Boosting for high-dimensional linear models

Peter Bühlmann
ETH Zürich, Switzerland

February 2, 2004

Abstract

We prove that boosting with the squared error loss, L_2Boosting, is consistent for very high-dimensional linear models, where the number of predictor variables is allowed to grow essentially as fast as $O(\exp(\text{sample size}))$, assuming that the true underlying regression function is sparse in terms of the ℓ_1-norm of the regression coefficients. In the language of signal processing, this means consistency for de-noising using a strongly overcomplete dictionary if the underlying signal is sparse in terms of the ℓ_1-norm.

L_2Boosting is computationally attractive. We propose an AIC-based estimate for tuning, namely choosing the number of boosting iterations. This makes L_2Boosting computationally even more attractive since boosting is not required to be run multiple times for cross-validation as commonly used in practice.

We demonstrate L_2Boosting for simulated data, where the predictor dimension is large in comparison to sample size, and for a difficult tumor-classification problem with gene expression microarray data.

Heading: Boosting for high-dimensional regression

1 Introduction

Freund and Schapire's (1996) AdaBoost algorithm for classification has attracted much attention in the machine learning community (see also Schapire (2002) and the references therein) as well as in related areas in statistics (Breiman, 1998; Friedman et al., 2000), mainly because of its good empirical performance in a variety of data sets. Boosting methods have been originally introduced as multiple prediction schemes, averaging estimated predictions from re-weighted data. Later, Breiman (1998, 1999) noted that the AdaBoost algorithm can be viewed as a gradient descent optimization technique in function space. This important insight opened a new perspective, namely to use boosting methods in other contexts than classification. For example, Friedman (2001) developed boosting methods for regression where boosting is implemented as an optimization with the squared error loss function: this is what we call L_2Boosting. It is essentially the same as Mallat and Zhang's (1993) matching pursuit algorithm in signal processing.

Recently, Efron et al. (2004) made for linear models a connection between L_2Boosting and the ℓ_1-penalized Lasso (Tibshirani, 1996) or basis pursuit (Chen et al., 1999) method. Roughly speaking, L_2Boosting approximately yields the set of all Lasso solutions (when varying over the penalty parameter). This intriguing insight may be useful to get a rough picture about L_2Boosting: it does variable selection and shrinkage, similar to Lasso. However, it should be stated clearly that the methods are not the same.

As the main result, we prove here that L_2Boosting for linear models yields consistent estimates in the very high-dimensional context, where the number of predictor variables is allowed to grow essentially as fast as $O(\exp(\text{sample size}))$, assuming that the true underlying regression function is sparse in terms of the ℓ_1-norm of the regression coefficients. This result is, to our knowledge, the first about boosting in the presence of (fast) growing dimension of the predictor. Some consistency results for boosting with fixed predictor dimension include Mannor et al. (2002), Jiang (2004), Lugosi and Vayatis (2004) as well as Zhang and Yu (2003). We believe that it is exactly for the case of high-dimensional predictors where boosting, among other methods, has a substantial advantage over more classical approaches, as demonstrated with some empirical examples in Bühlmann and Yu (2003); notably, many real data-sets nowadays are of such high-dimensional nature. Consistent estimation for very high-dimensional, but sparse functions may be achieved with other methods than boosting: for example, Bickel and Levina (2003) prove Bayes risk consistent classification of diagonal linear discriminant analysis for very high-dimensional predictors whose covariance matrix satisfies a regularity constraint in terms of its condition number. We think that besides the well-documented good empirical performance of boosting, it is important to identify it as a method which can consistently recover very high-dimensional, sparse functions.

Of course, we can also think of our result as a consistency property for de-noising using L_2Boosting with a strongly overcomplete dictionary. In contrast to a complete dictionary, e.g. Fourier- or wavelet-basis, the strongly overcomplete noisy case is not well understood. Our result yields at least the basic property of consistency.

L_2Boosting has an important computational advantage over Lasso (although the Lars algorithm from Efron et al. (2004) is even faster). Instead of having to search for a best penalty (tuning) parameter in Lasso (over a grid of candidate values), which requires solving many linear programming problems (up to numerical convergence), we can do one

sweep of boosting, which involves only fitting simple least squares regression many times. Moreover, for the tuning parameter in boosting, which is the number of boosting iterations, we develop some easily computable definition of degrees of freedom for L_2Boosting, and we then propose its use in the (corrected) AIC criterion. Unlike cross-validation, our AIC-type tuning estimator does not require boosting to be run multiple times. This makes the fully data-driven boosting computationally attractive and much faster than cross-validating Lasso (and sometimes faster than cross-validating the Lars algorithm from Efron et al. (2004)).

We demonstrate on some simulated examples how our L_2Boosting performs for (high-dimensional) linear models, in comparison to ordinary least squares, forward variable selection and a method which has been designed for high-dimensional regression (Goldenshluger and Tsybakov, 2001). We also consider a difficult tumor-classification problem with gene expression microarray data: the predictive accuracy of L_2Boosting is compared with four other, commonly used classifiers for microarray data, and we briefly indicate the interpretation of the L_2Boosting-fit along the lines of a linear model fit.

2 L_2Boosting with componentwise linear least squares

To explain boosting for linear models, consider a regression model

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i, \ i = 1, \ldots, n,$$

with p predictor variables (the jth component of a p-dimensional vector x is denoted by $x^{(j)}$) and a random, mean zero error term ε. More precise assumptions for the model are given in section 3.

We first specify a base procedure: given some input data $\{(X_i, U_i);\ i = 1, \ldots, n\}$, where $U_1, \ldots, U_n$ denote some (pseudo-)response variables which are not necessarily the original $Y_1, \ldots, Y_n$, the base procedure yields an estimated function

$$\hat{g}(\cdot) = \hat{g}_{(\mathbf{X},\mathbf{U})}(\cdot),$$

based on $\mathbf{X} = (X_1^T, \ldots, X_n^T)^T$, $\mathbf{U} = (U_1, \ldots, U_n)^T$. Here, we will exclusively consider the componentwise linear least squares base procedure:

$$\hat{g}_{(\mathbf{X},\mathbf{U})}(x) = \hat{\beta}_{\hat{\mathcal{S}}} x^{(\hat{\mathcal{S}})}, \ \hat{\beta}_j = \frac{\sum_{i=1}^n U_i X_i^{(j)}}{\sum_{i=1}^n (X_i^{(j)})^2} \ (j = 1, \ldots, p),$$
$$\hat{\mathcal{S}} = \operatorname*{arg\,min}_{1 \le j \le p} \sum_{i=1}^{n} (U_i - \hat{\beta}_j X_i^{(j)})^2. \tag{2.1}$$

Thus, the componentwise linear least squares base procedure performs a linear least squares regression against the one selected predictor variable which reduces residual sum of squares most.

Boosting using the squared error loss, L_2Boosting, has a simple structure. Boosting algorithms using other loss functions are described in Friedman (2001).

L_2Boost algorithm

Step 1 (initialization). Given data $\{(X_i, Y_i); i = 1, \ldots, n\}$, apply the base procedure yielding the function estimate

$$\hat{F}^{(1)}(\cdot) = \hat{g}(\cdot),$$

where $\hat{g} = \hat{g}_{(\mathbf{X},\mathbf{Y})}$ is estimated from the original data. Set $m = 1$.

Step 2. Compute residuals $U_i = Y_i - \hat{F}^{(m)}(X_i)$ $(i = 1, \ldots, n)$ and fit the real-valued base procedure to the current residuals. The fit is denoted by $\hat{g}_{m+1}(\cdot) = \hat{g}_{(X,U)}(\cdot)$ which is an estimate based on the original predictor variables and the current residuals.
Update

$$\hat{F}^{(m+1)}(\cdot) = \hat{F}^{(m)}(\cdot) + \hat{g}_{m+1}(\cdot).$$

Step 3 (iteration). Increase the iteration index m by one and repeat Step 2 until a stopping iteration M is achieved.

The estimate $\hat{F}^{(M)}(\cdot)$ is an estimator of the regression function $\mathbb{E}[Y|X = \cdot]$. L_2Boosting is nothing else than repeated least squares fitting of residuals (cf. Friedman (2001), Bühlmann and Yu (2003)). With $m = 2$ (one boosting step), it has already been proposed by Tukey (1977) under the name "twicing". In the non-stochastic context, the L_2Boosting algorithm is known as "Matching Pursuit" (Mallat and Zhang, 1993) which is popular in signal processing for fitting overcomplete dictionaries.

It is often better to use small step sizes: we advocate here to use the step-size ν in the update of $\hat{F}^{(m+1)}$ in step 2 which then becomes

$$\hat{F}^{(m+1)}(\cdot) = \hat{F}^{(m)}(\cdot) + \nu \hat{g}_{m+1}(\cdot), \ 0 < \nu \leq 1, \tag{2.2}$$

where ν is constant during boosting iterations and small, e.g. $\nu = 0.1$. The parameter ν can be seen as a shrinkage parameter or alternatively, describing the step-size when up-dating $\hat{F}^{(m+1)}(\cdot)$ along the function $\hat{g}_{m+1}(\cdot)$. Small step-sizes (or shrinkage) make the boosting algorithm slower and require a larger number M of iterations. However, the computational slow-down often turns out to be advantageous for better out-of-sample empirical prediction performance, cf. Friedman (2001), Bühlmann and Yu (2003). There are also some theoretical reasons to use boosting with ν (infinitesimally) small (Efron et al. 2004).

2.1 Stopping the boosting iterations

Boosting needs to be stopped at a suitable number of iterations, to avoid overfitting. The computationally efficient AIC_c criterion in (2.3) below can be used in our context where the base procedure has linear components.

Our goal here is to assign degrees of freedom for boosting. Denote by

$$\mathcal{H}^{(j)} = \mathbf{X}^{(j)}(\mathbf{X}^{(j)})^T / \|\mathbf{X}^{(j)}\|^2, \ j = 1, \ldots, p,$$

the $n \times n$ hat-matrix for the linear least squares fitting operator using the jth predictor variable $\mathbf{X}^{(j)} = (X_1^{(j)}, \ldots, X_n^{(j)})^T$ only; $\|x\|^2 = x^T x$ denotes the Euclidean norm for a vector $x \in \mathbb{R}^p$. It is then straightforward to show (cf. Bühlmann and Yu, 2003) that the L_2Boosting hat-matrix, when using the step size $0 < \nu \leq 1$, equals,

$$\mathcal{B}_m = I - (I - \nu \mathcal{H}^{(\hat{\mathcal{S}}_1)})(I - \nu \mathcal{H}^{(\hat{\mathcal{S}}_2)}) \ldots (I - \nu \mathcal{H}^{(\hat{\mathcal{S}}_m)}),$$

where $\hat{\mathcal{S}}_i \in \{1, \ldots, d\}$ denotes the component which is selected in the componentwise least squares base procedure in the ith boosting iteration.

We can now use a corrected version of AIC (cf. Hurvich et al. (1998)) to define a stopping rule of boosting:

$$\begin{aligned} AIC_c(m) &= \log(\hat{\sigma}^2) + \frac{1 + \mathrm{trace}(\mathcal{B}_m)/n}{1 - (\mathrm{trace}(\mathcal{B}_m) + 2)/n}, \\ \hat{\sigma}^2 &= n^{-1} \sum_{i=1}^{n} (Y_i - (\mathcal{B}_m \mathbf{Y})_i)^2, \ \mathbf{Y} = (Y_1, \ldots, Y_n)^T. \end{aligned} \tag{2.3}$$

An estimate for the number of boosting iterations is then

$$\hat{M} = \operatorname*{arg\,min}_{1 \leq m \leq m_{upp}} AIC_c(m),$$

where m_{upp} is a large, upper bound for the candidate number of boosting iterations.

3 Consistency of L_2Boosting for high-dimensional linear model

We present here a consistency result for L_2Boosting in linear models where the number of predictors is allowed to grow very fast as the sample size n increases. Consider the model

$$\begin{aligned} Y_i &= f_n(X_i) + \varepsilon_i, \ i = 1, \ldots, n, \\ f_n(x) &= \sum_{j=1}^{p_n} \beta_{j,n} x^{(j)}, \ x \in \mathbb{R}^{p_n}, \end{aligned} \tag{3.1}$$

where $X_1, \ldots, X_n$ are i.i.d. with $\mathbb{E}|X^{(j)}|^2 \equiv 1$ for all $j = 1, \ldots, p_n$ and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d., independent from $\{X_s; 1 \leq s \leq n\}$, with $\mathbb{E}[\varepsilon] = 0$. The number of predictors p_n is allowed to grow with sample size n. Therefore, also the predictor $X_i = X_{i,n}$ and the response $Y_i = Y_{i,n}$ depend on n, but we usually ignore this in the notation. The scaling of the predictor variables $\mathbb{E}|X^{(j)}|^2 = 1$ is not necessary for running L_2Boosting, but it allows to identify the magnitude of the coefficients $\beta_{j,n}$ (see also assumption (A1) below).

We make the following assumptions.

(A1) The dimension of the predictor in model (3.1) satisfies $p_n = O(\exp(Cn^{1-\xi}))$ $(n \to \infty)$, for some $0 < \xi < 1$, $0 < C < \infty$.

(A2) $\sup_{n \in \mathbb{N}} \sum_{j=1}^{p_n} |\beta_{j,n}| < \infty$.

(A3) $\sup_{1 \leq j \leq p_n} \|X^{(j)}\|_\infty < \infty$, where $\|X\|_\infty = \sup_{\omega \in \Omega} |X(\omega)|$ (Ω denotes the underlying probability space).

(A4) $\mathbb{E}|\varepsilon|^s < \infty$ for some $s > 2/\xi$ with ξ from (A1).

Assumption (A1) allows for a very large predictor dimension relative to the sample size n. Assumption (A2) is a ℓ_1-norm sparseness condition (it could be generalized to $\sum_{j=1}^{p_n} |\beta_{j,n}| \to \infty$ slow enough as $n \to \infty$, at the expense of other restrictions on p_n and m_n in Theorem 1 below). Even if p_n grows, all predictors may be relevant but most of them contribute only with small magnitudes (small $|\beta_{j,n}|$). Assumption (A2) holds for regressions where the number of effective predictors is finite and fixed: that is, the number of $\beta_{j,n} \neq 0$ is independent from n and finite. Assumption (A3) about the boundedness of the predictor variables can be relaxed at the price of a more restrictive growth of $p = p_n$, see Remark 2 below.

Theorem 1 *Consider the model (3.1) satisfying (A1)-(A4). Then, the boosting estimate $\hat{F}^{(m)}(\cdot) = \hat{F}_n^{(m)}(\cdot)$ with the componentwise linear learner from (2.1) satisfies: for some sequence $(m_n)_{n\in\mathbb{N}}$, which is allowed to be random and depending on the realizations of the data, with $m_n = o_P(n^{\xi/4})$ $(n \to \infty)$,*

$$\mathbb{E}_X|\hat{F}_n^{(m_n)}(X) - f_n(X)|^2 = o_P(1) \ (n \to \infty),$$

where X denotes a new predictor variable, independent of and with the same distribution as the data.

A proof is given in section 6. Theorem 1 says that L_2Boosting recovers the true sparse regression function even if the number of predictor variables is essentially exponentially increasing with sample size n. Notably, no assumptions are needed on the correlation structure of the predictor variables.

Remark 1. The restriction about the increase of the boosting iteration m_n as $n \to \infty$ is probably far from the fastest possible whose exploration is beyond our scope. However, the qualitative behavior of having m_n to grow slower with a fast growing p_n ($\xi > 0$ small) seems correct since a large p_n would more easily lead to overfitting (assuming a fixed "signal intensity" $\sup_{n\in\mathbb{N}} \sum_{j=1}^{p_n} |\beta_{j,n}|$). Moreover, the fact that $m = m_n$ is allowed to be random yields a theoretical framework which is closer to practice where m is chosen via the AIC_c statistic in (2.3) and hence depending on the realizations of the data.

Remark 2. Assumption (A2) requires boundedness of the predictor variables. Theorem 1 also holds under the assumption

$$\sup_{1\le j\le p_n} \mathbb{E}|X^{(j)}|^s < \infty \text{ for some } s \ge 4$$

(and using another restriction for $m = m_n$) if the growth of dimension is restricted to $p_n = O(n^\alpha)$ where $\alpha = \alpha(s) > 0$ is a number, depending on the number of existing moments s, which converges monotonically to ∞ as s increases, i.e. any polynomial growth of p_n is allowed if the number of moments s is sufficiently large.

4 Numerical results

4.1 Low-dimensional regression surface within high-dimensional predictor space

We consider the model

$$X \sim \mathcal{N}_{10}(0, V),\ Y = f(X) + \varepsilon,$$
$$f(X) = a(V)(1 + 5X_1 + 2X_2 + X_3),\ a(V) \text{ a constant},\ \varepsilon \sim \mathcal{N}(0, 2^2). \quad (4.1)$$

The covariance matrix for the predictor variable X and the constant $a(V)$ are chosen as:

$$V = I_{10},\ a(V) \equiv 1 \quad (4.2)$$

for uncorrelated predictors; or for block-correlated predictors,

$$V = \begin{pmatrix} 1 & b & c & 0 & \dots & \dots & \dots & 0 \\ b & 1 & b & c & 0 & \dots & \dots & 0 \\ c & b & 1 & b & c & 0 & \dots & 0 \\ 0 & c & b & 1 & b & c & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & c & b & 1 & b & c \\ 0 & \dots & \dots & 0 & c & b & 1 & b \\ 0 & \dots & \dots & \dots & 0 & c & b & 1 \end{pmatrix},$$
$$b = 0.677, c = 0.323,\ a(V) = 0.779. \quad (4.3)$$

The constant $a(V)$ is such that the signal to noise ratio $\mathrm{Var}(f(X))/\sigma_\varepsilon^2$ is the same for both model specifications. The model (4.1) with either specification (4.2) or (4.3) has only 3 effective predictors, all of them contributing to the regression function with different magnitudes (different coefficients). We choose sample size $n = 20$, i.e. we generate 20 i.i.d. realizations $(X_i, Y_i),\ i = 1, \dots 20$ from the model. Relative to the number of predictor variables $p = 10$, the problem is high-dimensional with a low-dimensional (effective $p = 3$) true underlying structure.

We use L_2Boosting, using shrinkage factor $\nu = 0.1$ (see (2.2)) and the corrected AIC criterion for stopping the boosting iterations (see (2.3)). We compare it with forward variable selection for optimizing the classical AIC criterion and with ordinary least squares (OLS) without variable selection. Table 4.1 and Figure 4.1 report the mean squared error $\mathrm{MSE} = \mathbb{E}[(\hat{f}(X) - f(X))^2]$ where X is a new test observation, independent from,

predictor	L_2Boost	forward var. sel.	OLS
uncorrelated (4.2)	2.318 (0.238)	3.648 (0.421)	5.674 (0.556)
correlated (4.3)	1.649 (0.181)	2.893 (0.373)	5.674 (0.556)

Table 4.1: Mean squared error $\mathbb{E}[(\hat{f}(X) - f(X))^2]$ for L_2Boosting, forward variable selection and ordinary least squares in model (4.1) with specifications (4.2) and (4.3). Estimated standard errors from independent model simulations are given in parentheses.

but with the same distribution as the training data. All results are based on 50 model simulations. Figure 4.1 displays the resistance of boosting against overfitting and also the good performance of the corrected AIC criterion in (2.3) for stopping the boosting iterations.

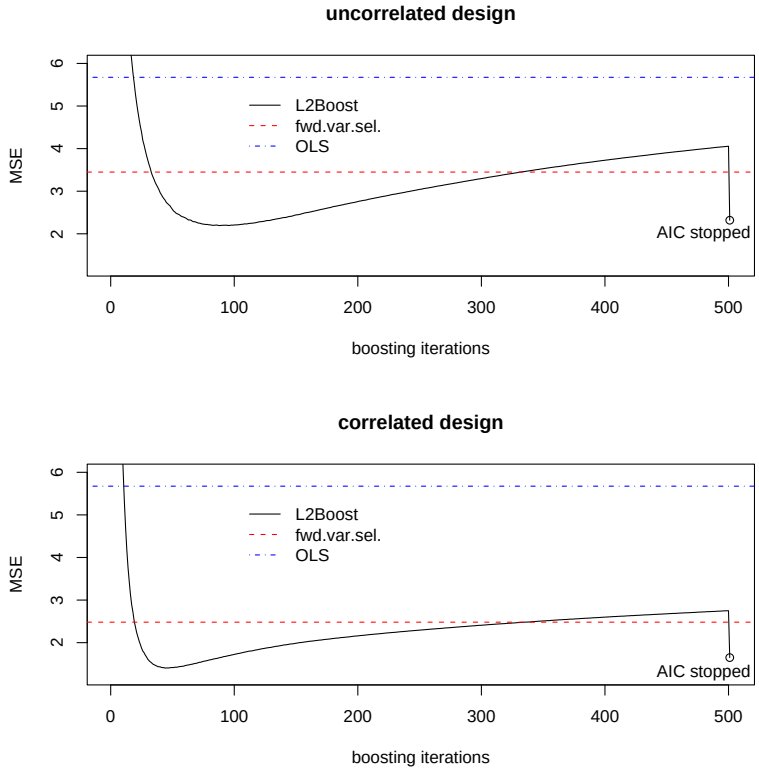

Figure 4.1: Mean squared error $\mathbb{E}[(\hat{f}(X) - f(X))^2]$ for L_2Boosting as a function of boosting iterations (solid line), forward variable selection (dashed line) and ordinary least squares (dashed-dotted line) in model (4.1) with specifications (4.2) (top panel) and (4.3) (bottom panel). The performance when estimating the number of boosting iterations with the corrected AIC criterion is indicated by the circle "AIC stopped".

4.2 High-dimensional regression surface with ℓ_1 coefficients

We consider here a regression model which fits into the theory of an adaptive estimation procedure for high-dimensional linear regression, presented by Goldenshluger and Tsybakov (2001).

The model is

$$X \sim \mathcal{N}_p(0, I), \;\; Y = \sum_{j=1}^{p} \beta_j X^{(j)} + \varepsilon, \;\; \varepsilon \sim \mathcal{N}(0,1),$$
$$\beta_1, \ldots, \beta_p \text{ independent}, \;\; \beta_j \sim \mathcal{N}(0, \sigma_j^2), \tag{4.4}$$

where ε, X and $\beta_1, \ldots, \beta_p$ are independent of each other. The values σ_j^2 are decreasing as j increases. Thus, absolute values of the regression coefficients $|\beta_j|$ have a tendency to become small for large j. A precise description of the model is given in Appendix A. To summarize, the model is such that $p = p_n$ and $\beta_j = \beta_{j,n}$ $(j = 1, \ldots, p_n)$ depend on n, satisfying

$$\sup_{n \in \mathbb{N}} \sum_{j=1}^{p_n} |\beta_{j,n}| < \infty,$$

which is our assumption (A1) from section 3. Sample size is chosen as $n = 100$ and the resulting dimension of the predictor then equals $p = 23$.

We use L_2Boosting, using shrinkage $\nu = 0.1$ (see (2.2)) and the estimated number of boosting iterations with the corrected AIC criterion as in (2.3), and we compare it with forward variable selection for optimizing the classical AIC criterion, with ordinary least squares (OLS) without variable selection and with the procedure from Goldenshluger and Tsybakov (2001). Table 4.2 and Figure 4.2 display the results. All the results are based

	L_2Boost	G&T method	forward var. sel.	OLS
MSE	0.132 (0.006)	0.195 (0.047)	0.279 (0.019)	0.313 (0.017)

Table 4.2: Mean squared error $\mathbb{E}[(\hat{f}(X) - f(X))^2]$ for L_2Boosting, the method from Goldenshluger and Tsybakov (G&T), forward variable selection and ordinary least squares in model (4.4). Estimated standard errors from independent model simulations are given in parentheses.

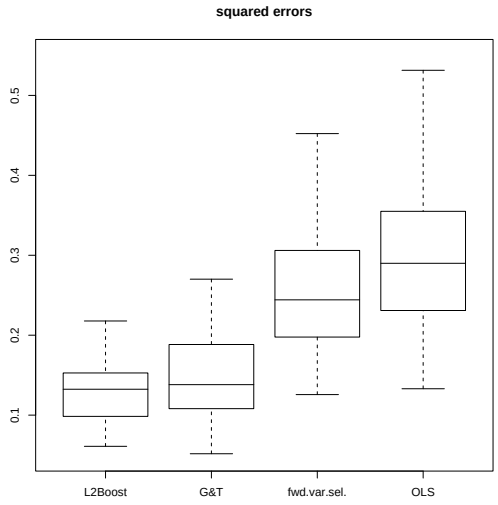

Figure 4.2: Boxplots (without outliers) of squared errors $(\hat{f}(X) - f(X))^2$ for L_2Boosting, the method from Goldenshluger and Tsybakov (G&T), forward variable selection and ordinary least squares in model (4.4).

on 50 independent model simulations. The method from Goldenshluger and Tsybakov (2001) produced one outlier with very large squared error (not shown in the boxplot), but Figure 4.2 still shows the substantial advantage of boosting.

Moreover, the method from Goldenshluger and Tsybakov (2001) depends on the indexing of the predictor variables and is tailored for regression problems where the coefficients β_j have a tendency to decay as j increases (e.g. in time series where j indicates the jth lagged variable). All other methods are not depending on indexing the predictor variables. We also ran the method from Goldenshluger and Tsybakov (2001) on the same model but with index-reversed regression coefficients

$$\beta_1, \dots \beta_{23} = \tilde{\beta}_{23}, \dots, \tilde{\beta}_1, \ \tilde{\beta}_j \text{ as in (4.4)}. \tag{4.5}$$

The mean squared error was then

$$\text{MSE for G\&T method with (4.5):} \quad 0.224 \ (0.025)$$

which shows very clearly the sensitivity of indexing the variables.

4.3 Gene expression microarray data

We consider a dataset which monitors $p = 7129$ gene expressions in 49 breast tumor samples using the Affymetrix technology, see West et al. (2001). After thresholding to a floor of 100 and a ceiling of 16,000 expression units, we applied a base 10 log-transformation and standardized each experiment to zero mean and unit variance. For each sample, a binary response variable is available, describing the status of lymph node involvement in breast cancer. The data are available at `http://mgm.duke.edu/genome/dna_micro/work/`.

We use L_2Boosting although the data has the structure of a binary classification problem; Bühlmann and Yu (2003) argue why L_2Boosting is also a reasonable procedure for binary classification (but we have not given a proof for Theorem 1 in case of heteroscedastic errors which would be needed for the classification case). The only modification is the AIC stopping criterion: instead of (2.3), we use

$$AIC(m) = -2 \cdot \text{log-likelihood} + 2 \cdot \text{trace}(\mathcal{B}_m),$$

with the Bernoulli log-likelihood. Instead of L_2Boost, we could also use the LogitBoost algorithm (Friedman et al., 2000): for stopping, the penalty-term in the AIC criterion above might then need some modification since LogitBoost involves another operator than $\mathcal{B}_m$.

We estimate the classification performance by a cross-validation scheme where we randomly divide the 49 samples into balanced training- and test-data of sizes $2n/3$ and $n/3$, respectively, and we repeat this 50 times. We compare L_2Boosting with AIC-stopping (as described above) with four other classification methods: 1-nearest neighbors, diagonal linear discriminant analysis, support vector machine (from the R-package `e1071`) with radial basis kernel, and a forward selection penalized logistic regression model. For 1-nearest neighbors, diagonal linear discriminant analysis and support vector machines, we use the 200 genes which have the best Wilcoxon score in a two-sample problem (estimated from the training dataset only), which is recommended to improve the classification performance,

	L_2Boost	FPLR	1-NN	DLDA	SVM
misclassifications	30.50%	35.25%	43.25%	36.12%	36.88%

Table 4.3: Cross-validated misclassification rates for lymph node breast cancer data. L_2Boosting is with linear least squares and AIC-stopping (L_2Boost), forward variable selection penalized logistic regression (FPLR), 1-nearest-neighbor rule (1-NN), diagonal linear discriminant analysis (DLDA) and a support vector machine (SVM); the latter three are based on 200 best genes (on each training dataset) according to a Wilcoxon score.

see Dudoit et al. (2002). Our L_2Boosting and the forward variable selection penalized regression are run without pre-selection of genes.

For this difficult classification problem, our L_2Boosting with componentwise linear least squares performs well. It is also interesting to note that the minimal cross-validated misclassification rate as a function of boosting iterations is 29.25%. It shows that the AIC-stopping rule is very accurate for this example. The only method which we know to be better is the recently proposed Pelora algorithm (Dettling and Bühlmann, 2003) which does supervised gene grouping: its misclassification rate is 27.88%.

We also show in Figure 4.3 the estimated regression coefficients for the 42 genes which have been selected during the boosting iterations until AIC-stopping; the AIC-curve is also shown in Figure 4.3. For comparing the influence of different genes, we show scaled coefficients $\hat{\beta}\sqrt{\mathrm{Var}\, X^{(j)}}$ which are the coefficients when standardizing the genes to unit variance. There is one gene whose positive expression strongly points towards the class with $Y = 0$ (having negative estimated regression coefficient) and there are 5 genes whose

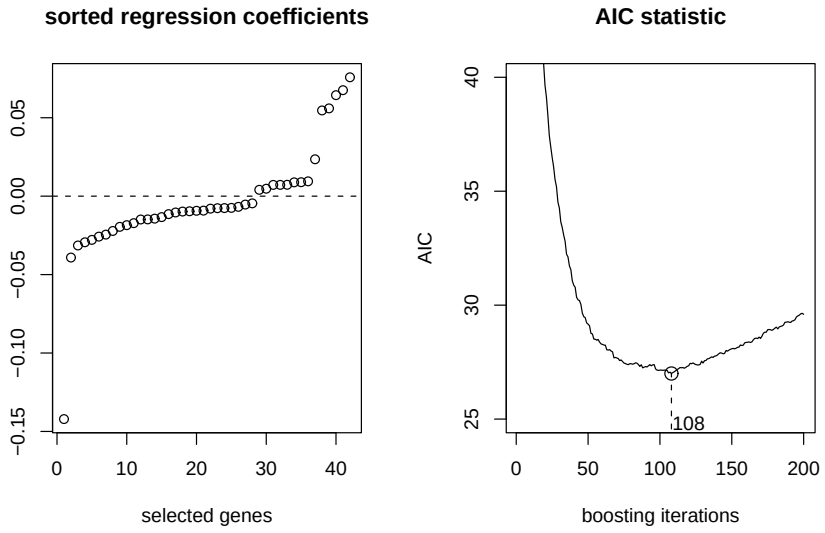

Figure 4.3: Lymph node breast cancer data. Left: scaled regression coefficients $\hat{\beta}_j\sqrt{\mathrm{Var}\, X^{(j)}}$ (plotted in increasing order) from L_2Boosting for the selected 42 genes. Right: AIC-statistic as a function of L_2Boosting iterations with minimum at 108.

positive expressions (individually) point towards the class with $Y = 1$. The smallest standardized regression coefficient corresponds to a gene which appears as the second best when ranking all the genes with the score of a two-sample Wilcoxon test; the five largest standardized coefficients correspond to the Wilcoxon-based ranks 7, 6, 1, 121, 3 among all the genes. But it should be emphasized that, as usual, our estimated regression model goes well beyond describing the effects of single genes.

5 Conclusions

We consider L_2Boosting for fitting linear models. The method does variable selection and shrinkage, a property which is very useful in practical applications. This indicates that L_2Boosting is related to the ℓ_1-penalized Lasso, but the methods are not the same.

As a useful device, we propose a simple estimate for the number of boosting iterations, which is the tuning parameter of the method, by using a corrected AIC_c criterion. This makes the computationally efficient boosting even more attractive, since we do not have to run boosting multiple times in a cross-validation set-up.

We then present some theory for very high-dimensional regression (or for de-noising with strongly overcomplete dictionaries), saying that if the underlying true regression function is sparse in terms of the ℓ_1-norm of the regression coefficients, the L_2Boosting method consistently estimates the true regression function, even when the number of predictor variables grows like $p = O(\exp(n^{1-\xi})$ for some (small) $\xi > 0$. Notably, no assumptions are made one the correlation structure of the predictors. Thus, we identify L_2Boosting as a method which is able, under mild assumptions, to consistently recover very high-dimensional, sparse functions.

6 Proofs

6.1 A population version

The L_2Boosting algorithm has a population version which is known as "matching pursuit" (Mallat and Zhang, 1993) or "weak greedy algorithm" (cf. Temlyakov (2000)).

Consider the Hilbert space $L_2(P) = \{f; \|f\|_2^2 = \int f(x)^2 dP(x) < \infty\}$ with inner product $\langle f, g\rangle = \int f(x)g(x)dP(x)$. Here, the probability measure P is generating the predictor X in model (3.1). To be precise, the probability measure $P = P_n$ depends on n since the dimensionality of X is growing with n: we are actually looking at a sequence of Hilbert spaces $L_2(P_n)$ but we often ignore this notationally (a uniform bound in (6.5) will be a key result to deal with such sequences of Hilbert spaces).

Denote the components of X by

$$g_j(x) = x^{(j)},\ j = 1, \ldots, p_n.$$

Define the following sequence of remainder functions, called matching pursuit or weak greedy algorithm:

$$\begin{aligned} &R^0 f = f, \\ &R^m f = R^{m-1} f - \left\langle R^{m-1} f, g_{\mathcal{S}_m} \right\rangle g_{\mathcal{S}_m},\ m = 1, 2, \ldots \end{aligned} \tag{6.1}$$

where $\mathcal{S}_m$ would be ideally chosen as

$$\mathcal{S}_m = \text{argmax}_{1 \le j \le p_n} |\left\langle R^{m-1} f, g_j \right\rangle|.$$

The choice function $\mathcal{S}_m$ is often infeasible to realize in practice. A weaker criterion is: for every m (under consideration), choose any $\mathcal{S}_m$, which satisfies

$$|\left\langle R^{m-1} f, g_{\mathcal{S}_m} \right\rangle| \ge b \cdot \sup_{1 \le j \le p_n} |\left\langle R^{m-1} f, g_j \right\rangle| \text{ for some } 0 < b \le 1. \tag{6.2}$$

Of course, the sequence $R^m f = R^{m,\mathcal{S}} f$ depends on $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m$ how we actually make the choice in (6.2). Again, we will ignore this notationally.

It easily follows that

$$f = \sum_{j=0}^{m-1} \left\langle R^j f, g_{\mathcal{S}_{j+1}} \right\rangle g_{\mathcal{S}_{j+1}} + R^m f,$$

and

$$\|R^m f\|_2^2 = \|R^{m-1} f\|_2^2 - |\left\langle R^{m-1} f, g_{\mathcal{S}_m} \right\rangle|^2 \tag{6.3}$$

6.1.1 Temlyakov's result

Temlyakov (2000) gives a uniform bound for the algorithm in (6.1) with (6.2).

If the function f is representable as

$$f(x) = \sum_j \beta_j g_j(x), \ \sum_j |\beta_j| \le B < \infty, \tag{6.4}$$

which is true by our assumption (A1), then

$$\|R^m f\| \le B(1 + mb^2)^{-b/(2(2+b))}, \ 0 < b \le 1 \text{ as in (6.2)}. \tag{6.5}$$

To make the point clear, this bound holds also for sequences $R^m f = R^{m,\mathcal{S},n} f$ which depend on the choice function $\mathcal{S}$ in (6.2) and on the sample size n (since $X \sim P$ depends on n and also the function of interest $f = f_n$): all we have to assume is the condition (6.4).

6.2 Asymptotic analysis as sample size increases

The L_2Boosting algorithm can be represented analogously to (6.1). We introduce the following notation:

$$\langle f, g \rangle_{(n)} = n^{-1} \sum_{i=1}^n f(X_i) g(X_i), \text{ and } \|f\|_{(n)}^2 = n^{-1} \sum_{i=1}^n f(X_i)^2$$

for functions $f, g : \mathbb{R}^{p_n} \to \mathbb{R}$. As before, we denote by $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ the vector of response variables.

Define

$$\hat{R}_n^1 f = f - \left\langle \mathbf{Y}, g_{\hat{\mathcal{S}}_1} \right\rangle_{(n)} g_{\hat{\mathcal{S}}_1},$$
$$\hat{R}_n^m f = \hat{R}_n^{m-1} f - \left\langle \hat{R}_n^{m-1}, g_{\hat{\mathcal{S}}_m} \right\rangle_{(n)} g_{\hat{\mathcal{S}}_m},\ m = 2, 3, \ldots,$$

where

$$\hat{\mathcal{S}}_1 = \underset{1 \le j \le p_n}{\arg\max} |\langle \mathbf{Y}, g_j \rangle_{(n)}|,$$
$$\hat{\mathcal{S}}_m = \underset{1 \le j \le p_n}{\arg\max} \left| \left\langle \hat{R}_n^{m-1} f, g_j \right\rangle_{(n)} \right|,\ m = 2, 3, \ldots$$

Note that we emphasize here the dependence of $\hat{R}_n^m$ on n since finite-sample estimates $\left\langle \hat{R}_n^{m-1} f, g_j \right\rangle_{(n)}$ are involved.

The strategy is now to establish a finite-sample analogue of (6.2), and then invoke Temlyakov's (2000) result from (6.5).

6.2.1 Uniform laws of large numbers

Lemma 1 *Under the assumptions (A1)-(A4), with* $0 < \xi < 1$ *as in (A1),*

(i) $\sup_{1 \le j,k \le p_n, n \in \mathbb{N}} |n^{-1} \sum_{i=1}^n g_j(X_i) g_k(X_i) - \mathbb{E}[g_j(X) g_k(X)]| = \zeta_{n,1} = O_P(n^{-\xi/2})$,

(ii) $\sup_{1 \le j \le p_n, n \in \mathbb{N}} |n^{-1} \sum_{i=1}^n g_j(X_i) \varepsilon_i| = \zeta_{n,2} = O_P(n^{-\xi/2})$,

(iii) $\sup_{1 \le j \le p_n, n \in \mathbb{N}} |n^{-1} \sum_{i=1}^n f(X_i) g_j(X_i) - \mathbb{E}[f(X_i) g_j(X)]| = \zeta_{n,3} = O_P(n^{-\xi/2})$,

(iv) $\sup_{1 \le j \le p_n, n \in \mathbb{N}} |n^{-1} \sum_{i=1}^n g_j(X_i) Y_i - \mathbb{E}[g_j(X) Y]| = \zeta_{n,4} = O_P(n^{-\xi/2})$,

(v) $|n^{-1} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)^2]| = \zeta_{n,5} = O_P(n^{-\xi/2})$,

(vi) $|n^{-1} \sum_{i=1}^n f(X_i) \varepsilon_i - \mathbb{E}[f(X) \varepsilon]| = \zeta_{n,6} = O_P(n^{-\xi/2})$

(vii) $|n^{-1} \sum_{i=1}^n \varepsilon_i^2 - \mathbb{E}[\varepsilon^2]| = \zeta_{n,7} = O_P(n^{-1/2})$

Proof: For assertion (i), denote by $M = \sup_j \|g_j(X)\|_\infty$, see assumption (A3). Then, Bernstein's inequality yields for every $\gamma > 0$,

$$\begin{aligned} & \mathbb{P}[n^{\xi/2} \sup_{1 \le j,k \le p_n, n \in \mathbb{N}} |n^{-1} \sum_{i=1}^n g_j(X_i) g_k(X_i) - \mathbb{E}[g_j(X) g_k(X)]| > \gamma] \\ \le\ & p_n^2 2 \exp\left(- \frac{\gamma^2 n^{1-\xi}}{2(\sigma_g^2 + M^2 \gamma n^{-\xi/2})} \right), \end{aligned}$$

where σ_g^2 is an upper bound for $\mathrm{Var}(g_j(X) g_k(X))$ for all j, k (e.g. $\sigma_g^2 = M^4$). Since $p_n^2 = O(\exp(2C(n^{1-\xi})))$, the right-hand side of the inequality above becomes arbitrarily small for n sufficiently large and $\gamma > 0$ large.

For proving assertion (ii), we have to deal with the unboundedness of the ε_i's in order to apply Bernstein's inequality. Define the truncated variables

$$\varepsilon_i^{tr} = \begin{cases} \varepsilon_i, & \text{if } |\varepsilon_i| \le M_n \\ \text{sign}(\varepsilon_i) M_n, & \text{if } |\varepsilon_i| > M_n. \end{cases}$$

Then, for $\gamma > 0$,

$$\begin{aligned} & \mathbb{P}[n^{\xi/2} \sup_{1\le j\le p_n, n\in\mathbb{N}} |n^{-1}\sum_{i=1}^n g_j(X_i)\varepsilon_i| > \gamma] \\ \le\ & \mathbb{P}[n^{\xi/2} \sup_{1\le j\le p_n, n\in\mathbb{N}} |n^{-1}\sum_{i=1}^n g_j(X_i)\varepsilon_i^{tr} - \mathbb{E}[g_j(X)\varepsilon^{tr}]| > \gamma/3] \\ +\ & \mathbb{P}[n^{\xi/2} \sup_{1\le j\le p_n, n\in\mathbb{N}} |n^{-1}\sum_{i=1}^n g_j(X_i)(\varepsilon_i - \varepsilon_i^{tr})| > \gamma/3] \\ +\ & \mathbb{P}[n^{\xi/2} \sup_{1\le j\le p_n, n\in\mathbb{N}} |n^{-1}\sum_{i=1}^n \mathbb{E}[g_j(X_i)(\varepsilon_i - \varepsilon_i^{tr})]| > \gamma/3] \\ =\ & I + II + III, \end{aligned}$$

since $\mathbb{E}[g_j(X)\varepsilon] = \mathbb{E}[g_j(X)]\mathbb{E}[\varepsilon] = 0$. We can bound I again by using Bernstein's inequality:

$$I \le p_n 2 \exp\left(-\frac{\gamma^2/9n^{1-\xi}}{2(\sigma_g^2 + M_n\gamma/3n^{-\xi/2})} \right), \tag{6.6}$$

where σ_g^2 is an upper bound for $\text{Var}(g_j(X)\varepsilon^{tr})$ (e.g. $\sup_j \|g_j(X)\|_\infty^2 \mathbb{E}|\varepsilon|^2$). When using

$$M_n = n^{\xi/2},$$

we can make the right hand since in (6.6) arbitrarily small since $p_n = O(\exp(Cn^{1-\xi}))$: thus, for every $\delta > 0$,

$$I \le \delta \text{ for } n \text{ sufficiently large, } \gamma \text{ sufficiently large.} \tag{6.7}$$

A bound for II can be obtained as follows:

$$\begin{aligned} II &\le \mathbb{P}[\text{some } |\varepsilon_i| > M_n] \le n\mathbb{P}[|\varepsilon| > M_n] \le nM_n^{-s}\mathbb{E}|\varepsilon|^s \\ &= O(n^{1-s\xi/2}) = o(1)\ (n\to\infty) \end{aligned} \tag{6.8}$$

since $s > 2/\xi$ by assumption (A4).

For III we use the bound

$$III \le \mathbb{1}_{[n^{\xi/2}\sup_j |\mathbb{E}[g_j(X)(\varepsilon - \varepsilon^{tr})]| > \gamma/3]}. \tag{6.9}$$

Note that by the independence of ε (and ε^{tr}) from $g_j(X)$,

$$\mathbb{E}[g_j(X)(\varepsilon - \varepsilon^{tr})] = \mathbb{E}[g_j(X)]\mathbb{E}[\varepsilon - \varepsilon^{tr}].$$

Hence, an upper bound is

$$|\mathbb{E}[g_j(X)(\varepsilon - \varepsilon^{tr})]| \leq M|\mathbb{E}[\varepsilon - \varepsilon^{tr}]|.$$

The latter can be bounded as

$$\begin{aligned}
&|\mathbb{E}[\varepsilon - \varepsilon^{tr}]| \leq |\int_{|x|>M_n} (\text{sign}(x)M_n - x)dP_\varepsilon(x)| \leq \int \mathbb{I}_{[|x|>M_n]}(M_n + |x|)dP_\varepsilon(x) \\
= \;& M_n\mathbb{P}[|\varepsilon| > M_n] + \int |x|\mathbb{I}_{[|x|>M_n]}dP_\varepsilon(x) \\
\leq \;& M_n^{1-s}\mathbb{E}|\varepsilon|^s + (\mathbb{E}|\varepsilon|^2)^{1/2}(\mathbb{P}[|\varepsilon| > M_n])^{1/2} \\
= \;& O(M_n^{1-s}) + O(M_n^{-s/2}) = o(M_n^{-1}) = o(n^{-\xi/2})
\end{aligned}$$

since $s > 2/\xi > 2$ $(0 < \xi < 1)$. Hence, by using (6.9):

$$III = 0 \text{ for } n \text{ sufficiently large, } \gamma > 0 \text{ sufficiently large,}$$

and together with (6.7) and (6.8), this proves assertion (ii).

Assertion (iii) follows from (i):

$$\begin{aligned}
&\sup_{1\leq j\leq p_n, n\in\mathbb{N}} |n^{-1}\sum_{i=1}^n f(X_i)g_j(X_i) - \mathbb{E}[f(X)g_j(X)]| \\
\leq \;& \sum_{r=1}^{p_n} |\beta_{r,n}| \sup_{1\leq j,k\leq p_n, n\in\mathbb{N}} |n^{-1}\sum_{i=1}^n g_j(X_i)g_k(X_i) - \mathbb{E}[g_j(X)g_k(X)]| \\
\leq \;& \sum_{r=1}^{p_n} |\beta_{r,n}| \sup_{1\leq j,k\leq p_n, n\in\mathbb{N}} |n^{-1}\sum_{i=1}^n g_j(X_i)g_k(X_i) - \mathbb{E}[g_j(X)g_k(X)]| \\
\leq \;& \sum_{r=1}^{p_n} |\beta_{r,n}|\zeta_{n,1} = O_P(n^{-\xi/2}).
\end{aligned}$$

Assertion (iv) follows from (ii) and (iii):

$$\begin{aligned}
&\sup_{1\leq j\leq p_n, n\in\mathbb{N}} |n^{-1}\sum_{i=1}^n g_j(X_i)Y_i - \mathbb{E}[g_j(X)Y]| \\
\leq \;& \sup_{1\leq j\leq p_n, n\in\mathbb{N}} |n^{-1}\sum_{i=1}^n f(X_i)g_j(X_i) - \mathbb{E}[f(X)g_j(X)]| \\
+ \;& \sup_{1\leq j\leq p_n, n\in\mathbb{N}} |n^{-1}\sum_{i=1}^n g_j(X_i)\varepsilon_i| \leq \zeta_{n,3} + \zeta_{n,2} = O_P(n^{-\xi/2}).
\end{aligned}$$

.

Assertion (v) follows from (i):

$$\begin{aligned}
&|n^{-1}\sum_{i=1}^n f(X_i)^2 - \mathbb{E}[f(X)^2]| \\
\leq \;& (\sum_{j=1}^{p_n} |\beta_{j,n}|)^2 \sup_{1\leq j,k\leq p_n, n\in\mathbb{N}} |n^{-1}\sum_{i=1}^n g_j(X_i)g_k(X_i) - \mathbb{E}[g_j(X)g_k(X)]| \\
\leq \;& (\sum_{j=1}^{p_n} |\beta_{j,n}|)^2\zeta_{n,1} = O_P(n^{-\xi/2}).
\end{aligned}$$

Assertion (vi) follows by

$$|n^{-1}\sum_{i=1}^{n} f(X_i)\varepsilon_i - \mathbb{E}[f(X)\varepsilon]| \leq \sum_{j=1}^{p_n} |\beta_{j,n}|\zeta_{n,2} = O_P(n^{-\xi/2}).$$

Finally, assertion (vii) is trivial. □

6.2.2 Recursive analysis of L_2Boosting

Denote by

$$\zeta_n = \max\{\zeta_{n,1}, \zeta_{n,2}, \zeta_{n,3}, \zeta_{n,4}, \zeta_{n,5}, \zeta_{n,6}, \zeta_{n,7}\} = O_P(n^{-\xi/2})$$

which is a bound for all assertions (i)-(vii) in Lemma 1. Also, we denote by ω a realization of all n data-points and we often abbreviate $m_n(\omega)$ by m.

Lemma 2 *Under the assumptions of Lemma 1, there exists a constant $0 < C_* < \infty$, independent from n and m, such that*

$$\sup_{1\leq j\leq p_n} |\left\langle \hat{R}_n^m f, g_j \right\rangle_{(n)} - \left\langle \hat{R}_n^m f, g_j \right\rangle| \leq m\zeta_n U_n,$$

where U_n is a random variable satisfying $0 \leq U_n \leq C_*$ on the set $A_n = \{\omega;\ |\zeta_n(\omega)| < 1\}$.

Note that Lemma 1 implies that $\mathbb{P}[A_n] \to 1$ $(n \to \infty)$. The constant C_* is depending on $\sup_{n\in\mathbb{N}} \sum_{j=1}^{p_n} |\beta_{j,n}|$ and $\mathrm{Var}(\varepsilon)$.

Proof: We proceed recursively. For $m = 1$, and using the definition of $\hat{R}_n^1 f$,

$$\begin{aligned} & \sup_{1\leq j\leq p_n} |\left\langle \hat{R}_n^1 f, g_j \right\rangle_{(n)} - \left\langle \hat{R}_n^1 f, g_j \right\rangle| \\ \leq & \sup_{1\leq j\leq p_n} |\langle f, g_j\rangle_{(n)} - \langle f, g_j\rangle| + |\left\langle \mathbf{Y}, g_{\hat{S}_1} \right\rangle_{(n)}| \sup_{1\leq j\leq p_n} |\left\langle g_{\hat{S}_1}, g_j \right\rangle_{(n)} - \left\langle g_{\hat{S}_1}, g_j \right\rangle| \\ = & I + II. \end{aligned}$$

From Lemma 1 (iii)

$$I \leq \zeta_n. \tag{6.10}$$

Regarding II, we can bound the first factor using the Cauchy-Schwarz inequality,

$$|\left\langle \mathbf{Y}, g_{\hat{S}_1} \right\rangle_{(n)}| \leq \|\mathbf{Y}\|_{(n)} \|g_{\hat{S}_1}\|_{(n)}. \tag{6.11}$$

Furthermore,

$$\|g_{\hat{S}_1}\|_{(n)}^2 = \|g_{\hat{S}_1}\|^2 + (\|g_{\hat{S}_1}\|_{(n)}^2 - \|g_{\hat{S}_1}\|^2) \leq \|g_{\hat{S}_1}\|^2 + \zeta_n = 1 + \zeta_n, \tag{6.12}$$

due to Lemma 1 (i) and the norming of the predictor $\|g_j\| = 1$ for all j. Similarly,

$$\begin{aligned} & \|\mathbf{Y}\|_{(n)}^2 = \|Y\|^2 + (\|f + \varepsilon\|_{(n)}^2 - \|f + \varepsilon\|^2) \\ \leq & \|Y\|^2 + |\|f\|_{(n)}^2 - \|f\|^2| + |\|\varepsilon\|_{(n)}^2 - \|\varepsilon\|^2| + 2|\langle f, \varepsilon\rangle_{(n)}| \\ \leq & \|Y\|^2 + 4\zeta_n. \end{aligned}$$

Thus, the bound in (6.11) becomes

$$|\left\langle \mathbf{Y}, g_{\hat{\mathcal{S}}_1} \right\rangle_{(n)}| \leq (\|Y\|^2 + 4\zeta_n)^{1/2}(1+\zeta_n)^{1/2}. \tag{6.13}$$

The second factor in II can be bounded by Lemma 1 (i):

$$\sup_{1\leq j\leq p_n} |\left\langle g_{\hat{\mathcal{S}}_1}, g_j \right\rangle_{(n)} - \left\langle g_{\hat{\mathcal{S}}_1}, g_j \right\rangle| \leq \sup_{1\leq j,k\leq p_n} |\langle g_j, g_k\rangle_{(n)} - \langle g_j, g_k\rangle| \leq \zeta_n. \tag{6.14}$$

Hence by (6.13),

$$II \leq \zeta_n(1+\zeta_n)^{1/2}(\|Y\|^2 + 4\zeta_n)^{1/2},$$

and therefore, by (6.10),

$$\begin{aligned} & \sup_{1\leq j\leq p_n} |\left\langle \hat{R}_n^1 f, g_j \right\rangle_{(n)} - \left\langle \hat{R}_n^1 f, g_j \right\rangle| \leq \zeta_n + \zeta_n(1+\zeta_n)^{1/2}(\|Y\|^2 + 4\zeta_n)^{1/2} \\ \leq\ & \zeta_n(1+C_{*1}) \text{ on the set } A_n. \end{aligned} \tag{6.15}$$

for some C_{*1} which depends on $\|Y\|^2 = \|f\|^2 + \mathrm{Var}(\varepsilon) \leq (\sum_{j=1}^{p_n} |\beta_{j,n}|)^2 + \mathrm{Var}(\varepsilon)$, e.g. we can choose it depending on $\sup_n \sum_{j=1}^{p_n} |\beta_{j,n}|$ and $\mathrm{Var}(\varepsilon)$.

For general m, by the definition of $\hat{R}_n^m f$,

$$\begin{aligned} & \sup_{1\leq j\leq p_n} |\left\langle \hat{R}_n^m f, g_j \right\rangle_{(n)} - \left\langle \hat{R}_n^m, g_j \right\rangle| \\ \leq\ & \sup_{1\leq j\leq p_n} |\left\langle \hat{R}_n^{m-1} f, g_j \right\rangle_{(n)} - \left\langle \hat{R}_n^{m-1} f, g_j \right\rangle| + \|\hat{R}_n^{m-1} f\|_{(n)} \|g_{\hat{\mathcal{S}}_m}\|_{(n)} \zeta_n \\ =\ & III + IV, \end{aligned} \tag{6.16}$$

using the analogous reasoning as in (6.11) with the Cauchy-Schwarz inequality and invoking the analogous bound as in (6.14)

$$\sup_{1\leq j\leq p_n} |\left\langle g_{\hat{\mathcal{S}}_m}, g_j \right\rangle_{(n)} - \left\langle g_{\hat{\mathcal{S}}_m}, g_j \right\rangle| \leq \zeta_n.$$

The first term III will be controlled from an induction with $m-1$ instead of m, as used in the last displayed formula of the proof.

For the second term IV, we develop a bound for $\|\hat{R}_n^m f\|_{(n)}$:

$$\begin{aligned} & \|\hat{R}_n^m f\|_{(n)}^2 = \|\hat{R}_n^{m-1} f\|_{(n)}^2 - 2|\left\langle \hat{R}_n^{m-1} f, g_{\hat{\mathcal{S}}_m} \right\rangle_{(n)}|^2 + |\left\langle \hat{R}_n^{m-1} f, g_{\hat{\mathcal{S}}_m} \right\rangle_{(n)}|^2 \|g_{\hat{\mathcal{S}}_m}\|_{(n)}^2 \\ =\ & \|\hat{R}_n^{m-1} f\|_{(n)}^2 - |\left\langle \hat{R}_n^{m-1} f, g_{\hat{\mathcal{S}}_m} \right\rangle_{(n)}|^2 (2 - \|g_{\hat{\mathcal{S}}_m}\|_{(n)}^2). \end{aligned} \tag{6.17}$$

Since

$$\|g_{\hat{\mathcal{S}}_m}\|_{(n)}^2 \leq 1 + \zeta_n$$

(see (6.12)), we obtain for (6.17):

$$\begin{aligned} & \|\hat{R}_n^m f\|_{(n)}^2 \leq \|\hat{R}_n^{m-1} f\|_{(n)}^2 - |\left\langle \hat{R}_n^{m-1} f, g_{\hat{\mathcal{S}}} \right\rangle_{(n)}|^2 (1-\zeta_n) \\ \leq\ & \|\hat{R}_n^{m-1} f\|_{(n)}^2 \text{ on the set } A_n = \{\omega;\ |\zeta_n(\omega)| < 1\}. \end{aligned}$$

Proceeding recursively,

$$\begin{aligned} & \|\hat{R}_n^m f\|_{(n)}^2 \leq \|\hat{R}_n^1 f\|_{(n)}^2 \text{ on the set } A_n \\ \leq \; & (\|f\|_{(n)} + |\left\langle \mathbf{Y}, g_{\hat{\mathcal{S}}_1} \right\rangle| \|g_{\hat{\mathcal{S}}_1}\|_{(n)})^2 \leq \|f\|^2 + C_{*2} \text{ on the set } A_n. \end{aligned}$$

for some constant C_{*2}, independent from m, n and which we can choose depending on $\sup_n \sum_{j=1}^{p_n} |\beta_{j,n}|$ and $\mathrm{Var}(\varepsilon)$ (the last inequality follows from Lemma 1). Hence

$$\begin{aligned} & IV \leq \zeta_n (1+\zeta_n)^{1/2} (\|f\|^2 + C_{*2}) \text{ on the set } A_n, \\ \leq \; & \zeta_n C_{*3} \text{ on the set } A_n \end{aligned} \tag{6.18}$$

for some constant C_{*3} which we can choose depending on $\sup_n \sum_{j=1}^{p_n} |\beta_{j,n}|$ and $\mathrm{Var}(\varepsilon)$.

Now set $C_* = \max(1 + C_{*1}, C_{*3})$. We then get inductively, by (6.15) and repeatedly using (6.16) and (6.18),

$$\begin{aligned} & \sup_{1 \leq j \leq p_n} |\left\langle \hat{R}_n^m f, g_j \right\rangle_{(n)} - \left\langle \hat{R}_n^m, g_j \right\rangle| \leq (m-1)\zeta_n C_* + \zeta_n C_* \text{ on the set } A_n \\ = \; & m \zeta_n C_* \text{ on the set } A_n, \end{aligned}$$

which completes the proof of Lemma 2. □

We are now ready to establish a finite-sample analogue of (6.2). We have

$$\left\langle \hat{R}_n^m f, g_j \right\rangle_{(n)} = \left\langle \hat{R}_n^m f, g_j \right\rangle + (\left\langle \hat{R}_n^m f, g_j \right\rangle_{(n)} - \left\langle \hat{R}_n^m f, g_j \right\rangle).$$

Hence, by invoking Lemma 2 we get

$$\sup_{1 \leq j \leq p_n, n \in \mathbb{N}} |\left\langle \hat{R}_n^m f, g_j \right\rangle_{(n)}| \geq \sup_{1 \leq j \leq p_n, n \in \mathbb{N}} |\left\langle \hat{R}_n^m f, g_j \right\rangle| - m\zeta_n C_* \text{ on the set } A_n. \tag{6.19}$$

Consider the set $B_n = \{\omega;\ \sup_{1 \leq j \leq p_n, n \in \mathbb{N}} |\left\langle \hat{R}_n^m f, g_j \right\rangle| > 2m\zeta_n C_*\}$. Then, by (6.19),

$$\sup_{1 \leq j \leq p_n, n \in \mathbb{N}} |\left\langle \hat{R}_n^m f, g_j \right\rangle_{(n)}| \geq 0.5 \sup_{1 \leq j \leq p_n, n \in \mathbb{N}} |\left\langle \hat{R}_n^m f, g_j \right\rangle| \text{ on the set } A_n \cap B_n. \tag{6.20}$$

We can now invoke Temlyakov's result in (6.5), since the condition (6.2) holds on the set $A_n \cap B_n$ (as established in (6.20)),

$$\begin{aligned} & \|\hat{R}_n^m f\| = (\mathbb{E}_X[(\hat{F}_n^{(m)}(X) - f(X))^2])^{1/2} \leq B(1 + m/4)^{-1/10} \\ = \; & o(1) \text{ on the set } A_n \cap B_n \end{aligned} \tag{6.21}$$

by choosing $m = m_n(\omega) \to \infty$ $(n \to \infty)$ (slow enough) for $\omega \in A_n \cap B_n$.

For $\omega \in B_n^C = \{\omega;\ \sup_{1 \leq j \leq p_n, n \in \mathbb{N}} |\left\langle \hat{R}_n^m f, g_j \right\rangle| \leq 2m\zeta_n C_*\}$, by using formula (5.2) from Temlyakov (2000) with b_m as defined there,

$$\begin{aligned} & \|\hat{R}_n^m f\|^2 \leq \sup_{1 \leq j \leq p_n, n \in \mathbb{N}} |\left\langle \hat{R}_n^m f, g_j \right\rangle| b_m \leq \sup_{1 \leq j \leq p_n, n \in \mathbb{N}} |\left\langle \hat{R}_n^m f, g_j \right\rangle| (1 + m\|Y\|) \\ \leq \; & (m+1) \sup_{1 \leq j \leq p_n, n \in \mathbb{N}} |\left\langle \hat{R}_n^m f, g_j \right\rangle| C,\ C > 0 \text{ a constant} \\ \leq \; & 2m(m+1)\zeta_n C_* C \text{ on the set } B_n^C. \end{aligned} \tag{6.22}$$

For bounding the number b_m, we have used the norm reducing property in (6.3).

The proof of Theorem 1 is then complete by (6.21), by (6.22) together with the assumption $m = o_P(n^{\xi/4})$ and $\zeta_n = O_P(n^{-\xi/2})$ from Lemma 1, and by observing that $\mathbb{P}[(A_n \cap B_n) \cup B_n^C] \geq \mathbb{P}[A_n] \to 1$ $(n \to \infty)$ due to Lemma 1. □

References

[1] Bickel, P. and Levina, E. (2003). Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. Preprint. Department of Statistics, Univ. of Calif., Berkeley.

[2] Breiman, L. (1998). Arcing classifiers. Ann. Statist. **26**, 801–849 (with discussion).

[3] Breiman, L. (1999). Prediction games & arcing algorithms. Neural Computation **11**, 1493-1517.

[4] Bühlmann, P. and Yu, B. (2003). Boosting with the L_2loss: regression and classification. J. Amer. Statist. Assoc. **98**, 324–339.

[5] Chen, S.S., Donoho, D.L., Saunders, M.A. (1999). Atomic decomposition by basis pursuit. SIAM J. Scient. Comp. **20**(1), 33–61.

[6] Dettling, M. and Bühlmann, P. (2003). Finding predictive gene groups from microarray data. To appear in J. Multiv. Anal.

[7] Dudoit, S., Fridlyand, J. and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. J. Amer. Statist. Assoc. **97**, 77–87.

[8] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. To appear in Ann. Statist.

[9] Freund, Y. and Schapire, R.E. (1996). Experiments with a new boosting algorithm. In Machine Learning: Proc. Thirteenth International Conference, pp. 148–156. Morgan Kauffman, San Francisco.

[10] Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Ann. Statist. **29**, 1189–1232.

[11] Friedman, J.H., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. Ann. Statist. **28**, 337–407 (with discussion).

[12] Goldenshluger, A. and Tsybakov, A. (2001). Adaptive prediction and estimation in linear regression with infinitely many predictors. Ann. Statist. **29**, 1601–1619.

[13] Hurvich, C.M., Simonoff, J.S. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J. Roy. Statist. Soc., Ser. B, **60**, 271–293.

[14] Jiang, W. (2004). Process consistency for AdaBoost. To appear in Ann. Statist.

[15] Lugosi, G. and Vayatis, N. (2004). On the Bayes-risk consistency of regularized boosting methods. To appear in Ann. Statist.

[16] Mallat, S and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Proc. **41**, 3397–3415.

[17] Mannor, S., Meir, R. and Zhang, T. (2002). The consistency of greedy algorithms for classification. To appear in COLT (fifteenth annual conference on computational learning theory).

[18] Schapire, R. E. (2002). The boosting approach to machine learning: an overview. In MSRI Workshop on Nonlinear Estimation and Classification (D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick and B. Yu, Eds.). Springer, New York. Press.

[19] Temlyakov, V.N. (2000). Weak greedy algorithms. Adv. Comp. Math. **12**, 213–227.

[20] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc., Ser. B, **58**, 267–288.

[21] Tukey, J.W. (1977). *Exploratory data analysis*. Addison-Wesley, Reading, MA.

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

[23] West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J., Marks, J., Nevins, J. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Nat. Acad. Sci. **98**, 11462–11467.

[24] Zhang, T. and Yu, B. (2003). Boosting with early stopping: convergence and consistency. Technical Report 635, Dept. of Statistics, Univ. of Calif., Berkeley. Available from http://www.stat.berkeley.edu/users/binyu/publications.html

Appendix A: The model (4.4)

The model (4.4) is as follows. Define $a_j = j^{0.51}$. Let the parameter κ be the solution of the equation $\sigma_\varepsilon^2 n^{-1} \sum_{j=1}^{\infty} a_j \lambda_j = \kappa$, where we denote by $\lambda_j = (1 - \kappa a_j)_+$. For $n = 100$, the solution is $\kappa = 0.199$. Determine the predictor dimension $p = \max_j\{a_j \le \kappa^{-1}\} = 23$. The variances are

$$\sigma_j^2 = \lambda_j (n\kappa a_j)^{-1},\ j = 1, \ldots, 23,\ n = 100.$$

It can be shown that such regression coefficients belong with high probability to $\{(\beta_{j,n})_j;\ \sum_{j=1}^{p_n} a_j^2 \beta_{j,n}^2 \le 1\}$ (note that $p = p_n$ depends on n via the parameter $\kappa = \kappa_n$).