots, W_{b_n}; W_{2b_n+1}, \ldots, W_{3b_n}; \ldots\right).$$

Similarly denote by $\widetilde{\mathbf{Q}}$ *the distribution of the corresponding odd blocks*

$$\left(\tilde{W}_1, \ldots, \tilde{W}_{b_n}; \tilde{W}_{2b_n+1}, \ldots, \tilde{W}_{3b_n}; \ldots\right).$$

For any measurable function h *on* $\mathbb{R}^{b_n\mu_n}$ *with bound* H*, we have*

$$\left| \mathbb{E}_{\mathbf{Q}} h\left(W_1, \ldots\right) - \mathbb{E}_{\widetilde{\mathbf{Q}}} h\left(\tilde{W}_1, \ldots\right) \right| \leq H\left(\mu_n - 1\right) \beta_W\left(b_n\right).$$

The same result holds for the even blocks $(W_{b_n+1}, \ldots, W_{2b_n}; W_{3b_n+1}, \ldots, W_{4b_n}; \ldots)$.

Consider the function h on $\mathbb{R}^{b_n\mu_n}$ such that

$$h(x_1, \ldots, x_{b_n\mu_n}) = \sup_g \left| \frac{1}{n} \left(\sum_{i=1}^{b_n\mu_n} g(x_i) - b_n\mu_n \mathbb{E} g(x_1) \right) \right|.$$

We then obtain $h\left(W_1, \ldots, W_{b_n}, W_{2b_n+1}, \ldots, W_{3b_n}, \ldots\right) = \sup_g \left| \frac{1}{n} \sum_{\mathcal{O}_{\mu_n}} \xi_{j,g} \right|$, and $h(W_{b_n+1}, \ldots, W_{2b_n}, W_{3b_n+1}, \ldots, W_{4b_n} \ldots) = \sup_g \left| \frac{1}{n} \sum_{\mathcal{E}_{\mu_n}} \xi_{j,g} \right|$. Applying the lemma and noting that $H = \phi(\lambda)/2$, we have, using (10),

$$\mathbb{E} \sup_g \left| \frac{1}{n} \sum_{i=1}^{n} g\left(W_i\right) \right| \leq \mathbb{E} \sup_g \left| \frac{1}{n} \sum_{\mathcal{O}_{\mu_n}} \tilde{\xi}_{j,g} \right| + \frac{\phi(\lambda)}{2} \mu_n \beta_W\left(b_n\right) + \mathbb{E} \sup_g \left| \frac{1}{n} \sum_{\mathcal{E}_{\mu_n}} \tilde{\xi}_{j,g} \right| + \frac{\phi(\lambda)}{2} \mu_n \beta_W\left(b_n\right) + \frac{\phi(\lambda)(2b_n)}{n}.$$

As the $\tilde{\xi}_{j,g}$'s of the odd blocks are independent, the $\tilde{\xi}_{j,g}$'s of the even blocks are independent, and $\tilde{\xi}_{j,g}$'s from odd and even blocks have the same distribution, we have

$$\mathbb{E} \sup_g \left| \frac{1}{n} \sum_{\mathcal{O}_{\mu_n}} \tilde{\xi}_{j,g} \right| = \mathbb{E} \sup_g \left| \frac{1}{n} \sum_{\mathcal{E}_{\mu_n}} \tilde{\xi}_{j,g} \right|.$$

We then obtain (9). ∎

The odd blocks $\tilde{\xi}_{j,g}$'s being independent, we can then use standard symmetrization and contraction principle similarly as in [19]. We then obtain:

Step 2: Symmetrization and Contraction Principle

$$\mathbb{E} \sup_{g \in \mathcal{G}_\lambda} \left| \frac{1}{n} \sum_{j \in \mathcal{O}_{\mu_n}} \tilde{\xi}_{j,g} \right| \leq 2 \cdot b_n \lambda \phi'(\lambda) \, \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j=1}^{\mu_n} s_j Z_{1,j}(f) \right|, \tag{11}$$

where (s_i) is a Rademacher sequence, i.e. a sequence of independent random variables taking the values ± 1 with probability $1/2$.

Step 3: Maximal Inequality. We show that there exists a constant $c_1 > 0$ such that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j=1}^{\mu_n} s_j Z_{1,j}(f) \right| \leq \frac{c_1 \sqrt{\mu_n}}{n}. \tag{12}$$

Proof: Denote $(h_1, \ldots, h_N)$ by h_1^N. One can write

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j=1}^{\mu_n} s_j Z_{1,j}(f) \right| = \frac{1}{n} \mathbb{E} \sup_{N \geq 1} \sup_{h_1^N \in \mathcal{H}^N} \sup_{\substack{\alpha_1, \ldots, \alpha_N \\ \alpha_i \geq 0, \sum \alpha_i = 1}} \left| \sum_{j=1}^{\mu_n} \sum_{k=1}^{N} \alpha_k s_j \tilde{Y}_{(2j-2)b_n+1} \cdot h_k\left(\tilde{X}_{(2j-2)b_n+1}\right) \right|.$$

Since $\tilde{Y}_{(2j-2)b_n+1}$ and $\tilde{Y}_{(2j'-2)b_n+1}$ are i.i.d. for all $j \neq j'$ (they come from different blocks), and (s_j) is a Rademacher sequence, $\left(s_j \tilde{Y}_{(2j-2)b_n+1} h_k\left(\tilde{X}_{(2j-2)b_n+1}\right)\right)_{j=1,\ldots,\mu_n}$ has the same distribution as $\left(s_j h_k\left(\tilde{X}_{(2j-2)b_n+1}\right)\right)_{j=1,\ldots,\mu_n}$. Hence

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j=1}^{\mu_n} s_j Z_{1,j}(f) \right| = \frac{1}{n} \mathbb{E} \sup_{N \geq 1} \sup_{h_1^N \in \mathcal{H}^N} \sup_{\substack{\alpha_1, \ldots, \alpha_N \\ \alpha_i \geq 0, \sum \alpha_i = 1}} \left| \sum_{j=1}^{\mu_n} \sum_{k=1}^{N} s_j \alpha_k h_k\left(\tilde{X}_{(2j-2)b_n+1}\right) \right|.$$

By the same argument as used in ([19], p.53) on the maximum of a linear function over a convex polygon, the supremum is achieved when $\alpha_k = 1$ for some k. We arrive at

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j=1}^{\mu_n} s_j Z_{1,j}(f) \right| = \frac{1}{n} \mathbb{E} \sup_{h \in \mathcal{H}} \left| \sum_{j=1}^{\mu_n} s_j h\left(\tilde{X}_{(2j-2)b_n+1}\right) \right|. \tag{13}$$

To upper bound the right-hand side of (13), we use the the following inequality on the Rademacher complexity (see [27], Corollary 2.2.8):

For a class of functions $\mathcal{G} = \{g : \mathcal{X}\}$ if $0 \in \mathcal{G}$ then we have

$$\mathbb{E} \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty (\log N_2(\epsilon, n, \mathcal{G}))^{1/2} d\epsilon, \right. \tag{14}$$

where c' is a universal constant and $N_2(\epsilon, n, \mathcal{G})$ is the uniform $\mathcal{L}_2$ covering number of $\mathcal{G}$.

Noting that for all $j \neq j'$, $h(\tilde{X}_{(2j-2)b_n+1})$ and $h(\tilde{X}_{(2j'-2)b_n+1})$ are i.i.d. we can apply the inequality of (14) to upper bound the right-hand side of (13) and obtain

$$\begin{aligned} &\frac{1}{n} \mathbb{E} \sup_{h \in \mathcal{H}} \left| \sum_{j=1}^{\mu_n} s_j h\left(\tilde{X}_{(2j-2)b_n+1}\right) \right| \\ &\quad \leq \frac{1}{n} \mathbb{E} \sup_{h \in \mathcal{H} \cup \{0\}} \left| \sum_{j=1}^{\mu_n} s_j h\left(\tilde{X}_{(2j-2)b_n+1}\right) \right| \\ &\quad \leq \frac{c' \sqrt{\mu_n}}{n} \int_0^\infty (\log N_2(\epsilon, n, \mathcal{H} \cup \{0\}))^{1/2} d\epsilon, \end{aligned}$$

where c' is a universal constant and $N_2(\epsilon, n, \mathcal{H} \cup \{0\})$ is the uniform $\mathcal{L}_2$ covering number of $\mathcal{H} \cup \{0\}$. As $\mathcal{H}$ has finite VC-dimension (by Assumption A3), there exists a positive constant w such that $N_2(\epsilon, n, \mathcal{H} \cup \{0\}) = O(\epsilon^{-w})$ (by [27], Theorem 2.6.7). Hence

$$\int_0^\infty (\log N_2(\epsilon, n, \mathcal{H} \cup \{0\}))^{1/2} d\epsilon < \infty.$$

and (12) follows. ∎

Step 5: Establishing (2). Combining (9),(11), and (12), we have

$$\begin{aligned} &\mathbb{E} \sup_{g \in \mathcal{G}_\lambda} \left| \frac{1}{n} \sum_{i=1}^{n} g(W_i) - \mathbb{E} g(W_1) \right| \\ &\quad \leq 4 b_n \lambda \phi'(\lambda) \frac{c_1 \sqrt{\mu_n}}{n} + \phi(\lambda) \left(\mu_n \beta_W(b_n) + \frac{2 b_n}{n} \right). \end{aligned}$$

Take $b_n = n^b$, with $0 \leq b < 1$. By (8), we obtain $\mu_n \leq n^{1-b}/2$. Besides, as we assumed that the sequence $\underline{W}$ is algebraically β-mixing (see Definition 2), $\beta_W(n) = O\left(n^{-r_\beta}\right)$. Then $\mu_n \beta_W(b_n) = O\left(n^{1-b(1+r_\beta)}\right)$, and we arrive at (2). ∎

C. Proof of Lemma 2

Step 1: Working with Independent Blocks and Symmetrization. For any $b \in [0, 1)$, $\alpha \in (0, 1-b)$, let

$$\epsilon_n = \frac{3(2c_1 + n^{\alpha/2}) \lambda \phi'(\lambda)}{n^{(1-b)/2}}. \tag{15}$$

We show

$$\begin{aligned} &\mathbb{P}\left(\sup_{g \in \mathcal{G}_\lambda} \left| \frac{1}{n} \sum_{i=1}^{n} g(W_i) - \mathbb{E} g(W_1) \right| > \epsilon_n \right) \\ &\quad \leq 2 \mathbb{P}\left(\sup_{g \in \mathcal{G}_\lambda} \left| \frac{1}{n} \sum_{j \in \mathcal{O}_{\mu_n}} \tilde{\xi}_{j,g} \right| > \epsilon_n / 3 \right) + O\left(n^{1-b(1+r_\beta)}\right) \end{aligned} \tag{16}$$

Proof: Without loss of generality, assume again that $\mathbb{E} g(W_1) = \mathbb{E} g\left(\tilde{W}_1\right) = 0$. Then,

$$\frac{1}{n} \sum_{i=1}^{n} g(W_i) = \frac{1}{n} \sum_{\mathcal{O}_{\mu_n}} \xi_{j,g} + \frac{1}{n} \sum_{\mathcal{E}_{\mu_n}} \xi_{j,g} + R,$$

where R is the remainder term consisting of a sum of at most $2b_n$ terms. Set $b_n = n^b$, with $0 \leq b < 1$. Since by Assumption A2, ϕ is strictly convex and differentiable, we have $\forall x, y,\ \phi(y) \geq \phi(x) + \phi'(x)(y - x)$. Recall that, by Assumption A2, $\phi(0) = 1$. Then, by setting $y = 0$ and $x = \lambda$, we obtain $\lambda \phi'(\lambda) \geq \phi(\lambda) - 1$. With ϵ_n as in (15), and since $\lambda \phi'(\lambda) \geq \phi(\lambda) - 1$, we automatically

obtain $\phi(\lambda)b_n = o(n\epsilon_n)$. This implies that R can be made smaller than $\epsilon_n/3$. Notice that $\mu_n\beta_W(b_n) = O(n^{1-b(1+r_\beta)})$ (for the same reasons as in Appendix B). Then, by applying Lemma 3, with h being the indicator function of the events $\left\{\sup_g |n^{-1}\sum_{\mathcal{O}_{\mu_n}}\xi_{j,g}| \geq \epsilon_n/3\right\}$ and $\left\{\sup_g |n^{-1}\sum_{\mathcal{E}_{\mu_n}}\xi_{j,g}| \geq \epsilon_n/3\right\}$ respectively, we conclude. ∎

Step 2: McDiarmid's Bounded Difference Inequality. The $\tilde{\xi}_{j,g}$'s of the odd blocks being independent, we can apply McDiarmid's bounded difference inequality [6] (Theorem 9.2) on the function $\sup_{g\in\mathcal{G}_\lambda}|\frac{1}{n}\sum_{j\in\mathcal{O}_{\mu_n}}\tilde{\xi}_{j,g}|$ which depends on $\tilde{\xi}_{1,g}, \tilde{\xi}_{3,g}\ldots, \tilde{\xi}_{2\mu_n-1,g}$. Then combining (11) and (12) from the proof of Lemma 1, and setting $b_n = n^b$ we obtain that for ϵ_n as in (15), there exists a constant $c_2 > 0$ such that,

$$\mathbb{P}\Big(\sup_{g\in\mathcal{G}_\lambda}\Big|\frac{1}{n}\sum_{j\in\mathcal{O}_{\mu_n}}\tilde{\xi}_{j,g}\Big| > \epsilon_n/3\Big) \leq \exp(-4c_2n^\alpha). \tag{17}$$

Note that $c_2 = (1-1/\phi(\lambda_0))^2$ for λ_0 such that $0 < \lambda_0 < \lambda$. Combining (16) and (17) we obtain (3). ∎

D. Proof of Theorem 1

Let $\bar{f}_\lambda$ be a function in $\mathcal{F}$ minimizing C^λ. With $f_n = \hat{f}_n^{\lambda_n}$, we have

$$C(\lambda_n f_n) - C^* = \Big(C^{\lambda_n}(\hat{f}_n^{\lambda_n}) - C^{\lambda_n}(\bar{f}_{\lambda_n})\Big) + \Big(\inf_{f\in\lambda_n\mathcal{F}} C(f) - C^*\Big).$$

Since $\lambda_n \to \infty$, the second term on the right-hand side converges to zero by Assumption A3. By [6], Lemma 8.2, we have $C^{\lambda_n}(\hat{f}_n^{\lambda_n}) - C^{\lambda_n}\left(\bar{f}_{\lambda_n}\right) \leq 2\sup_{f\in\mathcal{F}}|C^{\lambda_n}(f) - C_n^{\lambda_n}(f)|$. If $r_\beta \leq 1$, then by Lemma 2, $\sup_{f\in\mathcal{F}}|C^{\lambda_n}(f) - C_n^{\lambda_n}(f)| \to 0$ in probability if $\lim_{n\to\infty}\lambda_n\phi'(\lambda_n)n^{(\alpha+b-1)/2} = 0$ and $b > 1/(1+r_\beta)$. Hence if Assumption A4 holds, $C(\lambda_n f_n) \to C^*$ in probability.

If $r_\beta > 1$ then $\sup_{f\in\mathcal{F}}|C^{\lambda_n}(f) - C_n^{\lambda_n}(f)| \to 0$ with probability 1 if $\lim_{n\to\infty}\lambda_n\phi'(\lambda_n)n^{(\alpha+b-1)/2} = 0$ and $b > 2/(1+r_\beta)$. Hence if Assumption A4 holds, $C(\lambda_n f_n) \to C^*$ with probability 1. The theorem follows by applying Lemma 5 in [19]. ∎

E. Proof of Theorem 2

Note that Theorem 2 generalizes the result of [19]. However, due to the nature of the sampling, the proof of Theorem 2 differs from the one for i.i.d. sampling as we have here less flexibility with various parameters.

For any n and any λ define $\delta(n)$ and $\epsilon(\lambda, n)$ as

$$\delta(n) = \exp(-4c_2n^\alpha) + O\left(n^{1-b(r_\beta+1)}\right),$$
$$\epsilon(\lambda,n) = \frac{3(2c_1 + n^{\alpha/2})\lambda\phi'(\lambda)}{n^{(1-b)/2}},$$

where α, c_1 and c_2 are as in Lemma 2, and for any $b \in \left(\frac{1}{1+r_\beta}, 1\right)$ if $r_\beta \leq 1$, or for any $b \in \left(\frac{1}{2+r_\beta}, 1\right)$ if $r_\beta > 1$. Define the following empirical cost and true cost minimizers.

$$\hat{f}_n^\lambda = \arg\min_{f\in\mathcal{F}} C_n^\lambda(f) \quad \text{and} \quad \bar{f}^\lambda = \arg\min_{f\in\mathcal{F}} C^\lambda(f).$$

Now, consider the penalized cost function $\tilde{C}_n^\lambda(f) = C_n^\lambda(f) + \epsilon(\lambda, n)$. As before, let $K_n = \left\{1, \ldots, \lfloor n^{b'}\rfloor\right\}$, where $b' \in \big(0, b(1+r_\beta)-1\big)$ if $r_\beta \leq 1$, or where $b' \in \big(0, b(1+r_\beta)-2\big)$ if $r_\beta > 1$. Let $k^* = \arg\min_{k\in K_n}\tilde{C}_n^{\lambda_k}(\hat{f}_n^{\lambda_k})$, and $f_n = \hat{f}_n^{\lambda_{k^*}}$.

Step 1: We first show that

$$\mathbb{P}\left(\min_{k\in K_n} C^{\lambda_k}(f_n) \leq \min_{k\in K_n}\tilde{C}_n^{\lambda_k}(\hat{f}_n^{\lambda_k})\right) \geq 1-\delta'(n), \tag{18}$$

with $\delta'(n) = n^{b'}\delta(n)$.

Proof: By Lemma 2, for any particular k, $\mathbb{P}(\forall f \in \mathcal{F}, C^{\lambda_k}(f) \leq \tilde{C}_n^{\lambda_k}(f)) \geq 1-\delta(n)$. Together with the union bound, this implies that

$$\mathbb{P}\left(\forall k \in K_n, \forall f\in\mathcal{F}, \quad C^{\lambda_k}(f) \leq \tilde{C}_n^{\lambda_k}(f)\right) \geq 1-\delta'(n).$$

Thus we obtain $\mathbb{P}\left(C^{\lambda_{k^*}}(f_n) \leq \tilde{C}_n^{\lambda_{k^*}}(f_n)\right) \geq 1-\delta'(n)$.
We conclude by noting that $\min_{k\in K_n} C^{\lambda_k}(f_n) \leq C^{\lambda_{k^*}}(f_n)$. ∎

Step 2: We then show that with

$$\bar{k} = \arg\min_{k\in K_n}\left(C^{\lambda_k}(\bar{f}^{\lambda_k}) + 2\epsilon(\lambda_k, n)\right),$$
$$\mathbb{P}\left(\tilde{C}_n^{\lambda_{\bar{k}}}(\hat{f}_n^{\lambda_{\bar{k}}}) \leq C^{\lambda_{\bar{k}}}(\bar{f}^{\lambda_{\bar{k}}}) + 2\epsilon(\lambda_{\bar{k}}, n)\right) \geq 1-\delta(n). \tag{19}$$

Proof: By applying Lemma 2, this time to upper bound the empirical cost of choosing $f = \bar{f}^{\lambda_{\bar{k}}}$, we obtain that, $\mathbb{P}\left(C_n^{\lambda_{\bar{k}}}(\bar{f}^{\lambda_{\bar{k}}}) \leq C^{\lambda_{\bar{k}}}(\bar{f}^{\lambda_{\bar{k}}}) + \epsilon(\lambda_{\bar{k}}, n)\right) \geq 1-\delta(n)$. Noting that by definition of $\hat{f}_n^{\lambda_{\bar{k}}}$, we have $C_n^{\lambda_{\bar{k}}}(\hat{f}_n^{\lambda_{\bar{k}}}) \leq C_n^{\lambda_{\bar{k}}}(\bar{f}^{\lambda_{\bar{k}}})$ and thus that $\tilde{C}_n^{\lambda_{\bar{k}}}(\hat{f}_n^{\lambda_{\bar{k}}}) \leq C_n^{\lambda_{\bar{k}}}(\bar{f}^{\lambda_{\bar{k}}}) + \epsilon(\lambda_{\bar{k}}, n)$, we conclude. ∎

Step 3: Proof of Theorem III-D. Notice that (18) implies that

$$\mathbb{P}\left(\min_{k\in K_n} C^{\lambda_k}(f_n) \leq \tilde{C}_n^{\lambda_{\bar{k}}}(\hat{f}_n^{\lambda_{\bar{k}}})\right) \geq 1-\delta'(n).$$

Combining this with (19), we obtain by the union bound that

$$\mathbb{P}\left(\min_{k\in K_n} C^{\lambda_k}(f_n) \leq \min_{k\in K_n}\left(C^{\lambda_k}(\bar{f}^{\lambda_k}) + 2\epsilon(\lambda_k, n)\right)\right) \geq 1-\delta(n)-\delta'(n).$$

In addition, we have by definition of $\bar{f}^{\lambda_k}$ that for all $k \in K_n$, $C^{\lambda_k}(\bar{f}^{\lambda_k}) \leq C^{\lambda_k}(f_n)$. Hence we have $\min_{k\in K_n} C^{\lambda_k}(\bar{f}^{\lambda_k}) \leq \min_{k\in K_n} C^{\lambda_k}(f_n)$.
Since Assumption A3 implies that $\lim_{n\to\infty}\inf_{k\in K_n} C^{\lambda_k}\left(\bar{f}^{\lambda_k}\right) = C^*$, we obtain

$$\lim_{n\to\infty}\inf_{k\in K_n} C^{\lambda_k}(f_n) = C^* \text{in probability if } r_\beta \leq 1, \text{and}$$

$$\lim_{n\to\infty}\inf_{k\in K_n} C^{\lambda_k}(f_n) = C^* \text{with probability 1 if } r_\beta > 1.$$

This implies that there exists a subsequence $\lambda_{k_1}, \lambda_{k_2}, \ldots$ such that $\lim_{n\to\infty} C^{\lambda_{k_n}}(f_n) = C^*$ in probability if $r_\beta \leq 1$, or almost surely if $r_\beta > 1$. Since $C^{\lambda_{k_n}}(f_n) = C(\lambda_{k_n} f_n)$, by applying [19], Lemma 5, the theorem follows. ∎

Remark regarding the set K_n In the case of i.i.d. observations, one does not need to restrict k^* to the set K_n (see [19]). However, in the case of β-mixing sequences, it seems that this restriction cannot be avoided at least for the proof methods that we have been using. Using elements of the proof of Lemma 2

in Section IV-C, one can show that for any $\epsilon > 0$ and any $\lambda_k > 0$, we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|C^{\lambda_k}(f) - C_n^{\lambda_k}(f)\right| > 3\epsilon + 6b_n\lambda_k\phi'(\lambda_k)\frac{c_1\sqrt{\mu_n}}{n}\right) \leq \exp\left(\frac{-4\epsilon^2 n}{\phi(\lambda_k)^2 b_n}\right) + \mu_n\beta_W(b_n), \quad (20)$$

where (b_n, μ_n) are given by (8), and $\beta_W(b_n) = O(n^{-r_\beta})$, provided that

$$\frac{\phi(\lambda_k)b_n}{n} = o\left(3\epsilon + 6b_n\lambda_k\phi'(\lambda_k)\frac{c_1\sqrt{\mu_n}}{n}\right).$$

In Step 1 above, suppose that we do not restrict $k \in K_n$. Then to be able to make the statement that

$$\mathbb{P}\left(\inf_{k\geq 1} C^{\lambda_k}(f_n) \leq \inf_{k\geq 1}\tilde{C}_n^{\lambda_k}(\hat{f}_n^{\lambda_k})\right) \geq 1 - \delta'(n),$$

for some $\delta'(n) < 1$ such that $\lim_{n\to\infty}\delta'(n) \to 0$, and for some penalized cost function $\tilde{C}_n^{\lambda_k}(\cdot)$, we need b_n in (20) to depend on k. So we cannot simply set $b_n = n^b$ as in the proof of Lemma 2. However, since $b_n(k)$ is the block size, it cannot exceed the sample size n. This imposes an upper limit on the values that k can take. For instance, let the block size be of the form $b_n(k) = n^b k^{\frac{d}{r_\beta+1}}$, for some constant $d > 1$. Then $\sum_{k=1}^{\infty}\mu_n(k)\beta_W(b_n(k)) < \infty$. However, since $b_n(k)$ is the block size, we need $k < n^c$, where $c = \frac{(1-b)(r_\beta+1)}{d}$. Hence it appears that we cannot avoid the restriction to a set such as K_n.

REFERENCES

[1] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *J. Amer. Statist. Associat.*, vol. 101, no. 473, pp. 138–156, 2006.

[2] P. Bartlett and M. Traskin, "AdaBoost is consistent," *J. Mach. Learn. Res.*, vol. 8, pp. 2347–2368, Oct. 2007.

[3] P. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," in *Proc. 14th Annu. Conf. Comput. Learn. Theory*, 2001, pp. 224–240.

[4] L. Breiman, "Some infinity theory of predictor ensembles," Dept. Statist., Univ. California, Berkeley, CA, USA, Tech. Rep. 579, 2000.

[5] P. Bühlmann, "Consistency for L_2 boosting and matching pursuit with trees and tree-type basis functions," Seminar fur Statistik, ETH, Zürich, Switzerland, Tech. Rep. 109, 2002.

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

[7] P. Doukhan, *Mixing Properties and Examples*. New York, NY, USA: Springer-Verlag, 1995.

[8] Y. Freund, "Boosting a weak learning algorithm by majority," *Inf. Comput.*, vol. 121, no. 2, pp. 256–285, 1995.

[9] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[10] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Ann. Statist.*, vol. 28, no. 2, pp. 337–374, 2000.

[11] L. Györfi, W. Härdle, P. Sarda, and P. Vieu, *Nonparametric Curve Estimation from Time Series* (Lecture Notes in Statistics). Berlin, Germany: Springer-Verlag, 1989.

[12] A. Irle, "On the consistency in nonparametric estimation under mixing assumptions," *J. Multivariate Anal.*, vol. 60, no. 1, pp. 123–147, 1997.

[13] W. Jiang, "Does boosting overfit: Views from an exact solution," Dept. Statist., Northwestern Univ., Evanston, IL, USA, Tech. Rep. 00-03, 2000.

[14] W. Jiang, "Process consistency for AdaBoost," *Ann. Statist.*, vol. 32, no. 1, pp. 13–29, 2004.

[15] M. Kearns and L. G. Valiant, "Learning boolean formulae or finite automata is as hard as factoring," Harvard University Aiken Computation Laboratory, Cambridge, MA, USA, Tech. Rep. TR-14-88, 1988.

[16] V. Koltchinskii and D. Panchenko, "Rademacher processes and bounding the risk of function learning," in *High Dimensional Probability II*. New York, NY, USA: Springer-Verlag, 2000, pp. 443–459.

[17] M. Ledoux and N. Talagrand, *Probability in Banach Spaces*. New York, NY, USA: Springer-Verlag, 1991.

[18] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," *Neural Netw.*, vol. 6, no. 6, pp. 861–867, 1993.

[19] G. Lugosi and N. Vayatis, "On the Bayes-risk consistency of boosting methods," *Ann. Statist.*, vol. 32, no. 1, pp. 30–55, 2004.

[20] S. Mannor, R. Meir, and T. Zhang, "Greedy algorithms for classification—Consistency, convergence rates, and adaptivity," *J. Mach. Learn. Res.*, vol. 4, pp. 713–741, Oct. 2003.

[21] R. Meir, "Nonparametric time series prediction through adaptative model selection," *Mach. Learn.*, vol. 39, no. 1, pp. 5–34, 2000.

[22] R. Meir and T. Zhang, "Generalization error bounds for Bayesian mixture algorithms," *J. Mach. Learn. Res.*, vol. 4, pp. 839–860, Oct. 2003.

[23] D. Modha and E. Masry, "Memory-universal prediction of stationary random processes," *IEEE Trans. Inf. Theory*, vol. 44, no. 1, pp. 117–133, Jan. 1998.

[24] G. G. Roussas, "Nonparametric estimation in mixing sequences of random variables," *J. Statist. Plan. Inference*, vol. 18, no. 2, pp. 135–149, 1988.

[25] R. E. Schapire, "The strength of weak learnability," *Mach. Learn.*, vol. 5, no. 2, pp. 197–227, 1990.

[26] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Proc. MSRI Workshop Nonlinear Estimat. Classification*, Berkeley, CA, USA, 2002, pp. 149–172.

[27] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes* (Springer Series in Statistics). New York, NY, USA: Springer-Verlag, 1996.

[28] M. Vidyasagar, *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*, 2nd ed. London, U.K.: Springer-Verlag, 2002.

[29] B. Yu, "Some results on empirical processes and stochastic complexity," Ph.D. dissertation, Dept. Statist., Univ. California, Berkeley, CA, USA, Apr. 1990.

[30] B. Yu, "Density estimation in the L^∞ norm for dependent data with applications to the Gibbs Sampler," *Ann. Statist.*, vol. 21, no. 2, pp. 711–735, 1993.

[31] B. Yu, "Rate of convergence for empirical processes of stationary mixing sequences," *Ann. Probab.*, vol. 22, no. 1, pp. 94–116, 1994.

[32] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," *Ann. Statist.*, vol. 32, no. 1, pp. 56–85, 2004.

[33] T. Zhang and B. Yu, "Boosting with early stopping: Convergence and consistency," *Ann. Statist.*, vol. 33, no. 4, pp. 1538–1579, 2005.

Aurélie C. Lozano received the M.S./Dipl.Ing. degree in Communication Systems from the Swiss Federal Institute of Technology Lausanne (EPFL) in 2001, and the M.A. and Ph.D. degrees in Electrical Engineering from Princeton University respectively in 2004 and 2007. Since 2007, Dr. Lozano has been a Research Staff Member in the Machine Learning Group at the IBM T.J. Watson Research Center. Her research interests include machine learning, statistics and data mining. Her current focus is on high dimensional data analysis and predictive modeling, with applications including biology, environmental sciences, business and infrastructure analytics, and social media analytics.

Sanjeev R. Kulkarni (M'91–SM'96–F'04) received the B.S. in Mathematics, B.S. in E.E., M.S. in Mathematics from Clarkson University in 1983, 1984, and 1985, respectively, the M.S. degree in E.E. from Stanford University in 1985, and the Ph.D. in E.E. from M.I.T. in 1991. From 1985 to 1991, he was a Member of the Technical Staff at M.I.T. Lincoln Laboratory. Since 1991, he has been with Princeton University where he is currently Professor of Electrical Engineering, and an affiliated faculty member in the Department of Operations Research and Financial Engineering and the Department of Philosophy. He spent January 1996 as a research fellow at the Australian National University, 1998 with Susquehanna International Group, and Summer 2001 with Flarion Technologies. Prof. Kulkarni received an ARO Young Investigator Award in 1992, an NSF Young Investigator Award in 1994. He is a Fellow of the IEEE and has served as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY. Prof. Kulkarni's research interests include statistical pattern recognition, nonparametric statistics, learning and adaptive systems, information theory, wireless networks, and image/video processing.

Robert E. Schapire received his ScB in math and computer science from Brown University in 1986, and his SM (1988) and PhD (1991) from MIT under the supervision of Ronald Rivest. After a short post-doc at Harvard, he joined the technical staff at AT&T Labs (formerly AT&T Bell Laboratories) in 1991 where he remained for eleven years. Since 2002, he has been on the faculty of Princeton University where he is currently the David M. Siegel '83 Professor in Computer Science. His awards include the 1991 ACM Doctoral Dissertation Award, the 2003 Gödel Prize and the 2004 Kanelakkis Theory and Practice Award (both of the last two with Yoav Freund). His main research interest is in theoretical and applied machine learning.