Sequential Procedures for Aggregating Arbitrary Estimators of a Conditional Mean

Florentina Bunea * and Andrew B. Nobel †

23 January, 2008

Abstract

In this paper we describe and analyze a sequential procedure for aggregating linear combinations of a finite family of regression estimates, with particular attention to linear combinations having coefficients in the generalized simplex. The procedure is based on exponential weighting, and has a computationally tractable approximation. Analysis of the procedure is based in part on techniques from the sequential prediction of non-random sequences. Here these techniques are applied in a stochastic setting to obtain cumulative loss bounds for the aggregation procedure. From the cumulative loss bounds we derive an oracle inequality for the aggregate estimator for an unbounded response having a suitable moment generating function. The inequality shows that the risk of the aggregate estimator is less than the risk of the best candidate linear combination in the generalized simplex, plus a complexity term that depends on the size of the coefficient set. The inequality readily yields convergence rates for aggregation over the unit simplex that are within logarithmic factors of known minimax bounds. Some preliminary results on model selection are also presented.

Appears in IEEE Transactions on Information Theory, vol.54, pp.1725-1735, 2008

*Florentina Bunea is with the Department of Statistics, Florida State University, Tallahassee, FL 32306-4330. Email: bunea@stat.fsu.edu. Corresponding author. Research partially supported by NSF-DMS 0706829.

†Andrew Nobel is with the Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-3260. Email: nobel@email.unc.edu

1 Introduction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of a jointly distributed pair (X, Y) where $X \in \mathbb{R}^d$ is a vector of real-valued covariates, and $Y \in \mathbb{R}$ is a real-valued response. Suppose that $EY^2 < \infty$, and let $f(x) = E(Y \mid X = x)$ be the regression function of Y on X. We do not assume that Y is related to $f(X)$ through a standard additive model. The predictive performance of a fixed estimate or data-dependent estimator of f will be measured in terms of its squared loss. Let ν denote the common d-variate distribution of the random vectors X_i. For a fixed estimate $F : \mathbb{R}^d \to \mathbb{R}$ let

$$||\widehat{F} - f||^2 = \int (\widehat{F}(x) - f(x))^2 \, d\nu(x)$$

be the $L_2(\nu)$ distance between F and f. The risk of an estimator $\widehat{F}$ depending on (X_1, Y_1), $\ldots$,(X_n, Y_n) is the expectation $E||\widehat{F} - f||^2$.

Let $\mathcal{F} = \{F_j : 1 \le j \le M\}$ be a family of fixed, bounded estimates of f. In what follows we will be interested in linear combinations of the base estimates F_j. For each $\lambda = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^M$ let

$$F_\lambda = \sum_{j=1}^{M} \lambda_j F_j.$$

For $c > 0$ let $\Lambda(c) \subseteq \mathbb{R}^M$ be the generalized simplex consisting of coefficients $\lambda = (\lambda_1, \ldots, \lambda_M)$ such that $\sum_{j=1}^m |\lambda_j| \le c$. Given observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ and the base estimates $\mathcal{F}$, we wish to construct an estimator $\widehat{F}$ whose performance is comparable to that of the best estimate F_λ with $\lambda \in \Lambda(c)$. To be more precise, we are interested in oracle-type inequalities of the form

$$E||\widehat{F} - f||^2 \le \inf_{\lambda \in \Lambda(c)} ||F_\lambda - f||^2 + C \, \frac{\Delta(n, M, c)}{n}, \tag{1}$$

where C is a constant independent of n and M. The first term on the right hand side of (1) is the optimal risk of the estimates F_λ, which is the natural target for $\widehat{F}$. The quantity $\Delta(n, M, c)$ represents the cost associated with a data-driven search among the estimates F_λ for a combination having minimal risk. It is important to emphasize the requirement in (1) that the leading coefficient of the infimum be equal to one, as most existing oracle inequalities have leading coefficients greater than one. This issue is discussed in more detail below.

Our analysis of (1) begins with the simpler problem of model selection, where the goal is oracle inequalities of the form

$$E||\widehat{F} - f||^2 \le \min_{1 \le j \le M} ||F_j - f||^2 + C \, \frac{\Delta_0(n, M)}{n} \quad \text{(MS)}. \tag{2}$$

We note that the bound (1) yields inequalities for the set Λ_1 of sub-convex coefficients $(\lambda_1, \ldots, \lambda_M)$ such that $\lambda_j \geq 0$ and $\sum_{j=1}^{M} \lambda_j \leq 1$. Indeed, as $\Lambda_1 \subseteq \Lambda(1)$, inequality (1) yields bound

$$E||\widehat{F} - f||^2 \ \leq \ \inf_{\lambda \in \Lambda_1} ||F_\lambda - f||^2 \ + \ C \, \frac{\Delta_1(n, M)}{n} \qquad \text{(CVX)} \tag{3}$$

Minimax bounds on the quantities $\Delta_0(n, M)$ and $\Delta_1(n, M)$ associated with model selection and sub-convex combinations, respectively, have been established by Tsybakov [29], who showed that

$$\Delta_0(n, M) \ \geq \ \ln M \text{ and} \tag{4}$$

$$\Delta_1(n, M) \ \geq \ M \cdot I\{M \leq \sqrt{n}\} \ + \ \sqrt{n \, \log(1 + M/\sqrt{n})} \cdot I\{M > \sqrt{n}\} \tag{5}$$

These bounds were obtained for the nonparametric regression model $Y = f(X) + \epsilon$, where X and ϵ are independent and ϵ is a Gaussian term with mean zero and known variance. Here we consider the more general model in which Y and X need only be jointly distributed. Nevertheless, we use the minimax bounds (4) and (5) as a (conservative) benchmark for the performance of our estimators.

Over the past decade, there have been a number of papers describing methods that achieve, or nearly achieve, the minimax optimal bounds. Much of this work yields bounds of the type

$$E||\widehat{F} - f||^2 \ \leq \ K \inf_{F \in \mathcal{H}} ||F - f||^2 \ + \ C \, \frac{\Delta(n, M, \mathcal{H})}{n}, \tag{6}$$

with $K > 1$. Inequalities of this sort, in which $\Delta(n, M, \mathcal{H})$ is the minimax aggregation cost for model selection ($\mathcal{H} = \{F_\lambda : \lambda \in \Lambda_1\}$) and sub-convex combinations ($\mathcal{H}$, have been obtained by Yang [34, 35, 36] using sequential aggregation methods different from those described below, and by Bunea, Tsybakov and Wegkamp [5] using penalized least squares procedures. Wegkamp [32] considers a data-adaptive model selection procedure and establishes (6), with the optimal $\Delta(n, M, \mathcal{F}) = \log M$. Birgé [4] studies an aggregation method that utilizes convex weights, and establishes (3) for with an aggregation cost that is optimal for $M > \sqrt{n}$ and suboptimal for $M \leq \sqrt{n}$. In all these cases there is a tradeoff between the size of the constants K and C. As K approaches one, C becomes large, and the bound is valuable only for very large n. As C approaches one, K increases, and the connection between the risk of $\widehat{F}$ and that of the best $F \in \mathcal{H}$ is weakened accordingly. In general, bounds of type (6) can not guarantee that $\widehat{F}$ provides improvement over the original estimators, even accounting for the cost of aggregation.

This motivates interest in aggregation and other methods that yield a bound (1) with leading constant one and the minimax, or near minimax, remainder term. The literature in

this area is still growing. The minimax (CVX) bound for $M > \sqrt{n}$ can be obtained by least squares convex aggregation, as in Juditsky and Nemirovski [19] and Nemirovski [25], or a sequential mirror descent algorithm, as in Juditsky *et al.* [18]. The minimax (CVX) bound for both cases has been obtained by Audibert [1] using a PAC-Bayesian aggregation method, and by Koltchinskii [21] who proposed aggregation via risk minimization with Rademacher complexity penalties. Catoni [11] established the minimax (MS) bound for a sequential Bayesian estimator. We also mention here the closely related work of Leung and Barron [22], who consider i.i.d. observations from the simplified model $Y = \theta + W$, where $\theta \in \mathbb{R}$ is an unknown mean and W is a Gaussian error term of mean zero and known variance. Using a convex, non-sequential, aggregation method based on exponential weights, they establish the empirical norm analog of the (MS) bound; their focus is on best fit rather than best prediction.

1.1 Overview and Motivation

In this paper we show that estimators with near optimal performance as in (3) and (2) can be obtained by a simple sequential aggregation procedure. We assume that the base estimates $\mathcal{F}$ are bounded, and that the response Y has a suitable moment generating function. Under these conditions, the aggregation procedure yields an estimator $\widehat{F}$ satisfying (1) with aggregation cost $CM \ln(cMn \ln n)$, where C is independent of M, n and c. When $c = 1$ and $M < \sqrt{n}$, $\widehat{F}$ satisfies the convex bound (3) with an error term that is within a logarithmic factor of the minimax bound. We also establish a number of oracle inequalities for model selection (MS); these yield an estimator that comes within a logarithmic factor of the minimax rate in (3) when $M > \sqrt{n}$.

The aggregation procedure described here is an extension of existing aggregation methods for the sequential prediction of deterministic individual sequences. Representative work and related references can be found in Vovk [30, 31], Littlestone and Warmuth [23], Freund [13], Cesa-Bianchi *et al.* [8], Cesa-Bianchi *et al.* [7], Haussler *et al.* [17], Yamanishi [33], Singer and Feder [27], Cesa-Bianchi and Lugosi [9], and Azoury and Warmuth [2]. See also the surveys of Foster and Vohra [12] and Merhav and Feder [24].

Aggregation methods from the deterministic prediction literature have been applied to several stochastic prediction problems. Györfi, Lugosi and Morvai [16] applied aggregation methods to exhibit a randomized universal predictor for binary ergodic processes. Related ideas are used by Gyorfi and Lugosi [15], and Nobel [26], to define universal prediction schemes for unbounded ergodic processes. In both cases, unboundedness is addressed by

applying results for bounded sequences to blocks of observations, and letting the bound grow slowly with sample size. Yang [37] considered the problem of sequential prediction in a non-stationary time series setting with covariate-response type observations. Building on the analysis of Catoni [11], he established an oracle type inequality for the prediction problem, with leading constant one, under suitable conditions on the moment generating function of the errors.

1.2 Outline

The rest of the paper is organized as follows. The next section presents a sequential aggregation based estimator for model selection and establishes its performance for bounded and unbounded responses, and for dependent observations. Section 3 is devoted to an analogous aggregation based estimator for the generalized simplex. The results of this section extend those of Cesa-Bianchi and Lugosi [10] to a stochastic setting where the response Y is not bounded and performance is evaluated with respect to the square loss. Oracle inequalities for the estimator are established in our main theorem, Theorem 1, under conditions on the moment generating function of the response Y. The results of Theorem 1 yield oracle inequalities for convex aggregation that are within a logarithmic factor of the minimax bound in (3) when $|\mathcal{F}| \leq \sqrt{n}$. Near-minimax bounds for (3) in the case $|\mathcal{F}| > \sqrt{n}$ are established in Theorem 2 using the model selection results in Section 2. The proofs of the results in Section 2 are given in Appendix I, while Appendices II and III contain the proofs of Theorems 1 and 2, respectively.

2 Model Selection Bound for a Sequential Procedure

Here we describe a procedure for aggregating estimates of a conditional mean function. Aggregation is based on a data set of size n. For each $k \geq 1$ we produce a pre-estimator from $(X_1, Y_1), \ldots, (X_k, Y_k)$. The pre-estimate is a convex combination of either the base estimates in $\mathcal{F}$ or linear combinations of the base estimates. The weights assigned to different estimates are data dependent. The final estimate is the average of the n pre-estimators.

2.1 Aggregation Procedure 1: Discrete Weights

Given: Base estimates $\mathcal{F} = \{F_1, \ldots, F_M\}$, and observations $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Step 1: Fix a constant $\eta > 0$. Let $C_k(F_j) = \sum_{i=1}^{k} (Y_i - F_j(X_i))^2$. At time $0 \le k \le n-1$ assign to each $F_j \in \mathcal{F}$ a weight

$$w_k(F_j) = \frac{\exp\{-\eta\, C_k(F_j)\}}{\sum_{1 \le l \le M} \exp\{-\eta\, C_k(F_l)\}} \tag{7}$$

Note that $w_0(\cdot) = M^{-1}$. Let the k'th pre-estimate

$$\widehat{G}_k(x) = \sum_{1 \le j \le M} w_k(F_j)\, F_j(x). \tag{8}$$

be a weighted sum of the base estimates.

Step 2: Average the pre-estimates $\widehat{G}_0, \ldots, \widehat{G}_{n-1}$ to produce the final aggregate

$$\widehat{F}(x) = \frac{1}{n} \sum_{k=0}^{n-1} \widehat{G}_k(x) \tag{9}$$

By definition, the weights $w_k(F_j)$ are positive and sum to one. Thus each pre-estimate $\widehat{G}_k$ is a convex combination of the functions F_j with data-dependent weights based on Z^k. Likewise, $\widehat{F}(x)$ can be written as a convex combination $\sum_{1 \le j \le M} \hat{\pi}(j)\, F_j$ with $\hat{\pi}(j) = n^{-1} \sum_{k=0}^{n-1} w_k(F_j)$. ¿From (7) it is clear that the weight $w_k(F_j)$ assigned to an estimate F_j depends on the relative cumulative prediction loss of F_j on the observations (X_1, Y_1), $\ldots$, (X_k, Y_k). Estimates with poor predictive performance are given less weight than those that perform well. The parameter η controls the sensitivity of the weights to the predictive performance of the estimates. Larger values of η lead to more aggressive down-weighting of poor performers. For easy reference we refer to this algorithm as Procedure 1.

Procedure 1 is an extension to the stochastic regression setting of aggregation methods developed for prediction of bounded deterministic sequences (*cf.* Kivinen and Warmuth [20], Singer and Feder [27]). Similar procedures for sequential prediction of stochastic sequences have been studied by Gyorfi and Lugosi [15], and Yang [37]. The estimator of Procedure 1 was proposed and analyzed by Catoni [11], who established a minimax optimal oracle inequality for $\widehat{F}$ when the response Y is unbounded and satisfies a moment generating function condition. (See Section 3 for more details.) Here we present an alternative analysis, which yields, with little effort, the optimal oracle inequality for $\widehat{F}$ when Y is bounded. The same analysis yields oracle type inequalities for stationary observations (see Theorem 3), and with minor modifications, one may also obtain bounds like those in Theorem 5 of Yang [37] for non-stationary observations. The case of unbounded Y can be treated via a simple truncation argument, which yields oracle inequalities that differ from the minimax bounds only by logarithmic factors when Y has a moment generating function. Elements of the

approach here and that of Catoni [11] are used in Section 3 to establish oracle inequalities for coefficients taking values in the simplex and generalized simplex.

2.2 Performance Bound

The analysis of Procedure 1 is given in Appendix I. It is shown in Lemma 1 that the cumulative loss of the pre-estimates $\widehat{G}_k$ is bounded above by the minimum cumulative loss of the available estimates plus additional error terms. This bound is a straightforward extension to the regression setting of arguments developed by Singer and Feder [27, 28] for the prediction of individual sequences (see also [15]). A similar bound covering a variety of convex loss functions can be found in Kivinen and Warmuth [20]. Related arguments for individual sequences were used by Littlestone and Warmuth [23], who established cumulative loss bounds for a number of simple sequential updating strategies related to gradient descent. However, the leading coefficient of the competitive target in their bounds is greater than one, and this prevents the bounds from being converted to a oracle inequality involving the standard risks $||F - f||$. The cumulative loss bound in Lemma 1 leads to the following oracle inequality for $\widehat{F}$ when the response Y is bounded.

Proposition 1. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed. Assume that each of the estimates $F_k \in \mathcal{F}$ is bounded by $b > 0$ and that $|Y_i| \leq \gamma$ for some $\gamma \geq b$. If the parameter $\eta = (2(\gamma + b)^2)^{-1}$, then the aggregation procedure described above produces an estimate $\widehat{F}$ such that*

$$E||\widehat{F} - f||^2 \;\leq\; \min_{1 \leq j \leq M} ||f - F_j||^2 \;+\; \frac{2(\gamma + b)^2 \ln M}{n}. \tag{10}$$

The conclusion of Proposition 1 can be extended to unbounded response variables Y_i through an additional truncation step in the aggregation procedure. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. with Y possibly unbounded. Given $\gamma \geq b$, define

$$\tilde{Y}_k \;=\; \begin{cases} -\gamma & \text{if } Y_k < -\gamma \\ Y_k & \text{if } -\gamma \leq Y_k \leq \gamma \\ \gamma & \text{if } Y_k > \gamma. \end{cases} \tag{11}$$

by thresholding the value of Y_k at $+\gamma$ and $-\gamma$. Let $\widetilde{F}$ be the aggregate estimate obtained from $(X_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Y}_n)$.

Proposition 2. *Suppose that each of the estimates $F_j \in \mathcal{F}$ is bounded by $b > 0$ and that $EY^2 < \infty$. If $\gamma \geq b$ and $\eta = (2(\gamma + b)^2)^{-1}$, then the aggregate estimate $\widetilde{F}$ satisfies*

$$E||\widetilde{F} - f||^2 \;\leq\; \min_{1 \leq j \leq M} ||f - F_j||^2 \;+\; \frac{2(\gamma + b)^2 \ln M}{n} \;+\; 8(b + \gamma)\, E|Y|\, I\{|Y| \geq \gamma\}. \tag{12}$$

This immediately yields the following corollary. Let $W =: Y - E(Y|X)$.

Corollary 1. *Assume that there exists finite constants $b, L > 0$ such that $E\exp(|W|) \leq L$ and $|F_j|, |f| \leq b$. If $\gamma = 2b\ln n$ and $\eta = (2(\gamma + b)^2)^{-1}$, then*

$$E||\widetilde{F} - f||^2 \;\leq\; \min_{1\leq j\leq M} ||f - F_j||^2 \;+\; \frac{2b^2(2\ln n + 1)^2 \ln M}{n} \;+\; \frac{C_2}{n}, \tag{13}$$

where C_2 is positive constant depending on L and b, but not on n.

The proof of this corollary follows directly from Proposition 2 and Lemma 2 in Appendix I.

We note that the error term in (13) differs from the minimax lower bound $Cn^{-1}\ln M$, $C > 0$, only by logarithmic factors in n. This is the price paid for choosing an explicitly defined tuning parameter η depending on n and b.

Catoni [11][Theorem 3.6.1, page 87], showed that Procedure 1 applied directly to the unbounded responses Y leads to the minimax lower bound. This result requires a practical tradeoff: his tuning parameter, which is independent of n, depends on b and also on suitable bounds on the moment generating function of W.

Both our Proposition 2 and Theorem 3.6.1 in Catoni assume the existence of a moment generating function for W. If this assumption holds, either approach is valid. In particular, the choice of η given in Proposition 2 avoids the evaluation of the bounds defined in and (A4) below, at a relatively low theoretical cost, especially for large n. Related arguments can be found in Section 3 below.

2.3 Results in Non-IID Settings

The cumulative loss bound of Lemma 1 holds pointwise (on an ω-by-ω basis) and therefore applies to arbitrary stochastic (or individual) sequences. From the Lemma one may readily derive a number of corollaries under different dependence assumptions. Let $X_1, X_2, \ldots$ be a stationary process with values in $\mathbb{R}^d$, and let $\varepsilon_1, \varepsilon_2, \ldots$ be a stationary martingale difference sequence that is independent of $\{X_i\}$, and such that $E\varepsilon_i^2 < \infty$. The martingale difference assumption,

$$E[\varepsilon_k \,|\, \varepsilon_1, \ldots, \varepsilon_{k-1}] \;=\; 0 \;\text{ for }\; k \geq 1 \tag{14}$$

implies in particular that $E\varepsilon_i = 0$. Suppose in addition that the response Y_k is related to the covariate X_k through the additive model

$$Y_k \;=\; f(X_k) + \varepsilon_k, \tag{15}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is any measurable function such that $Ef(X)^2 < \infty$. Then f is the regression function of Y on X. Let $\hat{G}_0, \ldots, \hat{G}_{n-1}$ be defined using the aggregating method outlined in (7) and (8), with parameter $\eta > 0$.

Proposition 3. *Assume that each function $F_j \in \mathcal{F}$ is bounded in absolute value by $b < \infty$ and that $|Y_i| \leq \gamma$ with $\gamma \geq b$. If $\eta = (2(\gamma + b)^2)^{-1}$, then*

$$\frac{1}{n} \sum_{k=1}^{n} E(f(X_k) - \hat{G}_{k-1}(X_k))^2 \;\leq\; \min_{1 \leq j \leq M} E||f - F_j||^2 \;+\; \frac{2\,(\gamma + b)^2\, \ln M}{n} \tag{16}$$

Remark: The upper bound in (16) is identical to the upper bound in Proposition 1 above: the leading coefficients of $\min_{1 \leq j \leq M} E||f - F_j||^2$ is equal to one, and the remaining terms are as in the i.i.d. case. The dependence of the observations affects only the left hand side of (16). As X_k may be stochastically dependent upon the observations $Z_1, \ldots, Z_{k-1}$ used to produce $\hat{G}_{k-1}$, the term $E(f(X_k) - \hat{G}_{k-1}(X_k))^2$ is not necessarily equal to the expected risk of $\hat{G}_{k-1}$. However, $f(X_k)$ is still the optimal predictor of Y_k given X_k and the previous observations Z^{k-1}, and $\hat{G}_{k-1}(X_k)$ is a competing (aggregate) predictor of Y_k given X_k and Z^{k-1}. Accordingly, we can view $E(f(X_k) - \hat{G}_{k-1}(X_k))^2$ as the expected predictive risk of G at time k. Then the right hand side of (16) is simply the average expected predictive risk of G on $Z_1, \ldots, Z_n$. This quantity satisfies the same upper bound as the risk of the aggregate estimator $\widehat{G}$ in the i.i.d. case, even though the covariates X_i are assumed only to be stationary.

3 Oracle Inequality for the Generalized Simplex

Here we describe a continuous version of the algorithm presented in Section 2. Data dependent weights are now assigned to linear combinations of the base estimates, rather than individual estimates. These weights are used, as before, to construct n pre-estimates the average of which yields the final estimator. Fix $c \geq 1$ and define

$$\Lambda = \{\lambda \in R^M : \sum_{j=1}^{M} |\lambda_j| \leq c\} \tag{17}$$

3.1 Aggregation Procedure 2: Continuous Weights

Input: Base estimates $\mathcal{F}$, coefficients Λ and observations $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Step 1: Fix a constant $\eta > 0$. At each time $0 \leq k \leq n-1$ assign to each linear combination $F_\lambda \in \mathcal{F}(\Lambda)$ a weight

$$w_k(\lambda) \;=\; \frac{\exp\{-\eta\, C_k(F_\lambda)\}}{\int_\Lambda \exp\{-\eta\, C_k(F_{\lambda'})\}\, d\lambda'}, \tag{18}$$

where $d\lambda$ denotes the Lebesgue measure on Λ and $C_k(F_\lambda) = \frac{1}{k}\sum_{i=1}^{k}(Y_i - F_\lambda(X_i))^2$. Define the k'th pre-estimate by

$$\widetilde{G}_k(x) \;=\; \int_\Lambda F_\lambda(x)\, w_k(\lambda)\, d\lambda. \tag{19}$$

Step 2: Average the pre-estimates $\widetilde{G}_0, \ldots, \widetilde{G}_{n-1}$ to produce the final aggregate

$$\widehat{F}(x) \;=\; \frac{1}{n}\sum_{k=0}^{n-1}\widetilde{G}_k(x). \tag{20}$$

By definition, $\int_\Lambda w_k(\lambda)d\mu(\lambda) = 1$ for every $k \geq 0$. Thus $w_k(\cdot)$ can be interpreted as a probability density on Λ, and both the pre-estimates and final aggregate are linear combinations of the elements of $\mathcal{F}$. Also note that the pre-estimates defined in Step 1 can be viewed as posterior means with respect to the density $w_k(\lambda)$. Thus they can be computed via standard Metropolis-Hastings type algorithms as described, for instance, in Chapter 11, pages 283 - 310, of Gelman et al (2003).

3.2 Results

The results of this section will be proved under the following assumptions. Recall that $W = Y - E(Y|X)$.

(A1) There exists $b > 0$ such that $|F_j|, |f| \leq b$.

(A2) There exists $0 < L < \infty$ such that $E\exp(|W|) \leq L$.

Remark. We note that assumption (A2) immediately implies the following two conditions.

(A3) For every $0 < a < 1$

$$E\exp(a|W|) \leq L.$$

(A4) For every $0 < a < 1$ there exists a constant $0 < V_a < \infty$ such that

$$\frac{E[W^2\exp(a|W|)]}{E\exp(a|W|)} \leq V_a.$$

Theorem 1. *Let $\widehat{F}$ be defined as in Procedure 2 with coefficients $\Lambda \subseteq \mathbb{R}^M$ defined in (17). Assume that (A1) and (A2) hold.*
(a) Let $\eta > 0$ be any number satisfying

$$\frac{1}{\eta} \geq \frac{L}{2}\exp(4\eta b^2c^2)(15b^2c^2 + 3V_{8bc\eta}), \tag{21}$$

for $V_{8bc\eta}$ as in (A4) above. Then for every $n \geq 1$ and for dominating constants B, B_1 independent of M and n we have:

$$E\|\widehat{F} - f\|^2 \leq \min_{\lambda \in \Lambda} E\|F_\lambda - f\|^2 + \frac{B_1 M \ln(Bn \ln n)}{n\eta}. \tag{22}$$

(b) Let $\eta = (2(\gamma_n + bc)^2)^{-1}$ with $\gamma_n = 2bc \ln n$. If Procedure 2 is applied to the truncated responses $\tilde{Y}_k$ defined in (11), the resulting estimator satisfies (22) with different values of B and B_1 also independent of M and n.

Remark. It is easy to check that the bound (21) holds when η is sufficiently small. As discussed after Corollary 1 above, one can use either the strategy in (a) or that in (b), the tradeoff being between the presence of extra logarithmic terms in the oracle bound versus the explicit form of η.

Vovk [31] considers a sequential ridge regression type estimators for individual sequences in the context of what he terms competitive online statistics. Azoury and Warmuth [2] also suggest variants of sequential ridge regression aggregation. Their Forward Algorithm coincides with Vovk's [31] in the linear regression context. They establish oracle inequalities for the cumulative loss of the resulting aggregate estimators that take the following form, in our notation. Let $U =: \max_{1 \leq i \leq n} Y_i^2$ and let $\hat{f}$ be the aggregate obtained by one of their procedures. They then show that

$$C_n(\hat{f}) \leq \inf_{\lambda \in \mathbb{R}^M} \left(C_n(F_\lambda) + a\|\lambda\|^2 \right) + dUM \ln(\frac{nb^2}{a} + 1),$$

where d is a small positive constant, b is a common bound on the base estimates, $\| \ \|$ is the Euclidean norm and $a > 0$ is a user specified tuning constant required by their algorithms.

When such bounds hold, one can use the arguments given in the proof of Proposition 1, to obtain expectation-type oracle inequalities that are minimax optimal within logarithmic factors, when U, and therefore the responses Y_i, are bounded and for coefficients λ with bounded Euclidean norm. If the responses are unbounded, one needs to resort to different strategies.

Our Procedure 2 offers a possible solution. It belongs (as Procedure 1) to a general class of methods often called Bayes' algorithms. Azoury and Warmuth [2] contains a summary of procedures of this type. To the best of our knowledge, the existing analyses of such procedures focus on asymptotic bounds on the cumulative risk of the aggregate, when the response is bounded (see, e.g., Freund [13] and Yamanishi [33]). The crucial point in such analyses is finding an upper bound on integrals of the type

$$-\frac{1}{\eta} \ln \int_\Lambda e^{-\eta C_n(F_\lambda)}.$$

The approach taken in the works mentioned above is to find an asymptotic (in n) approximation of this integral, typically using Laplace's method.

The analysis of our Procedure 2 also requires the investigation of this integral. In contrast with the existing methods, we provide a non-asymptotic upper bound for it. Our main tools are presented in in Lemma 5. Moreover, we also provide non-asymptotic upper bounds for the cumulative risk that allow us to obtain upper bounds on the risk when the response Y is unbounded and has a moment generating function. The results presented below for the generalized simplex appear to be new, and have not been investigated in the literature in this context; we refer to Theorem 3.3 of Section 3 in Cesa-Bianchi and Lugosi [10] for oracle bounds for the cumulative risk, in the case of bounded response Y and bounded exp-concave loss functions. Let $\Lambda_{1,M}$ be the set of sub-convex weight vectors $\lambda = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^M$ such that $\lambda_j \geq 0$ and $\sum_{j=1}^M \lambda_j \leq 1$. As $\Lambda_1 \subseteq \Lambda(1)$, Theorem 1 immediately yields an oracle inequality for Λ_1. For $M \leq \sqrt{n}$ the bound is within a logarithmic factor of the optimum value M/n given in (5). As Theorem 1 is non-asymptotic, (22) holds for each fixed sample size n. Suppose that $\mathcal{F}_n$ has cardinality $M_n = n^\tau$ for some $\tau \leq 1/2$ and let

$$\mathcal{H}_n = \left\{ \sum_{F_j \in \mathcal{F}_n} \lambda_j F_j : \lambda_j \geq 0 \text{ and } \sum_{j=1}^M \lambda_j \leq 1 \right\}.$$

Then Theorem 1 implies that $E||\widehat{F} - f||^2 \leq \inf_{F \in \mathcal{H}_n} ||F - f||^2 + O(n^{-1+\tau} \ln^2 n)$. This extends earlier results of Yang [34, 35, 36] who obtained similar bounds (with a leading constant greater than one) for coefficients having an l_1-norm bounded by one.

When $M > \sqrt{n}$, the estimator and bound of Theorem 1 can be improved. Let $m \leq M$ be a fixed integer that will be specified later, and let

$$\mathcal{H} \;=\; \left\{ \sum_{j=0}^M \lambda_j F_j : \text{each } \lambda_j \geq 0 \text{ is an integer multiple of } 1/m \text{ and } \sum_{j=0}^M \lambda_j = 1 \right\} \qquad (23)$$

where $F_0 = 0$. Then $\mathcal{F} \subseteq \mathcal{H}$ and no function in $\mathcal{H}$ can involve more than m of the base estimates $\{F_1, \ldots, F_M\}$. Let $\widehat{F}$ be the aggregate estimator obtained by applying Procedure 1 to the collection $\mathcal{H}$ with parameter η and m equal to the integer part of

$$\alpha_{n,M} = \frac{1}{2} \sqrt{\frac{n \ln 2}{\ln(1 + M/\sqrt{n})}}.$$

Note that $|\mathcal{H}|$ is less than or equal to the cardinality of the set of $(M+1)$-dimensional vectors having non-negative integer entries summing to m; thus $|\mathcal{H}| \leq \binom{M+m}{m}$.

The proof of the following theorem is given in the Appendix. Its proof is based on a Maurey-type argument (see, *e.g.*, Barron [3], Nemirovski [25] for other applications); we

follow the outline of Theorem 5 of Tsybakov [29] and Corollary 3.4 in Bunea, Tsybakov and Wegkamp [5].

Theorem 2. *(a) Assume that (A1) and (A2) hold. If $\gamma = 2b \ln n$, then Procedure 1 applied to the family $\mathcal{H}$ in (23) with $\eta = (2(\gamma + b)^2)^{-1}$, yields, for every $n \geq 1$ and $M \geq \sqrt{n}$*

$$E||\widehat{F} - f||_2^2 \ \leq \ \inf_{\lambda \in \Lambda_1} ||F_\lambda - f||^2 \ + \ 16(\gamma + b)^2 \sqrt{\left\{\log(1 + M/\sqrt{n})\right\}/n} + \frac{C}{n}, \qquad (24)$$

where C is a positive dominating constant independent of M and n. The bound is within a logarithmic factor of the minimax bound (5).

(b) If (A1) - (A2) hold then Procedure 1 applied to $\mathcal{H}$ in (23) with η defined in (21) yields, for every $n \geq 1$ and $M \geq \sqrt{n}$

$$E||\widehat{F} - f||_2^2 \ \leq \ \inf_{\lambda \in \Lambda_1} ||F_\lambda - f||^2 \ + \ D \sqrt{\left\{\log(1 + M/\sqrt{n})\right\}/n}, \qquad (25)$$

where D is a dominating constant depending on η and b, but independent of M and n. The bound is the minimax bound (5).

Theorem 2 provides another approach to obtaining minimax (CVX) bounds for $M > \sqrt{n}$, in addition to those mentioned in the Introduction.

4 Appendix I: Model Selection

4.1 Proof of Proposition 1

Let $C_n(\widehat{G}) = \sum_{k=0}^{n-1} (Y_{k+1} - \widehat{G}_k(X_{k+1}))^2$ denote the cumulative prediction loss on $Z_1, \ldots, Z_n$ of the pre-estimates defined in (8).

Lemma 1. *Suppose that each estimate $F_j \in \mathcal{F}$ is bounded in absolute value by $b > 0$ and that $|Y_i| \leq \gamma$ for some $\gamma \geq b$. Let $\widehat{G}_k$, $k \geq 0$ be the pre-estimates defined in (8) with $\eta = (2(\gamma + b)^2)^{-1}$. Then for every $n \geq 1$,*

$$C_n(\widehat{G}) \ \leq \ \min_{1 \leq j \leq M} C_n(F_j) \ + \ 2\,(\gamma + b)^2 \, \ln M \qquad (26)$$

with probability one.

Proof: We follow the arguments of Feder and Singer [27, 28], and Kivinen and Warmuth [20]. For $k = 0, \ldots, n$ define $V_k = \sum_{1 \leq j \leq M} \exp\{-\eta \, C_k(F_j)\}$. Then note that

$$\frac{V_{k+1}}{V_k} \ = \ \sum_{1 \leq j \leq M} w_k(F_j) \, e^{-\eta \, (Y_{k+1} - F_j(X_{k+1}))^2} \ \leq \ e^{-\eta \, (Y_{k+1} - \widehat{G}(X_{k+1}))^2} \qquad (27)$$

where the equality follows from (7), and the inequality follows from Jensen's inequality and the concavity of the function $h(u) = e^{-\eta u^2}$ on the interval $[-(\gamma + b), (\gamma + b)]$. Taking logarithms, rearranging terms and summing over k yields

$$C_n(\widehat{G}) \;\leq\; \sum_{k=0}^{n-1} -\frac{1}{\eta} \ln \frac{V_{k+1}}{V_k}. \tag{28}$$

The sum on the right hand side is equal to

$$\begin{aligned} -\frac{1}{\eta} \ln \frac{V_n}{V_0} \;&=\; -\frac{1}{\eta} \ln \left[\frac{1}{M} \sum_{1 \leq j \leq M} e^{-\eta\, C_n(F_j)} \right] \;\leq\; \frac{\ln M}{\eta} \;-\; \frac{1}{\eta} \ln \left[\max_{1 \leq j \leq M} e^{-\eta\, C_n(F_j)} \right] \\ &=\; 2\,(\gamma + b)^2 \ln M \;+\; \min_{1 \leq j \leq M} C_n(F_j). \end{aligned} \tag{29}$$

Combining (28) and (29) gives the result. ■

Following standard arguments, Lemma 1 can be extended to the setting where $\mathcal{F}$ is countable and one has a positive prior probability on each of its constituent estimates. For general bounded convex loss functions one may obtain weaker inequalities analogous to (26) (*cf.* Cesa-Bianchi (1999) and Cesa-Bianchi and Lugosi (2000)).

Proof of Proposition 1: Recall that $\widehat{F}$ is the average of the pre-estimates $\widehat{G}_k$, $k = 0, \ldots, n-1$. Using this definition, its risk can be bounded as follows:

$$\begin{aligned} E(\widehat{F}(X) - Y)^2 \;&=\; E \left(\frac{1}{n} \sum_{k=0}^{n-1} \widehat{G}_k(X) - Y \right)^2 \\ &\leq\; \frac{1}{n} \sum_{k=0}^{n-1} E \left(\widehat{G}_k(X) - Y \right)^2 \qquad (30) \\ &=\; \frac{1}{n} \sum_{k=0}^{n-1} E \left(\widehat{G}_k(X_{k+1}) - Y_{k+1} \right)^2 \qquad (31) \\ &=\; E \left[\frac{1}{n} \sum_{k=0}^{n-1} \left(\widehat{G}_k(X_{k+1}) - Y_{k+1} \right)^2 \right] \\ &=\; \frac{E C_n(\widehat{G})}{n}. \qquad (32) \end{aligned}$$

Here (30) follows from Jensen's inequality, (31) follows from the fact that the pair (X_{k+1}, Y_{k+1}) is independent of $\widehat{G}_k$, and (32) follows from the definition of cumulative loss. Applying

Lemma 1 to the final term above, we see that

$$E(\widehat{F}(X) - Y)^2 \;\leq\; E\left[\min_{1\leq j\leq M} \frac{1}{n}\sum_{k=1}^{n}(F_j(X_k) - Y_k)^2\right] \;+\; \frac{2(\gamma+b)^2 \ln M}{n}$$

$$\leq \min_{1\leq j\leq M} E(F_j(X) - Y)^2 \;+\; \frac{2(b+\gamma)^2 \ln M}{n} \tag{33}$$

As (X, Y) is independent of $Z_1, \ldots, Z_n$, it follows that $E(\widehat{F}(X) - Y)^2 = E||\widehat{F} - f||^2 + E(Y - f(X))^2$, and similarly $E(F_j(X) - Y)^2 = ||F_j - f||^2 + E(Y - f(X))^2$. Subtracting $E(Y - f(X))^2$ from both sides of (33) gives the result. ■

4.2 Proof of Proposition 2

Using the cumulative risk bound of Lemma 1 and the proof of Proposition 1 one may readily establish that

$$\tilde{\Delta} \;=\; E(\tilde{Y} - \tilde{F}(X))^2 \;-\; \min_{1\leq j\leq M} E(\tilde{Y} - F_j(X))^2 \;\leq\; \frac{2(b+\gamma)^2 \ln M}{n}. \tag{34}$$

Moreover, an elementary calculation shows that

$$E(Y - \tilde{F}(X))^2 \;-\; \min_{1\leq j\leq M} E(Y - F_j(X))^2$$
$$\leq\; \tilde{\Delta} \;+\; 2E(Y - \tilde{Y})(\tilde{Y} - \tilde{F}(X)) \;-\; 2\max_{1\leq j\leq M} E(Y - \tilde{Y})(\tilde{Y} - F_j(X)).$$

The definition of $\tilde{Y}$ ensures that the absolute value of each of the final terms is less than $4(b+\gamma)E|Y|\, I\{|Y| \geq \gamma\}$, and the stated bound then follows from (34).

4.3 Proof of Corollary 1

The proof of Corollary 1 follows from Proposition 2 above and part (b) of the following lemma.

Lemma 2. *Assume that there exists a constant $L > 0$ such that $E\exp(|W|) \leq L$. Assume that there exits $b > 0$ such that $\|f\|_\infty \leq b$. Let $c \geq 1$ be a fixed constant. Then*

$$(a) \qquad E\, Y^2 I\{|Y| > 2bc\ln n\} \;\leq\; \frac{C}{n\ln n}$$
$$(b) \qquad E\,|Y| I\{|Y| > 2b\ln n\} \;\leq\; \frac{C_1}{n\ln n},$$

where C is a dominating constant depending on b, c and L but not on n, and C_1 is a dominating constant depending on b and L but not on n.

Proof of Lemma 2. An elementary calculation shows that for any $\gamma_n > 2b$

$$E\,Y^2 I\{|Y| > \gamma_n\} \;\leq\; 2b^2 P(|W| > \gamma_n/2) \,+\, 2E[W^2 I\{|W| > \gamma_n/2\}].$$

Since

$$EW^2 I\{|W| > \gamma_n/2\} \;=\; \frac{\gamma_n^2}{4} P(|W| > \gamma_n/2) + 2\int_{\gamma_n/2}^{\infty} tP(|W| \;> t)dt,$$

and $P(|W| > x) \leq Le^{-x}$, the choice $\gamma_n = 2bc\ln n$ leads to

$$E\,Y^2 I\{|Y| > 2bc\ln n\} \leq 4L\frac{(bc\ln n + 1)^2}{n^{bc}}.$$

Since we can assume, without loss of generality, that $b, c > 1$, we can always find a constant C such that the right hand side of the display above is smaller than $\frac{C}{n\ln n}$, which concludes the proof of the first statement. For the second statement we use similar reasoning to show that for any $\gamma_n \geq 2b$

$$E\,|Y| I\{|Y| > \gamma_n\} \;\leq\; bP(|W| > \gamma_n/2) \,+\, E|W| I\{|W| > \gamma_n/2\}.$$

This, combined with

$$E|W| I\{|W| > \gamma_n/2\} \;=\; \frac{\gamma_n}{2} P(|W| > \gamma_n/2) + \int_{\gamma_n/2}^{\infty} P(|W| \;> t)dt$$

yields the desired result, for $\gamma_n = 2b\ln n$. ■

4.4 Proof of Proposition 3

It follows directly from Lemma 1 and the stationarity of the observations Z_i that

$$\begin{aligned}
&\sum_{k=1}^{n} E(Y_k - \hat{G}_{k-1}(X_k))^2 \\
&\quad\leq\; n \min_{1\leq j\leq M} E(Y - F_j)^2 \,+\, 2\,(\gamma + b)^2\,\ln M &(35)\\
&\quad=\; n \min_{1\leq j\leq M} E||f - F_j||^2 \,+\, n\,E(f(X) - Y)^2 \,+\, 2\,(\gamma + b)^2\,\ln M &(36)
\end{aligned}$$

Consider the summation on the left hand side of (36). Adding and subtracting $f(X_k)$ and expanding the square, the kth summand is equal to

$$E(f(X_k) - \hat{G}_{k-1}(X_k))^2 \,+\, E(Y_k - f(X_k))^2 \,+\, E(Y_k - f(X_k))(f(X_k) - \hat{G}_{k-1}(X_k)). \quad (37)$$

The second term above is simply $E(Y - f(X))^2$. Letting $U_k = (f(X_k) - \hat{G}_{k-1}(X_k))$, the third term above is equal to the expected value of

$$\begin{aligned} E[U_k\,(Y_k - f(X_k))\,|\,X_k, Z^{k-1}] &= U_k\,E[(Y_k - f(X_k))\,|\,X_k, Z^{k-1}] \\ &= U_k\,E[\varepsilon_k\,|\,X_k, Z^{k-1}] \\ &= U_k\,E[\varepsilon_k\,|\,\varepsilon_1, \ldots, \varepsilon_{k-1}] \qquad (38) \\ &= 0. \qquad (39) \end{aligned}$$

Here (38) follows from the independence of $\{\varepsilon_i\}$ and $\{X_i\}$, and (39) is a consequence of the martingale difference property (14). Combining (39) with (36) and (37) yields the result.

5 Appendix II: Proof of Theorem 1

The proof of Theorem 1 follows from the lemmas stated and proved below. Let $C_n(\widetilde{G}) = \sum_{k=0}^{n-1}(Y_{k+1} - \widetilde{G}_k(X_{k+1}))^2$ denote the cumulative prediction loss on $Z_1, \ldots, Z_n$ of the pre-estimates defined in (19).

Lemma 3. *Let $\widehat{F}$ be defined as in Procedure 2. Then*

$$E(\widehat{F}(X) - Y)^2 \le \frac{EC_n(\widetilde{G})}{n}.$$

Lemma 4. *Let $\widehat{F}$ be defined as in Procedure 2, with η satisfying (21). Assume (A1) - (A2) hold. Then*

$$\frac{EC_n(\widetilde{G})}{n} \le -\frac{1}{n\eta}E\left(\ln\int_\Lambda e^{-\eta\,C_n(F_\lambda)}\right) + \frac{M\ln(2c)}{n\eta} - \frac{\ln M!}{n\eta}.$$

Lemma 5.

$$\begin{aligned} -\frac{1}{n\eta}E\left(\ln\int_\Lambda e^{-\eta\,C_n(F_\lambda)}\right) \;\le\; & \min_{\lambda\in\Lambda} E(Y - F_\lambda(X))^2 + \frac{\ln M!}{n\eta} + \frac{M\ln(Bn\ln n)}{n\eta} + \frac{M+1}{n\eta} \\ & + 2E[Y^2\,I\{|Y| > 2bc\ln n\}], \end{aligned}$$

for $B = 8b^2c\eta$.

Proof of Theorem 1 (a)**.** The application of Lemmas 3, 4, 5 and 2 (a) immediately yields

$$E(\widehat{F}(X) - Y)^2 \;\le\; \min_{\lambda\in\Lambda} E(F_\lambda(X) - Y)^2 + \frac{B_1 M\ln(Bn\ln n)}{n\eta}. \qquad (40)$$

As (X, Y) is independent of $Z_1, \ldots, Z_n$, it follows that $E(\widehat{F}(X) - Y)^2 = E||\widehat{F} - f||^2 + E(Y - f(X))^2$, and similarly $E(F_\lambda(X) - Y)^2 = ||F_\lambda - f||^2 + E(Y - f(X))^2$. Subtracting $E(Y - f(X))^2$ from both sides of (40) gives the result. ■

Proof of Lemma 3: The proof is analogous to that of Proposition 1 above.

The proof of Lemma 4 requires additional notation and uses the result stated below. We first introduce the notation. Denote by E_k and Var_k the expectation and variance, respectively, with respect to the distribution $w_k(\lambda)$ defined in (18). For any $0 \leq \gamma \leq \eta$ we also consider the following distribution over Λ:

$$v_{k,\gamma}(\lambda) \;=\; \frac{\exp\{-(\eta+\gamma)C_k(F_\lambda)\}}{\int_\Lambda \exp\{-(\eta+\gamma)C_k(F_{\lambda'})\}\, d\mu(\lambda')}. \tag{41}$$

We denote by $E_{k,\gamma}$ the expected value under this distribution. Define the function

$$h(\lambda) = -[Y_{k+1} - F_\lambda(X_{k+1})]^2.$$

Define further

$$\mathrm{Var}_{k,\gamma}(h(\lambda)) = E_{k,\gamma}\left(h(\lambda) - E_{k,\gamma}(h(\lambda))\right)^2,$$

$$\mathrm{M}^3_{k,\gamma}(h(\lambda)) = E_{k,\gamma}\left(h(\lambda) - E_{k,\gamma}(h(\lambda))\right)^3.$$

and

$$V = 0 \vee \sup_{0\leq\gamma\leq\eta} \frac{\mathrm{M}^3_{k,\gamma}(h(\lambda))}{\mathrm{Var}_{k,\gamma}(h(\lambda))}. \tag{42}$$

Lemma 6.

$$\begin{aligned}\ln \int_\Lambda w_k(\lambda) \exp\left(-\eta[Y_{k+1} - F_\lambda(X_{k+1})]^2\right) d\lambda \;\;\leq\;\; & -\eta \int_\Lambda w_k(\lambda)[Y_{k+1} - F_\lambda(X_{k+1})]^2 \\ & + \frac{\eta^2}{2} e^{\eta V}\, Var_k(h(\lambda)) d\lambda.,\end{aligned}$$

Proof of Lemma 6. The proof follows directly from Lemma 3.6.1, page 85, in Catoni (2000). It suffices to replace $\nu(d\theta)$ by $w_k(\lambda)d\lambda$, $\nu_\gamma(d\theta)$ by $v_{k,\gamma}(\lambda)d\lambda$, and λ by our η. ■

Proof of Lemma 4. The proof of this lemma combines arguments of Lemma 1 above with those of Theorem 3.6.1 in Catoni (2000). The latter are needed since we can no longer use Jensen's inequality without truncating Y.

First notice that by adding and subtracting $\widetilde{G}_k$ in the bracket under the integral below we obtain

$$\int_\Lambda w_k(\lambda)[Y_{k+1} - F_\lambda(X_{k+1})]^2 d\lambda = [Y_{k+1} - \widetilde{G}_k(X_{k+1})]^2 + \mathrm{Var}_k(F_\lambda), \tag{43}$$

since

$$\widetilde{G}_k(x) \;=\; \int_\Lambda F_\lambda(x)\, w_k(\lambda)\, d\lambda = E_k(F_\lambda).$$

Therefore, by Lemma 6, we have

$$\begin{aligned}[Y_{k+1} - \widetilde{G}_k(X_{k+1})]^2 &\leq -\frac{1}{\eta}\ln\int_\Lambda w_k(\lambda)\exp\left(-\eta[Y_{k+1} - F_\lambda(X_{k+1})]^2\right)d\lambda - \mathrm{Var}_k(F_\lambda) \\ &\quad + \frac{\eta}{2}e^{\eta V}\mathrm{Var}_k(h(\lambda)).\end{aligned}$$

Summing over k and recalling the notation $C_n(\widetilde{G}) = \sum_{k=0}^{n-1}(Y_{k+1} - \widetilde{G}_k(X_{k+1}))^2$ we further obtain

$$\begin{aligned}C_n(\widetilde{G}) &\leq -\frac{1}{\eta}\sum_{k=0}^{n-1}\ln\int_\Lambda w_k(\lambda)\exp\left(-\eta[Y_{k+1} - F_\lambda(X_{k+1})]^2\right)d\lambda \\ &\quad + \sum_{k=0}^{n-1}\left(\frac{\eta}{2}e^{\eta V}\mathrm{Var}_k(h(\lambda)) - \mathrm{Var}_k(F_\lambda)\right). \qquad (44)\end{aligned}$$

Defining $W_k = \int_\Lambda \exp\left(-\eta\, C_k(F_\lambda)\right)\, d\lambda$ we notice that

$$\frac{W_{k+1}}{W_k} = \int_\Lambda w_k(\lambda)\, \exp(-\eta\left(Y_{k+1} - F_\lambda(X_{k+1})\right)^2)\, d\lambda,$$

and so

$$\sum_{k=0}^{n-1}\ln\int_\Lambda w_k(\lambda)\, \exp(-\eta\left(Y_{k+1} - F_\lambda(X_{k+1})\right)^2)\, d(\lambda) = \sum_{k=0}^{n-1}\ln\frac{W_{k+1}}{W_k} = \ln\frac{W_n}{W_0}. \qquad (45)$$

Since $W_0 = \frac{(2c)^M}{M!}$, displays (44) and (45) thus yield

$$\begin{aligned}E\left(\frac{C_n(\widetilde{G})}{n}\right) &\leq -\frac{1}{n\eta}E\ln\int_\Lambda \exp(-\eta\, C_n(F_\lambda))d\lambda + \frac{M\ln(2c)}{n\eta} - \frac{\ln M!}{n\eta} \\ &\quad + \frac{1}{n}\sum_{k=0}^{n-1}E\left(\frac{\eta}{2}e^{\eta V}\mathrm{Var}_k(h(\lambda)) - \mathrm{Var}_k(F_\lambda)\right). \qquad (46)\end{aligned}$$

A calculation as in Theorem 3.6.1 page 87 of Catoni (2000) shows that

$$\mathrm{Var}_k(h(\lambda)) \leq [15b^2c^2 + 3(Y_{k+1} - f(X_{k+1}))^2]\mathrm{Var}_k(F_\lambda). \qquad (47)$$

The last term in (46) is therefore at most

$$\frac{1}{n}\sum_{k=0}^{n-1}E\left\{\left(\frac{\eta}{2}e^{\eta V}[15b^2c^2 + 3(Y_{k+1} - f(X_{k+1}))^2] - 1\right)\mathrm{Var}_k(F_\lambda)\right\}.$$

To complete the proof of the lemma we show now that, for η defined in (21), the last term in the display above is negative. Let $Z_k = ((X_1, Y_1), \ldots, (X_k, Y_k))$. Then Z_k is independent of (X_{k+1}, Y_{k+1}), and $\mathrm{Var}_k(F_\lambda)$ is independent from the factor multiplying it. Therefore, dropping the subscript $k+1$ for clarity of notation, we only need to show that

$$E\left(\frac{\eta}{2}e^{\eta V}[15b^2c^2 + 3(Y - f(X))^2] - 1\right) \leq 0.$$

Recall the definition of V in (42) and let $Z(\lambda) = (h(\lambda) - E_{k,\gamma}(h(\lambda)))$. Then

$$\mathrm{M}^3_{k,\gamma}(h(\lambda)) \leq \sup_{\lambda\in\Lambda} |Z(\lambda)| \times \mathrm{Var}_{k,\gamma}(h(\lambda))$$

which implies $V \leq \sup_{\lambda\in\Lambda} |Z(\lambda)|$. Now note that

$$h(\lambda) = -[Y - f(X)]^2 - [f(X) - F_\lambda(X)]^2 - 2[Y - f(X)] \times [f(X) - F_\lambda(X)].$$

Since $E_{k,\gamma}(-[Y - f(X)]^2) = -[Y - f(X)]^2$ and $W = Y - E(Y|X) = Y - f(X)$, we obtain the bound

$$V \leq \sup_{\lambda\in\Lambda} |Z(\lambda)| \leq 4b^2c^2 + 8bc|Y - f(X)| = 4b^2c^2 + 8bc|W|. \tag{48}$$

Thus

$$E\left(\frac{\eta}{2}e^{\eta V}[15b^2c^2 + 3(Y - f(X))^2] - 1\right) \leq E\left(\frac{\eta}{2}e^{4\eta b^2c^2}e^{8bc\eta|W|}[15b^2c^2 + 3W^2] - 1\right).$$

Choosing η as in (21) makes the display above negative, which completes the proof of the lemma. ■.

Proof of Lemma 5. Let $\gamma_n = 2bc \ln n$. Define a bounded version of Y_k by thresholding its value at $+\gamma_n$ and $-\gamma_n$:

$$\tilde{Y}_k \;=\; \begin{cases} -\gamma_n & \text{if } Y_k < -\gamma_n \\ Y_k & \text{if } -\gamma_n \leq Y_k \leq \gamma_n \\ \gamma_n & \text{if } Y_k > \gamma_n. \end{cases}$$

For $v \in [-\gamma_n, \gamma_n]$ and re-denoting, for clarity, Y_k by Y we have

$$\begin{aligned} 0 \;\leq\; (Y - v)^2 \;-\; (\tilde{Y} - v)^2 \;&=\; (Y^2 - \tilde{Y}^2) \;-\; 2v(Y - \tilde{Y}) \\ &=\; (Y - \tilde{Y})\,(Y + \tilde{Y} - 2v) \\ &=\; |Y - \tilde{Y}|\,|Y + \tilde{Y} - 2v| \\ &\leq\; |Y - \tilde{Y}|\,(|Y| + |\tilde{Y}| + 2|v|) \\ &\leq\; |Y - \tilde{Y}|\,(2|Y| + 2|\tilde{Y}|) \\ &=\; (|Y| - |\tilde{Y}|)\,(2|Y| + 2|\tilde{Y}|) \\ &=\; 2(|Y|^2 - |\tilde{Y}|^2) \\ &=\; 2(|Y|^2 - |\tilde{Y}|^2)\,I\{|Y| > \gamma_n\} \\ &=\; 2|Y|^2\,I\{|Y| > \gamma_n\} \end{aligned} \tag{49}$$

Define $\tilde{C}_n(F_\lambda) = \sum_{k=1}^n (\tilde{Y}_k - F_\lambda(X_k))^2$. By inequality (49) with $v = F_\lambda(X) \in [-\gamma_n, \gamma_n]$, for each k,

$$(Y_k - F_\lambda(X_k))^2 \;-\; (\tilde{Y}_k - F_\lambda(X_k))^2 \;\leq\; 2Y_k^2\,I\{|Y_k| > \gamma_n\}.$$

Summing over k we obtain

$$C_n(F_\lambda) - \tilde{C}_n(F_\lambda) \;\le\; 2\sum_{k=1}^{n} Y_k^2\, I\{|Y_k| > \gamma_n\}.$$

Consequently

$$\begin{aligned} -\frac{1}{n\eta}\ln\int_\Lambda e^{-\eta\, C_n(\mathsf{F}_\lambda)}\, d\mu(\lambda) &= -\frac{1}{n\eta}\ln\int_\Lambda \frac{e^{-\eta\, C_n(\mathsf{F}_\lambda)}}{e^{-\eta\, \tilde{C}_n(\mathsf{F}_\lambda)}}\, e^{-\eta\, \tilde{C}_n(F_\lambda)}\, d\mu(\lambda) \\ &\le \frac{2}{n}\sum_{k=1}^{n} Y_k^2\, I\{|Y_k| > \gamma_n\} \;-\; \frac{1}{n\eta}\ln\int_\Lambda e^{-\eta\, \tilde{C}_n(F_\lambda)}\, d\mu(\lambda). \quad (50) \end{aligned}$$

In what follows we evaluate the integral in the right hand side of (50). Let $H(\lambda) = \exp(-\eta\, \tilde{C}_n(F_\lambda))$. As H is continuous on the compact set Λ, there exists $\lambda^* \in \Lambda$ such that $H(\lambda^*) = \max_{\lambda\in\Lambda} H(\lambda)$. Let

$$\begin{aligned} D &= \{\lambda \in \Lambda : H(\lambda) \ge \frac{1}{2} H(\lambda^*)\} \\ &= \left\{\lambda \in \Lambda : \tilde{C}_n(F_\lambda) - \tilde{C}_n(F_{\lambda^*}) \le \frac{\ln 2}{\eta}\right\} \\ &= \left\{\lambda \in \Lambda : \sum_{k=1}^{n} (F_{\lambda^*}(X_k) - F_\lambda(X_k)) \left(2\tilde{Y}_k - F_\lambda(X_k) - F_{\lambda^*}(X_k)\right) \le \frac{\ln 2}{\eta}\right\}. \end{aligned}$$

Denote the summand above by $A_{k,\lambda}$. Recall that since we fixed $\gamma_n > bc$ we have $|F_{\lambda^*}| \le bc \le \gamma_n$, $|F_\lambda| \le bc \le \gamma_n$ and $|\tilde{Y}_k| \le \gamma_n$. Notice then that

$$\sum_{k=1}^{n} A_{k,\lambda} \;\le\; \sum_{k=1}^{n} |A_{k,\lambda}| \le 4\gamma_n \sum_{k=1}^{n}\sum_{j=1}^{M} |\lambda_j - \lambda_j^*||F_j(X_k)| \le 4bn\gamma_n \sum_{j=1}^{M} |\lambda_j - \lambda_j^*|.$$

Define

$$\begin{aligned} A &= \left\{\lambda \in \Lambda : \sum_{j=1}^{M} |\lambda_j - \lambda_j^*| \le \frac{\ln 2}{4b\eta\gamma_n n}\right\} \\ &= \Lambda \cap \left\{\lambda \in R^M : \sum_{j=1}^{M} |\lambda_j - \lambda_j^*| \le \frac{\ln 2}{4b\eta\gamma_n n}\right\} \;=:\; \Lambda \cap B, \end{aligned}$$

and observe that $A \subseteq D$, and so $\mu(D) \ge \mu(A)$, where $\mu(D)$ denotes the Lebesgue measure of D. Then

$$\begin{aligned} \ln\int_\Lambda e^{-\eta\, \tilde{C}_n(F_\lambda)}\, d\mu(\lambda) &= \ln\int_\Lambda H(\lambda)\, d\mu(\lambda) \;\ge\; \ln\int_D H(\lambda)\, d\mu(\lambda) \quad (51) \\ &\ge \ln\left(\frac{1}{2} H(\lambda^*)\mu(D)\right) \;=\; -\eta \min_{\lambda\in\Lambda} \tilde{C}_n(F_\lambda) \;-\; \ln 2 \;+\; \ln\mu(D) \\ &\ge \;\; -\eta \min_{\lambda\in\Lambda} \tilde{C}_n(F_\lambda) \;-\; \ln 2 \;+\; \ln\mu(A). \end{aligned}$$

We evaluate now $\mu(A)$. It is enough to consider the extreme case when λ^* is one of the vertices of the generalized simplex Λ. The set A is then the intersection of the generalized simplex Λ and the generalized simplex B which is centered at some vertex of Λ. The intersection is therefore non-void and is a new generalized simplex, whose volume is the volume of B divided by 2^M. In this case

$$\mu(A) = \frac{(C_n)^M}{M!}.$$

Thus

$$\mu(A) \geq \frac{(C_n)^M}{M!}, \quad \text{for } C_n = \frac{\ln 2}{4b\eta\gamma_n n}. \tag{52}$$

Combining now (50), (51) and (52) we obtain

$$\begin{aligned} -\frac{1}{n\eta}\ln\int_\Lambda e^{-\eta\, C_n(\mathsf{F}_\lambda)}\, d\mu(\lambda) \;\;\leq\;\; & \frac{1}{n}\min_{\lambda\in\Lambda}\tilde{C}_n(F_\lambda) + \frac{2}{n}\sum_{k=1}^{n} Y_k^2\, I\{|Y_k| > \gamma_n\} \\ & + \frac{\ln M!}{n\eta} + \frac{M\ln(4b\eta n\gamma_n)}{n\eta} - \frac{M\ln\ln 2}{n\eta} + \frac{\ln 2}{n\eta}. \end{aligned} \tag{53}$$

The sum of the last two terms is bounded above by $\frac{M+1}{n\eta}$. Invoke again inequality (49) to see that for each k

$$(Y_k - F_\lambda(X_k))^2 \;\geq\; (\tilde{Y}_k - F_\lambda(X_k))^2,$$

so that $\tilde{C}_n(F_\lambda) \leq C_n(F_\lambda)$. To conclude the proof of this part, take expectations throughout.

Proof of Theorem 1 (b)**.** The proof of this part is a much simplified version of (a). It follows from the following lemmas. Let $\tilde{Y}, \tilde{Y}_k$ be given by (11), for $\gamma_n = 2bc\ln n$. In what follows $\widehat{F}$ denotes the aggregate obtained by applying Procedure 2 to $\tilde{Y}_k$ with $\eta = (2(\gamma_n + bc)^2)^{-1}$. Let $\tilde{C}_n(\widetilde{G}) = \sum_{k=0}^{n-1}(\tilde{Y}_{k+1} - \widetilde{G}_k(X_{k+1}))^2$. Define $\tilde{C}_n(F_\lambda) = \sum_{k=1}^{n}(\tilde{Y}_k - F_\lambda(X_k))^2$.

Lemma 7. *Under the assumptions of Lemma 2 and for a positive constant C independent of n*

$$E(Y - \hat{F}(X))^2 - \min_{\lambda\in\Lambda(c)} E(Y - F_\lambda(X))^2 \leq E(\tilde{Y} - \hat{F}(X))^2 - \min_{\lambda\in\Lambda(c)} E(\tilde{Y} - F_\lambda(X))^2 + \frac{C}{n}.$$

Proof. The result follows immediately from elementary calculations as in Proposition 2 and the definition of $\tilde{Y}$.

Lemma 8.

$$E(\widehat{F}(X) - \tilde{Y})^2 \leq \frac{E\tilde{C}_n(\widetilde{G})}{n}.$$

Proof. The proof is analogous to that of Proposition 1.

Lemma 9.

$$\frac{E\tilde{C}_n(\widetilde{G})}{n} \leq -\frac{1}{n\eta}E\left(\ln\int_\Lambda e^{-\eta\, C_n(F_\lambda)}\right) + \frac{M\ln(2c)}{n\eta} - \frac{\ln M!}{n\eta}.$$

Proof. We use same arguments as in Lemma 1 up to and including (28), where now we use the concavity of the function $h(u) = e^{-\eta u^2}$ on the interval $[-(\gamma_n + bc), (\gamma_n + bc)]$, as $|F_\lambda| \leq b\, c \leq \gamma$ for $\lambda \in \Lambda(c)$. To conclude the proof we recall that $W_0 = \frac{(2c)^M}{M!}$.

Remark. Note that for this choice of η we no longer require the implied conditions (A3) and (A4) used in Lemma 4 above.

Lemma 10.

$$-\frac{1}{n\eta}E\left(\ln\int_\Lambda e^{-\eta\,\tilde{C}_n(F_\lambda)}\right) \;\leq\; \min_{\lambda\in\Lambda} E(\tilde{Y} - F_\lambda(X))^2 + \frac{\ln M!}{n\eta} + \frac{M\ln(Bn\ln n)}{n\eta} + \frac{M+1}{n\eta}.$$

for $B = 8b^2c\eta$.

Proof. The proof is identical to that of Lemma 5 applied directly to $\tilde{Y}_k$. This completes the proof of this part and of the theorem. ■

Remark. An alternative proof of Theorem 1 (b) can be obtained by adapting Theorem 3.3 of Section 3 in Cesa-Bianchi and Lugosi [10] to this set up. This would yield the analogue of Lemma 1 for Procedure 2. The conclusion then follows by reasoning as in Propositions 1 and 2.

6 Appendix III: Proof of Theorem 2

Proof of Theorem 2 (a). Let $\lambda^* \in \Lambda_1$ be the value of λ that achieves $\inf_{\lambda\in\Lambda_1} ||F_\lambda - f||^2$. Define a probability mass function on the set $\{0, 1, \ldots, M\}$ as

$$\pi_j = \begin{cases} \lambda_j^* & \text{if } j \neq 0, \\ 1 - \sum_{j=1}^M \lambda_j^* & \text{if } j = 0. \end{cases}$$

Let H be a function chosen at random from the base functions $\mathcal{F}' = \{F_0, F_1, \ldots, F_M\}$ according to the distribution $P(H = F_j) = \pi_j$, $j = 0, \ldots, M$. Let $H^1, \ldots, H^m$ be independent copies of H and define the random function

$$G \;=\; \frac{1}{m}\sum_{i=1}^m H^i.$$

As each H^i is supported on the collection $\mathcal{F}'$, it is easy check that $P\{G \in \mathcal{H}\} = 1$. Let E' denote expectation with respect to the joint distribution of $H^1, \ldots, H^m$. For fixed $x \in \mathbb{R}$,

$$E'G(x) \;=\; E'H(x) \;=\; \sum_{j=1}^M \lambda_j^*\, F_j(x) \;=\; F_{\lambda^*}(x). \tag{54}$$

As a first step towards the desired convex aggregation bound, we use the argument of Nemirovski [25] to show that

$$E'||G - f||^2 \;\leq\; \inf_{\lambda \in \Lambda_1} ||F_\lambda - f||^2 \;+\; \frac{b^2}{m}. \tag{55}$$

To begin, note that for each x,

$$\begin{aligned} E'(G(x) - f(x))^2 &= E'(G(x) - F_{\lambda^*}(x))^2 \;+\; E'(f(x) - F_{\lambda^*}(x))^2 \\ &\quad +\, 2\, E'(G(x) - F_{\lambda^*}(x))\,(f(x) - F_{\lambda^*}(x)) \\ &= E'(G(x) - F_{\lambda^*}(x))^2 \;+\; (f(x) - F_{\lambda^*}(x))^2, \end{aligned} \tag{56}$$

where the last equality follows from (54) and the fact that f and F_{λ^*} are fixed under $E'(\cdot)$.

Consider the first term in (56). Let $\mathrm{Var}'(\cdot)$ denote the variance with respect to the joint distribution of $H^1, \ldots, H^m$. Then, as $E'H^i(x) = F_{\lambda^*}(x)$ and $H^1, \ldots, H^m$ are independent,

$$\begin{aligned} E'(G(x) - F_{\lambda^*}(x))^2 &= E'\left(\frac{1}{m}\sum_{i=1}^m (H^i(x) \;-\; F_{\lambda^*}(x))\right)^2 \\ &= \frac{1}{m^2} E'\left(\sum_{i=1}^m (H^i(x) \;-\; F_{\lambda^*}(x))\right)^2 \\ &= \frac{1}{m^2}\mathrm{Var}'\left(\sum_{i=1}^m H^i(x)\right) \;=\; \frac{1}{m^2}\sum_{i=1}^m \mathrm{Var}'(H_i(x)) \\ &= \frac{1}{m}\mathrm{Var}'(H(x)) \;\leq\; \frac{1}{m} E'H^2(x) \;\leq\; \frac{b^2}{m}. \end{aligned} \tag{57}$$

Combining the last inequality with (56), it follows that

$$E'(G(X) - f(X))^2 \;\leq\; \frac{b^2}{m} \;+\; (f(X) - F_{\lambda^*}(X))^2.$$

with probability one. Taking expectations of both sides of this inequality over X, and applying Fubini's theorem, we find that

$$E'||G - f||^2 \;\leq\; \frac{b^2}{m} \;+\; ||f - F_{\lambda^*}||^2.$$

By the definition of λ^*, the last inequality is equivalent to (55).

Now let $\widehat{F}$ be the aggregate estimator defined as in the statement of the theorem. Based on the arguments above, we may bound $E||\widehat{F} - f||^2$ as follows.

$$
\begin{aligned}
E||\widehat{F} - f||^2 &\leq \min_{F \in \mathcal{H}} ||F - f||^2 + \frac{2(\gamma+b)^2 \ln|\mathcal{H}|}{n} + \frac{C}{n} && (58)\\
&\leq E'||G - f||^2 + \frac{2(\gamma+b)^2 \ln|\mathcal{H}|}{n} + \frac{C}{n} && (59)\\
&\leq \inf_{\lambda \in \Lambda_1} ||F_\lambda - f||^2 + \frac{b^2}{m} + \frac{2(\gamma+b)^2 \ln|\mathcal{H}|}{n} + \frac{C}{n} && (60)\\
&\leq \inf_{\lambda \in \Lambda_1} ||F_\lambda - f||^2 + \frac{b^2}{m} + \frac{2(\gamma+b)^2 m}{n}\left[1 + \ln\left(1 + \frac{M}{m}\right)\right] + \frac{C}{n}\\
&\leq \inf_{\lambda \in \Lambda_1} ||F_\lambda - f||^2 + 4(\gamma+b)^2\left[\frac{1}{m} + \frac{m}{n}\ln\left(1 + \frac{M}{m}\right)\right] + \frac{C}{n} && (61)\\
&\leq \inf_{\lambda \in \Lambda_1} ||F_\lambda - f||^2 + 8(\gamma+b)^2\left[\frac{1}{x_{n,M}} + \frac{x_{n,M}}{n}\ln\left(1 + \frac{M}{x_{n,M}}\right)\right] + \frac{C}{n}.
\end{aligned}
$$

Here the inequality (58) follows from Corollary 1 applied to $\mathcal{H}$, (59) is a consequence of the fact that G is supported on $\mathcal{H}$, and (60) is an application of the bound (55). Stirling's formula implies that $|\mathcal{H}| = \binom{M+m}{m} \leq \left[\frac{e(m+M)}{m}\right]^m$, so that $\ln|\Lambda| \leq m(1 + \ln(1 + M/m))$. As $M > \sqrt{n}$ we have $m \leq \alpha_{n,M} \leq M/2$ and inequality (61) holds since $\ln(1 + \frac{M}{m}) > 1$ for any such m. To establish the last inequality, we note that $m \geq \alpha_{n,M}/2$, for $\alpha_{n,M} \geq 1$ and apply the inequality $\ln(1 + 2y) \leq 2\ln(1 + y)$, which holds for $y \geq 1$. To complete the proof, we follow in Corollary 3.4 in Bunea, Tsybakov and Wegkamp [5] and make use of the elementary inequality $\ln\left(1 + y\sqrt{\frac{\ln(1+y)}{\ln 2}}\right) \leq 2\ln(1 + y)$, for all $y \geq 1$. This yields

$$
\begin{aligned}
\ln\left(1 + \frac{M}{\alpha_{n,M}}\right) &\leq \ln\left(1 + 2\frac{M}{\sqrt{n}}\sqrt{\frac{\ln(1 + \frac{M}{\sqrt{n}})}{\ln 2}}\right)\\
&\leq 2\ln\left(1 + \frac{M}{\sqrt{n}}\sqrt{\frac{\ln(1 + \frac{M}{\sqrt{n}})}{\ln 2}}\right) \leq 4\ln\left(1 + \frac{M}{\sqrt{n}}\right).
\end{aligned}
$$

Combining the last display with the definition of $\alpha_{n,M}$ we obtain

$$
E||\widehat{F} - f||^2 \leq \inf_{\lambda \in \Lambda_1} ||F_\lambda - f||^2 + 16(\gamma+b)^2\sqrt{\frac{1}{n}\log\left(1 + \frac{M}{\sqrt{n}}\right)} + \frac{C}{n},
$$

which completes the proof of this part of the theorem.

Proof of Theorem 2 (b). The proof of this part is almost identical to that of Part (a). The only difference occurs in (58), where we now invoke Theorem 3.6.1 page 85 in Catoni (2000) to obtain

$$
E||\widehat{F} - f||^2 \leq \min_{F \in \mathcal{H}} ||F - f||^2 + \frac{D_1 \ln|\Lambda|}{n},
$$

where D_1 is a constant independent of n given by his theorem. Consequently, erasing the term $\frac{C}{n}$ throughout and replacing $2(\gamma+b)^2$ by D_1 and later by an appropriate multiplicative constant D we obtain the desired result. This concludes the proof of this theorem. ■

References

[1] J.-Y. Audibert, "Aggregated estimators and empirical complexity for least square regression," *Annales de l'Institut Henri Poincaré (B), Probability and Statistics*, 40: 685-736, 2003.

[2] S. Azoury and M. Warmuth, "Relative loss bounds for on-line density estimation with the exponential family of distributions," *Machine Learning*, 43: 211-245, 2001.

[3] A.R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information Theory*, 39: 930 – 945, 1993.

[4] L. Birgé, "Model selection via testing: an alternative to (penalized) maximum likelihood estimators," *Annales de l'Institut Henri Poincare (B) Probability and Statistics* , vol. 42, pp. 273 - 325, 2006.

[5] F. Bunea, A. B. Tsybakov and M. H. Wegkamp "Aggregation for Gaussian regression," *Annals of Statistics*, vol 35, pp. 1674 - 1697, 2007.

[6] N. Cesa-Bianchi, "Analysis of two gradient-based algorithms for on-line regression," *Proc. 12'th Annual Conference on Computational Learning Theory*, pp.163-170, ACM Press, 1999.

[7] N. Cesa-Bianchi, Y. Freund, D. Haussler, D.P. Helmbold, R.E. Schapire, and M.K. Warmuth, "How to use expert advice," *J. Assoc. Comp. Mach.*, vol. 44, pp. 427-485, 1997.

[8] N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, and M.K. Warmuth, "On-line prediction and conversion strategies," *Machine Learning*, vol. 25, pp. 71-110, 1996.

[9] N. Cesa-Bianchi and G. Lugosi, "On the prediction of individual sequences," *Ann. Statist.*, vol.27, pp.1865-1895, 2000.

[10] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning and games.* Cambridge University Press, 2006.

[11] O. Catoni, *Statistical Learning Theory and Stochastic Optimization.* Ecole d'Eté de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics, vol. 1851, Springer, New York, 2001.

[12] D.P. Foster and R. Vohra, "Regret in the on-line decision problem," *Games and Economic Behavior*, vol. 29, pp. 1084-1090, 1999.

[13] Y. Freund, "Predicting a binary sequence almost as well as the optimal biased coin," *Proc. 9th Annual Conference on Computational Learning Theory, pp 89-98.* ACM Press, New York, NY, 1996.

[14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin *Bayesian Data Analysis*. Chapman Hall/CRC 2003, Second edition, 2003.

[15] L. Györfi and G. Lugosi, "Strategies for sequential prediction of stationary time series," in *Modeling Uncertainty: an Examination of its Theory, Methods, and Applications*, M. Dror, P. L'Ecuyer and F. Szidarovszky Eds, Kluwer, 2001.

[16] L. Györfi, G. Lugosi, and G. Morvai, "A simple randomized algorithm for consistent sequential prediction of ergodic time series," *IEEE Trans. Info. Theory*, vol. 45, pp. 2642-2650, 1999.

[17] D.H. Haussler, J. Kivenen, and M.K. Warmuth, "Sequential prediction of individual sequences under general loss functions," *IEEE Trans. Info. Theory*, vol. 44, pp. 1906-1925, 1998.

[18] A. Juditsky, A. Nazin, A.B. Tsybakov and N. Vayatis, "Online aggregation with a mirror descent algorithm," *Problems of Information Transmission*, vol. 41, pp. 368- 384, 2005.

[19] A. Juditsky and A. Nemirovski, "Functional aggregation for nonparametric regression," *Annals of Statistics*, 28:681–712, 2000.

[20] J. Kivinen and M.K. Warmuth, "Averaging expert predictions," *Proceedings of the Fourth European Conference on Computational Learning Theory (EuroCOLT99)*, H.U. Simon and P. Fischer Eds, pp.153-167, Springer, Berlin, 1999.

[21] V. Koltchinskii, "Local Rademacher complexities and oracle inequalities in risk minimization," *Preprint*, 2004.

[22] G. Leung and A.R. Barron, "Information theory and mixing least-squares regressions," *IEEE Trans. Info. Theory*, vol. 52, 2006.

[23] N. Littlestone and M.K. Warmuth, "The weighted majority algorithm," *Info. and Comput.*, vol. 108, pp. 212-261, 1994.

[24] N. Merhav and M. Feder, "Universal Prediction," *IEEE Trans. Info. Theory*, vol. 44, pp. 2124-2147, 1998.

[25] A. Nemirovski, *Topics in non-parametric statistics.* In P. Bernard, editor, *Ecole d'Eté de Probabilités de Saint-Flour 1998*, volume XXVIII of *Lecture Notes in Mathematics*. Springer, New York, 2000.

[26] A.B. Nobel, "On optimal sequential decision schemes for general processes," *IEEE Transactions on Information Theory*, vol. 49, pp.83-98, 2003.

[27] A. Singer and M. Feder, "Universal linear prediction by model order weighting," *IEEE Transactions on Signal Processing*, vol. 47, pp.2685-2699, 1999.

[28] A. Singer and M. Feder, "Universal linear least squares prediction," *Proceedings of 2000 IEEE International Symposium on Information Theory*, 2000.

[29] A.B. Tsybakov, "Optimal rates of aggregation," *Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines. Lecture Notes in Artificial Intelligence*, v. 2777, p.303–313, 2003, Springer-Verlag, Heidelberg.

[30] V. Vovk, "Aggregating strategies," *Proc. 3'rd Annual Workshop on Computational Learning Theory*, pp. 371-383, Morgan Kaufman, San Mateo, 1990.

[31] V. Vovk, "Competitive On-line Statistics," *International Statistical Review*, 69: 213-248, 2001.

[32] M.H. Wegkamp, "Model selection in nonparametric regression," *Annals of Statistics*, 31: 252 – 273, 2003.

[33] K. Yamanishi, "Minimax relative loss analysis for sequential prediction algorithms using parametric hypotheses," Proceedings of *COLT 98*, ACM Press, pp.32-43, 1998.

[34] Y. Yang, "Adaptive Estimation in Pattern Recognition by Combining Different Procedures," *Statistica Sinica*, vol.10, pp.1069-1089, 2000.

[35] Y. Yang, "Adaptive regression by mixing," *JASA*, vol.96, pp.574-588, 2001.

[36] Y. Yang, "Aggregating regression procedures for a better performance, " *Bernoulli*, 10: 25 – 47, 2004.

[37] Y. Yang, "Combining forecasting procedures: some theoretical results," *Econometric Theory*, vol.20, pp.176-222, 2004.