Chapter 7
High-Dimensional Statistics

Jianqing Fan

7.1 Contributions of Peter Bickel to Statistical Learning

7.1.1 Introduction

Peter J. Bickel has made far-reaching and wide-ranging contributions to many areas of statistics. This short article highlights his marvelous contributions to high-dimensional statistical inference and machine learning, which range from novel methodological developments, deep theoretical analysis, and their applications. The focus is on the review and comments of his six recent papers in four areas, but only three of them are reproduced here due to limit of the space.

Information and technology make data collection and dissemination much easier over the last decade. High dimensionality and large data sets characterize many contemporary statistical problems from genomics and neural science to finance and economics, which give statistics and machine learning opportunities with challenges. These relatively new areas of statistical science encompass the majority of the frontiers and Peter Bickel is certainly a strong leader in those areas.

In response to the challenge of the complexity of data, new methods and greedy algorithms started to flourish in the 1990s and their theoretical properties were not well understood. Among those are the boosting algorithms and estimation of insintric dimensionality. In 2005, Peter Bickel and his coauthors gave deep theoretical foundation on boosting algorithms (Bickel et al. 2005; Freund and Schapire 1997) and novel methods on the estimation of intrinsic dimensionality (Levina and Bickel 2005). Another example is the use of LASSO (Tibshirani 1996) for high-dimensional variable selection. Realizing issues with biases of the

J. Fan (✉)
Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, 08540, USA
e-mail: jqfan@princeton.edu

J. Fan et al. (eds.), *Selected Works of Peter J. Bickel*, Selected Works in Probability and Statistics 13, DOI 10.1007/978-1-4614-5544-8_7,

Lasso estimate, Fan and Li (2001) advocated a family of folded concave penalties, including SCAD, to ameliorate the problem and critically analyzed its theoretical properties including LASSO. See also Fan and Lv (2011) for further analysis. Candes and Tao (2007) introduced the Dantzig selector. Zou and Li (2008) related the family folded-concave penalty with the adaptive LASSO (Zou 2006). It is Bickel et al. (2009) who critically analyzed the risk properties of the Lasso and the Dantzig selector, which significantly helps the statistics and machine learning communities on better understanding various variable selection procedures.

Covariance matrix is prominently featured in many statistical problems from network and graphic models to statistical inferences and portfolio management. Yet, estimating large covariance matrices is intrinsically challenging. How to reduce the number of parameters in a large covariance matrix is a challenging issue. In Economics and Finance, motivated by the arbitrage pricing theory, Fan et al. (2008) proposed to use the factor model to estimate the covariance matrix and its inverse. Yet, the impact of dimensionality is still very large. Bickel and Levina (2008a,b) and Rothman et al. (2008) proposed the use of sparsity, either on the covariance matrix or precision matrix, to reduce the dimensionality. The penalized likelihood method used in the paper fits in the generic framework of Fan and Li (2001) and Fan and Lv (2011), and the theory developed therein is applicable. Yet, Rothman et al. (2008) were able to utilize the specific structure of the covariance matrix and Gaussian distribution to get much deeper results. Realizing intensive computation of the penalized maximum likelihood method, Bickel and Levina (2008a,b) proposed a simple threshold estimator that achieves the same theoretical properties.

The papers will be reviewed in chronological order. They have high impacts on the subsequent development of statistics, applied mathematics, computer science, information theory, and signal processing. Despite young ages of those papers, a google-scholar search reveals that these six papers have around 900 citations. The impacts to broader scientific communities are evidenced!

7.1.2 Intrinsic Dimensionality

A general consensus is that high-dimensional data admits lower dimensional structure. The complexity of the data structure is characterized by the intrinsic dimensionality of the data, which is critical for manifold learning such as local linear embedding, Isomap, Lapacian and Hessian Eigenmaps (Brand 2002; Donoho and Grimes 2003; Roweis and Saul 2000; Tenenbaum et al. 2000). These nonlinear dimensionality reduction methods go behond traditional methods such as principal component analysis (PCA), which deals only with linear projections, and multidimensional scaling, which focuses on pairwise distances.

The techniques to estimate the intrinsic dimensionality before Levina and Bickel (2005) are roughly two groups: eigenvalue methods or geometric methods. The former are based on the number of eigenvalues greater than a given threshold. They fail on nonlinear manifolds. While localization enhances the applicability of

PCA, local methods depend strongly on the choice of local regions and thresholds (Verveer and Duin 1995). The latter exploit the geometry of the data. A popular metric is the correlation dimension from fractal analysis. Yet, there are a couple of parameters to be tuned.

The main contributions of Levina and Bickel (2005) are twofolds: It derives the maximum likelihood estimate (MLE) from a statistical prospective and gives its statistical properties. The MLE here is really the local MLE in the terminology of Fan and Gijbels (1996). Before this seminal work, there are virtually no formal statistical properties on the estimation of intrinsic dimensionality. The methods were often too heuristical and framework was not statistical.

The idea in Levina and Bickel (2005) is creative and statistical. Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be a random sample in R^p. They are embedded in an m-dimensional space via $\mathbf{X}_i = g(\mathbf{Y}_i)$, with unknown dimensionality m and unknown functions g, in which $\mathbf{Y}_i$ has a smooth density f in R^m. Because of nonlinear embedding g, we can only use the local data to determine m. Let R be small, which asymptotically goes to zero. Given a point $\mathbf{x}$ in R^p, the local information is summarized by the number of observations falling in the ball $\{\mathbf{z} : \|\mathbf{z} - \mathbf{x}\| \le t\}$, which is denoted by $N_{\mathbf{x}}(t)$, for $0 \le t \le R$. In other words, the local information around $\mathbf{x}$ with radius R is characterized by the process

$$\{N_{\mathbf{x}}(t) : 0 \le t \le R\}. \tag{7.1}$$

Clearly, $N_{\mathbf{x}}(t)$ is a binomial distribution with number of trial n and probability of success

$$P(\|\mathbf{X}_i - \mathbf{x}\| \le t) \approx f(\mathbf{x}) V(m) t^m, \quad \text{as } t \to 0, \tag{7.2}$$

where $V(m) = \pi^{m/2}[\Gamma(m/2+1)]^{-1}$ is the volume of the unit sphere in R^m. Recall that the approximation of the Binomial distribution by the Poison distribution. The process $\{N_{\mathbf{x}}(t) : 0 \le t \le R\}$ is approximately a Poisson process with the rate $\lambda(t)$, which is the derivative of (7.2), or more precisely

$$\lambda(t) = n f(\mathbf{x}) V(m) m t^{m-1} \tag{7.3}$$

The parameters $\theta = \log f(\mathbf{x})$ and m can be estimated by the maximum likelihood using the local observation (7.1).

Assuming $\{N_{\mathbf{x}}(t), 0 \le t \le R\}$ is the inhomogeneous Poisson process with rate $\lambda(t)$. Then, the log-likelihood of observing the process is given by

$$L(m, \theta) = \int_0^R \log \lambda(t) dN_{\mathbf{x}}(t) - \int_0^R \lambda(t) dt. \tag{7.4}$$

This can be understood by breaking the data $\{N_{\mathbf{x}}(t), 0 \le t \le R\}$ as the data

$$\{N(\Delta), N(2\Delta) - N(\Delta), \cdots, N(T\Delta) - N(T\Delta - \Delta)\}, \qquad \Delta = R/T \tag{7.5}$$

with a large T and noticing that the data above follow independent poisson distributions with mean $\lambda(j\Delta)\Delta$ for the j-th increment (The dependence on $\mathbf{x}$

is suppressed for brevity of notation). Therefore, using the Poisson formula, the likelihood of data (7.5) is

$$\prod_{j=1}^{T} \exp(-\lambda(j\Delta)\Delta)[\lambda(j\Delta)\Delta]^{dN(j\Delta)}/(dN(j\Delta)!)$$

where $dN(j\Delta) = N(j\Delta) - N(j\Delta - \Delta)$. Taking the logarithm and ignoring terms independent of the parameters, the log-likelihood of the observing data in (7.5) is

$$\sum_{j=1}^{T} [\log \lambda(j\Delta)] dN(j\Delta) - \sum_{j=1}^{T} \lambda(j\Delta)\Delta.$$

Taking the limit as $\Delta \to 0$, we obtain (7.4).

By taking the derivatives with parameters m and θ in (7.4) and setting them to zero, it is easy to obtain that

$$\hat{m}_R(\mathbf{x}) = \left\{ \log(R) - N_{\mathbf{x}}(R)^{-1} \int_0^R (\log t) dN_{\mathbf{x}}(t) \right\}^{-1}. \tag{7.6}$$

Let $T_k(\mathbf{x})$ be the distance of the k-th nearest point to $\mathbf{x}$. Then,

$$\hat{m}_R(\mathbf{x}) = \left\{ N_{\mathbf{x}}(R)^{-1} \sum_{j=1}^{N_{\mathbf{x}}(R)} \log[R/T_j(\mathbf{x})] \right\}^{-1}. \tag{7.7}$$

Now, instead of fixing distance R, but fixing the number of points k, namely, taking $R = T_k(\mathbf{x})$ for a given k, then, $N_{\mathbf{x}}(R) = k$ by definition and the estimator becomes

$$\hat{m}_k(\mathbf{x}) = \left\{ k^{-1} \sum_{j=1}^{k} \log[T_k(\mathbf{x})/T_j(\mathbf{x})] \right\}^{-1}. \tag{7.8}$$

Levina and Bickel (2005) realized that the parameter m is global whereas the estimate $\hat{m}_k(\mathbf{x})$ is local, depending on the location $\mathbf{x}$. They averaged out the n estimates at the observed data points and obtained

$$\hat{m}_k = n^{-1} \sum_{i=1}^{n} \hat{m}_k(\mathbf{X}_i). \tag{7.9}$$

To reduce the sensitivity on the choice of the parameter k, they proposed to use

$$\hat{m} = (k_2 - k_1 + 1)^{-1} \sum_{k=k_1}^{k_2} \hat{m}_k \tag{7.10}$$

for the given choices of k_1 and k_2.

The above discussion reveals that the parameter m was estimated in a semiparametric model in which $f(\mathbf{x})$ is fully nonparametric. Levina and Bickel (2005) estimates the global parameter m by averaging. Averaging reduces variances, but not biases. Therefore, it requires k to be small. However, when p is large, even with a small k, $T_k(\mathbf{x})$ can be large and so can be the bias. For semiparametric model, the work of Severini and Wong (1992) shows that the profile likelihood can have a better bias property. Inspired by that, an alternative version of the estimator is to use the global likelihood, which adds up the local likelihood (7.4) at each data point $\mathbf{X}_i$, i.e.

$$L(\theta_{\mathbf{x}_1}, \cdots, \theta_{\mathbf{x}_n}, m) = \sum_{i=1}^{n} L(\theta_{\mathbf{x}_i}, m). \tag{7.11}$$

Following the same derivations as in Levina and Bickel (2005), we obtain the maximum profile likelihood estimator

$$\hat{m}_R^* = \left\{ [\sum_{i=1}^{n} N_{\mathbf{x}_i}(R)]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{N_{\mathbf{x}_i}(R)} \log[R/T_j(\mathbf{x}_i)] \right\}^{-1}. \tag{7.12}$$

In its nearest neighbourhood form,

$$\hat{m}_k^* = \left\{ [n(k-2)]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} \log[T_k(\mathbf{x}_i)/T_j(\mathbf{x}_i)] \right\}^{-1}. \tag{7.13}$$

The reason for divisor $(k-2)$ instead of k is given in the next paragraph. It will be interesting to compare the performance of the method (7.13) with (7.9).

Levina and Bickel (2005) derived the asymptotic bias and variance of estimator (7.8). They advocated the normalization of (7.8) by $(k-2)$ rather than k. With this normalization, they derived that to the first order,

$$E(\hat{m}_k(\mathbf{x})) = m, \qquad \text{var}(\hat{m}_k(\mathbf{x})) = m^2/(k-3). \tag{7.14}$$

The paper has huge impact on manifold learning with a wide range of applications from patten analysis and object classification to machine learning and statistics. It has been cited nearly 200 times within 6 years of publication.

7.1.3 Generalized Boosting

Boosting is an iterative algorithm that uses a sequence of weak classifiers, which perform slightly better than a random guess, to build a stronger learner (Freund 1990; Schapire 1990), which can achieve the Bayes error rate. One of successful boosting algorithms is the AdaBoost by Freund and Schapire (1997). The algorithm

is powerful but appears heuristic at that time. It is Breiman (1998) who noted that the AdaBoost classifier can be viewed as a greedy algorithm for an empirical loss minimization. This makes a strong connection of the algorithm with statistical foundation that enables us to understand better theoretical properties.

Let $\{(\mathbf{X}_i, Y_i)\}_{i=1}^p$ be an i.i.d. sample where $Y_i \in \{-1, 1\}$. Let $\mathscr{H}$ be a set of weak learners. Breiman (1998) observed that the AdaBoost classifier is $\text{sgn}(F(\mathbf{X}))$, where F is found by a greedy algorithm minimizing

$$n^{-1} \sum_{i=1}^{n} \exp(-Y_i F(\mathbf{X}_i)), \tag{7.15}$$

over the class of function

$$\mathscr{F}_\infty = \bigcup_{k=1}^{\infty} \{ \sum_{j=1}^{k} \lambda_j h_j : \lambda_j \in \mathbb{R}, h_j \in \mathscr{H} \}.$$

The work of Bickel et al. (2005) generalizes the AdaBoost in two important directions: more general class of convex loss functions and more flexible class of algorithms. This enables them to study the convergence of the algorithms and classifiers in a unified framework. Let us state in the population version of their algorithms to simplify the notation. The goal is to find $F \in \mathscr{F}_\infty$ to minimize $w(F) = EW(YF)$ for a convex loss $W(\cdot)$. They proposed two relaxed Guass-Southwell algorithms, which are basically coordinatewise optimization algorithms in high-dimensional space. Given the current value F_m and coordinate h, one intends to minimize $W(F_m + \lambda h)$ over $\lambda \in \mathbb{R}$. The first algorithm is as follows: For given $\alpha \in (0, 1]$ and F_0, find inductively $F_1, F_2, \ldots$, by $F_{m+1} = F_m + \lambda_m h_m$, $\lambda_m \in \mathbb{R}$, $h_m \in \mathscr{H}$ such that

$$W(F_{m+1}) \le \alpha \min_{\lambda \in \mathbb{R}, h \in \mathscr{H}} W(F_m + \lambda h) + (1 - \alpha) W(F_m). \tag{7.16}$$

In particular, when λ_m and h_m minimize $W(F_m + \lambda h)$, then (7.16) is obviously satisfied with equality. The generalization covers the possibility that the minimum of $W(F_m + \lambda h)$ is not assumed or multiply assumed. The algorithm is very general in the sense that it does not even specify a way to find λ_m and h_m, but a necessary condition of (7.16) is that

$$W(F_{m+1}) \le W(F_m).$$

In other words, the target value decreases each iteration. The second algorithm is the same as the first one but requires

$$W(F_{m+1}) + \gamma \lambda_m^2 \le \alpha \min_{\lambda \in \mathbb{R}, h \in \mathscr{H}} [W(F_m + \lambda h) + \gamma \lambda^2] + (1 - \alpha) W(F_m). \tag{7.17}$$

Under such a broad class of algorithms, Bickel et al. (2005) demonstrated unambiguously and convincingly that the generalized boosting algorithm converges to the Bayes classifier. They further demonstrated that the generalized boosting

algorithms are consistent when the sample versions are used. In addition, they were able to derive the algorithmic speed of convergence, minimax rates of the convergence of the generalized boosting estimator to the Bayes classifier, and the minimax rates of the Bayes classification regret. The results are deep and useful. The work puts boosting algorithms in formal statistical framework and provides insightful understanding on the fundamental properties of the boosting algorithms.

Regularization of Covariation Matrices

It is well known that the sample covariance matrix has unexpected features when p and n are of the same order (Johnstone 2001; Marčcenko and Pastur 1967). Regularization is needed in order to obtain the desire statistical properties. Peter Bickel pioneered the work on the estimation of large covariance and led the development of the field through three seminal papers in 2008. Before Bickel's work, the theoretical work is very limited, often confining the dimensionality to be finite [with exception of Fan et al. (2008)], which does not reflect the nature of high-dimensionality. It is Bickel's work that allows the dimensionality grows much faster than sample size.

To regularize the covariance matrices, one needs to impose some sparsity conditions. The methods to explore sparsity are thresholding and the penalized quasi-likelihood approach. The former is frequently applied to the situations in which the sparsity is imposed on the elements which are directly estimable. For example, when the $p \times p$ covariance matrix Σ is sparse, a natural estimator is the following thresholding estimator

$$\hat{\Sigma}_t = (\hat{\sigma}_{i,j} I(|\hat{\sigma}_{i,j}| \geq t)) \tag{7.18}$$

for a thresholding parameter t. Bickel and Levina (2008b) considered a class of matrix

$$\left\{\Sigma : \sigma_{ii} \leq M, \quad \sum_{j=1}^{p} |\sigma_{ij}|^q \leq c_p, \forall i\right\}, \tag{7.19}$$

for $0 \leq q < 1$. In particular, when $q = 0$, c_p is the maximum number of nonvanishing elements in each row. They showed that when the data follow the Gaussian distribution and $t_n = M'(n^{-1}(\log p))^{1/2}$ for a sufficiently large constant M',

$$\|\hat{\Sigma}_{t_n} - \Sigma\| = O_p\left(c_p \left(n^{-1} \log p\right)^{(1-q)/2}\right), \tag{7.20}$$

and

$$p^{-1}\|\hat{\Sigma}_{t_n} - \Sigma\|_F^2 = O_p\left(c_p \left(n^{-1} \log p\right)^{1-q/2}\right). \tag{7.21}$$

uniformly for the class of matrices in (4.3), where $\|\mathbf{A}\|^2 = \lambda_{\max}(\mathbf{A}^T\mathbf{A})$ is the operator norm of a matrix $\mathbf{A}$ and $\|\mathbf{A}\|_F^2 = \sum_{i,j} a_{ij}^2$ is the Frobenius norm. Similar

results were derived when the distributions are sub-Gaussian or have finite moments or when t_n is chosen by cross-validation which is very technically challenging and novel. This along with Bickel and Levina (2008b) and El Karoui (2008) are the first results of this kind, allowing $p \gg n$, as long as c_p does not grow too fast.

When the covariance matrix admits a banded structure whose off-diagonal elements decay quickly:

$$\sum_{j:|i-j|>k} |\sigma_{ij}| \le Ck^{-\alpha}, \qquad \forall i \text{ and } k, \tag{7.22}$$

as arising frequently in time-series application including the covariance matrix of a weak-dependent stationary time series, Bickel and Levina (2008a) proposed a banding or more generally tapering to take advantage of prior sparsity structure. Let

$$\hat{\Sigma}_{B,k} = (\hat{\sigma}_{ij} I(|i-j| \le k)$$

be the banded sample covariance matrix. They showed that by taking $k_n \asymp (n^{-1}\log p)^{-1/(2(\alpha+1))}$,

$$\|\hat{\Sigma}_{B,k_n} - \hat{\Sigma}\| = O_p\left[(n^{-1}\log p)^{\alpha/(2(\alpha+1))}\right] = \|\hat{\Sigma}^{-1}_{B,k_n} - \hat{\Sigma}^{-1}\| \tag{7.23}$$

uniformly in the class of matrices (7.22) with additional restrictions that

$$c \le \lambda_{\min}(\Sigma) \le \lambda_{max}(\Sigma) \le C.$$

This again shows that large sparse covariance matrix can well be estimated even when $p \ge n$. The results are related to the estimation of spectral density (Fan and Gijbels 1996), but also allow non-stationary covariance matrices.

When the precision matrix $\Omega = \Sigma^{-1}$ is sparse, there is no easy way to apply thresholding rule. Hence, Rothman et al. (2008) appealed to the penalized likelihood method. Let $\ell_n(\theta)$ be the quasi-likelihood function based on a sample of size n and it is known that θ is sparse. Then, the penalized likelihood admits the form

$$\ell_n(\theta) + \sum_j p_\lambda(|\theta_j|). \tag{7.24}$$

Fan and Li (2001) advocated the use of folded-concave penalty p_λ to have a better bias property and put down a general theory. In particular, when the data $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are i.i.d. from $N(0, \Sigma)$, the penalized likelihood reduces to

$$\text{tr}(\Omega\hat{\Sigma}) - \log|\Omega| + \sum_{i,j} p_\lambda(|\omega_{ij}|), \tag{7.25}$$

where the matrix Ω is assumed to be sparse and is of primary interest. Rothman et al. (2008) utilized the fact that the diagonal elements are non-vanishing and

should not be penalized. They proposed the penalized likelihood estimator $\hat{\Omega}_\lambda$, which maximizes

$$\text{tr}(\Omega\hat{\Sigma}) - \log|\Omega| + \lambda \sum_{i \neq j} |\omega_{ij}|. \tag{7.26}$$

They showed that when $\lambda \asymp [(\log p)/n]^{1/2}$,

$$\|\hat{\Omega}_\lambda - \Omega\|_F^2 = O_P\left(\sqrt{\frac{(p+s)(\log p)}{n}}\right), \tag{7.27}$$

where s is the number of nonvanishing off diagonal elements. Note that there are $p + 2s$ nonvanishing elements in Ω and (7.27) reveals that each nonsparse element is estimated, on average, with rate $(n^{-1}(\log p))^{-1/2}$.

Note that thresholding and banding are very simple and easy to use. However, they are usually not semi-definite. Penalized likelihood can be used to enforce the positive definiteness in the optimization. It can also be applied to estimate sparse covariance matrices and sparse Chelosky decomposition; see Lam and Fan (2009).

The above three papers give us a comprehensive overview on the estimability of large covariance matrices. They have inspired many follow up work, including Levina et al. (2008), Lam and Fan (2009), Rothman et al. (2009), Cai et al. (2010), Cai and Liu (2011), and Cai and Zhou (2012), among others. In particular, the work inspires Fan et al. (2011) to propose an approximate factor model, allowing the idiosyncratic errors among financial assets to have a sparse covariance matrix, that widens significantly the scope and applicability of the strict factor model in finance. It also helps solving the aforementioned semi-definiteness issue, due to thresholding.

7.1.4 Variable Selections

Peter Bickel contributions to high-dimensional regression are highlighted by his paper with Ritov and Tsybakov (Bickel et al. 2009) on the analysis of the risk properties of the LASSO and Dantzig selector. This is done in least-squares setting on the nonparametric regression via basis approximations (approximate linear model) or linear model itself. This is based the following important observations in Bickel et al. (2009).

Recall that the LASSO estimator $\hat{\beta}_L$ minimizes

$$(2n)^{-1}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{7.28}$$

A necessary condition is that 0 belongs to the subgradient of the function (7.28), which is the same as

$$\|n^{-1}\mathbf{X}(\mathbf{Y}-\mathbf{X}\hat{\beta}_L)\|_\infty \le \lambda. \tag{7.29}$$

The Danzig selector (Candes and Tao 2007) is defined by

$$\hat{\beta}_D = \text{argmin}\Big\{\|\beta\|_1 : \|n^{-1}\mathbf{X}(\mathbf{Y}-\mathbf{X}\beta)\|_\infty \le \lambda\Big\}. \tag{7.30}$$

Thus, $\hat{\beta}_D$ satisfies (7.29), having a smaller L_1-norm than LASSO, by definition. They also show that for both the Lasso and the Danzig estimator, their estimation error δ satisfies

$$\|\delta_{J^c}\|_1 \le c\|\delta_J\|_1$$

with probability close to 1, where J is the subset of non-vanishing true regression coefficients. This leads them to define restricted eigenvalue assumptions.

For linear model, Bickel et al. (2009) established the convergence rates of

$$\|\hat{\beta}_D-\beta\|_p \text{ for } p\in[1,2] \quad \text{and} \quad \|\mathbf{X}(\hat{\beta}_D-\beta)\|_2. \tag{7.31}$$

The former is on the convergence rate of the estimator and the latter is on the prediction risk of the estimator. They also established the rate of convergence for the Lasso estimator. Both estimators admit the same rate of convergence under the same conditions. Similar results hold when the method is applied to nonparametric regression. This leads Bickel et al. (2009) to conclude that both the Danzig selector and Lasso estimator are equivalent.

The contributions of the paper are multi-fold. First of all, it provides a good understanding on the performance of the newly invented Danzig estimator and its relation to the Lasso estimator. Secondly, it introduced new technical tools for the analysis of penalized least-squares estimator. Thirdly, it derives various new results, including oracle inequalities, for the Lasso and the Danzig selector in both linear model and nonparametric regression model. The work has a strong impact on the recent development of the high-dimensional statistical learning. Within 3 years of its publications, it has been cited around 300 times!

References

Bickel PJ, Levina E (2008a) Regularized estimation of large covariance matrices. Ann Stat 36: 199–227

Bickel PJ, Levina E (2008b) Covariance regularization by thresholding. Ann Stat 36:2577–2604

Bickel PJ, Ritov Y, Zakai A (2005) Some theory for generalized boosting algorithms. J Mach Learn Res 7:705–732

Bickel PJ, Ritov Y, Tsybakov A (2009) Simultaneous analysis of Lasso and Dantzig selector. Ann Statist 37:1705–1732.

Brand M (2002) Charting a manifold. In: Advances in NIPS, vol 14. MIT Press, Cambridge, MA
Breiman L (1998) Arcing classifiers (with discussion). Ann Stat 26:801–849
Cai T, Liu W (2011) Adaptive thresholding for sparse covariance matrix estimation. J Am Stat Assoc 494:672–684
Cai T, Zhou H (2012) Minimax estimation of large covariance matrices under ℓ_1 norm (with discussion). Stat Sin, to appear
Cai T, Zhang C-H, Zhou H (2010) Optimal rates of convergence for covariance matrix estimation Ann Stat 38:2118–2144
Candes E, Tao T (2007) The Dantzig selector: statistical estimation when p is much larger than n (with discussion). Ann Stat 35:2313–2404
Donoho DL, Grimes C (2003) Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. Technical Report TR 2003-08, Department of Statistics, Stanford University, 2003
El Karoui N (2008) Operator norm consistent estimation of large-dimensional sparse covariance matrices. Ann Stat 36:2717–2756
Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman and Hall, London
Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360
Fan J, Lv J (2011) Non-concave penalized likelihood with NP-dimensionality. IEEE Inf Theory 57:5467–5484
Fan J, Fan Y, Lv J, (2008) Large dimensional covariance matrix estimation via a factor model. J Econ 147:186–197
Fan J, Liao Y, Mincheva M (2011) High dimensional covariance matrix estimation in approximate factor models. Ann Statist 39:3320–3356
Freund Y (1990) Boosting a weak learning algorithm by majority. In: Proceedings of the third annual workshop on computational learning theory. Morgan Kaufmann, San Mateo
Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55:119–139
Johnstone IM (2001) On the distribution of the largest eigenvalue in principal components analysis. Ann Stat 29:295–327
Lam C, Fan J (2009) Sparsistency and rates of convergence in large covariance matrices estimation. Ann Stat 37:4254–4278
Levina E, Bickel PJ (2005) Maximum likelihood estimation of intrinsic dimension. In: Saul LK, Weiss Y, Bottou L (eds) Advances in NIPS, vol 17. MIT Press, Cambridge, MA
Levina E, Rothman AJ, Zhu J (2008) Sparse estimation of large covariance matrices via a nested lasso penalty. Ann Stat Appl Stat 2:245–263
Marčcenko VA, Pastur LA (1967) Distributions of eigenvalues of some sets of random matrices. Math USSR-Sb 1:507–536
Rothman AJ, Bickel PJ, Levina E, Zhu J (2008) Sparse permutation invariant covariance estimation. Electron J Stat 2:494–515
Rothman AJ, Levina E, Zhu J (2009) Generalized thresholding of large covariance matrices. J Am Stat Assoc 104:177–186
Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
Schapire R (1990) Strength of weak learnability. Mach Learn 5:197–227
Severini TA, Wong WH (1992) Generalized profile likelihood and conditional parametric models. Ann Stat 20:1768–1802
Tenenbaum JB, de Silva V, Landford JC(2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323

Tibshirani R (1996) Regression shrinkage and selection via lasso. J R Stat Soc B 58:267–288
Verveer P, Duin R (1995) An evaluation of intrinsic dimensionality estimators. IEEE Trans PAMI 17:81–86
Zou H (2006) The adaptive Lasso and its oracle properties. J Am Stat Assoc 101:1418–1429
Zou H, Li R (2008) One-step sparse estimates in nonconcave penalized likelihood models (with discussion). Ann Stat 36:1509–1533

Maximum Likelihood Estimation of Intrinsic Dimension

Elizaveta Levina
Department of Statistics
University of Michigan
Ann Arbor MI 48109-1092
elevina@umich.edu

Peter J. Bickel
Department of Statistics
University of California
Berkeley CA 94720-3860
bickel@stat.berkeley.edu

Abstract

We propose a new method for estimating intrinsic dimension of a dataset derived by applying the principle of maximum likelihood to the distances between close neighbors. We derive the estimator by a Poisson process approximation, assess its bias and variance theoretically and by simulations, and apply it to a number of simulated and real datasets. We also show it has the best overall performance compared with two other intrinsic dimension estimators.

1 Introduction

There is a consensus in the high-dimensional data analysis community that the only reason any methods work in very high dimensions is that, in fact, the data are not truly high-dimensional. Rather, they are embedded in a high-dimensional space, but can be efficiently summarized in a space of a much lower dimension, such as a nonlinear manifold. Then one can reduce dimension without losing much information for many types of real-life high-dimensional data, such as images, and avoid many of the "curses of dimensionality". Learning these data manifolds can improve performance in classification and other applications, but if the data structure is complex and nonlinear, dimensionality reduction can be a hard problem.

Traditional methods for dimensionality reduction include principal component analysis (PCA), which only deals with linear projections of the data, and multidimensional scaling (MDS), which aims at preserving pairwise distances and traditionally is used for visualizing data. Recently, there has been a surge of interest in manifold projection methods (Locally Linear Embedding (LLE) [1], Isomap [2], Laplacian and Hessian Eigenmaps [3, 4], and others), which focus on finding a nonlinear low-dimensional embedding of high-dimensional data. So far, these methods have mostly been used for exploratory tasks such as visualization, but they have also been successfully applied to classification problems [5, 6].

The dimension of the embedding is a key parameter for manifold projection methods: if the dimension is too small, important data features are "collapsed" onto the same dimension, and if the dimension is too large, the projections become noisy and, in some cases, unstable. There is no consensus, however, on how this dimension should be determined. LLE [1] and its variants assume the manifold dimension

is provided by the user. Isomap [2] provides error curves that can be "eyeballed" to estimate dimension. The charting algorithm, a recent LLE variant [7], uses a heuristic estimate of dimension which is essentially equivalent to the regression estimator of [8] discussed below. Constructing a reliable estimator of intrinsic dimension and understanding its statistical properties will clearly facilitate further applications of manifold projection methods and improve their performance.

We note that for applications such as classification, cross-validation is in principle the simplest solution – just pick the dimension which gives the lowest classification error. However, in practice the computational cost of cross-validating for the dimension is prohibitive, and an estimate of the intrinsic dimension will still be helpful, either to be used directly or to narrow down the range for cross-validation.

In this paper, we present a new estimator of intrinsic dimension, study its statistical properties, and compare it to other estimators on both simulated and real datasets. Section 2 reviews previous work on intrinsic dimension. In Section 3 we derive the estimator and give its approximate asymptotic bias and variance. Section 4 presents results on datasets and compares our estimator to two other estimators of intrinsic dimension. Section 5 concludes with discussion.

2 Previous Work on Intrinsic Dimension Estimation

The existing approaches to estimating the intrinsic dimension can be roughly divided into two groups: eigenvalue or projection methods, and geometric methods. Eigenvalue methods, from the early proposal of [9] to a recent variant [10] are based on a global or local PCA, with intrinsic dimension determined by the number of eigenvalues greater than a given threshold. Global PCA methods fail on nonlinear manifolds, and local methods depend heavily on the precise choice of local regions and thresholds [11]. The eigenvalue methods may be a good tool for exploratory data analysis, where one might plot the eigenvalues and look for a clear-cut boundary, but not for providing reliable estimates of intrinsic dimension.

The geometric methods exploit the intrinsic geometry of the dataset and are most often based on fractal dimensions or nearest neighbor (NN) distances. Perhaps the most popular fractal dimension is the correlation dimension [12, 13]: given a set $S_n = \{x_1, \ldots, x_n\}$ in a metric space, define

$$C_n(r) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbf{1}\{\|x_i - x_j\| < r\}. \tag{1}$$

The correlation dimension is then estimated by plotting $\log C_n(r)$ against $\log r$ and estimating the slope of the linear part [12]. A recent variant [13] proposed plotting this estimate against the true dimension for some simulated data and then using this calibrating curve to estimate the dimension of a new dataset. This requires a different curve for each n, and the choice of calibration data may affect performance. The capacity dimension and packing numbers have also been used [14]. While the fractal methods successfully exploit certain geometric aspects of the data, the statistical properties of these methods have not been studied.

The correlation dimension (1) implicitly uses NN distances, and there are methods that focus on them explicitly. The use of NN distances relies on the following fact: if $X_1, \ldots, X_n$ are an independent identically distributed (i.i.d.) sample from a density $f(x)$ in $\mathbb{R}^m$, and $T_k(x)$ is the Euclidean distance from a fixed point x to its k-th NN in the sample, then

$$\frac{k}{n} \approx f(x)V(m)[T_k(x)]^m, \tag{2}$$

where $V(m) = \pi^{m/2}[\Gamma(m/2+1)]^{-1}$ is the volume of the unit sphere in $\mathbb{R}^m$. That is, the proportion of sample points falling into a ball around x is roughly $f(x)$ times the volume of the ball.

The relationship (2) can be used to estimate the dimension by regressing $\log \bar{T}_k$ on $\log k$ over a suitable range of k, where $\bar{T}_k = n^{-1}\sum_{i=1}^{n} T_k(X_i)$ is the average of distances from each point to its k-th NN [8, 11]. A comparison of this method to a local eigenvalue method [11] found that the NN method suffered more from underestimating dimension for high-dimensional datasets, but the eigenvalue method was sensitive to noise and parameter settings. A more sophisticated NN approach was recently proposed in [15], where the dimension is estimated from the length of the minimal spanning tree on the geodesic NN distances computed by Isomap.

While there are certainly existing methods available for estimating intrinsic dimension, there are some issues that have not been adequately addressed. The behavior of the estimators as a function of sample size and dimension is not well understood or studied beyond the obvious "curse of dimensionality"; the statistical properties of the estimators, such as bias and variance, have not been looked at (with the exception of [15]); and comparisons between methods are not always presented.

3 A Maximum Likelihood Estimator of Intrinsic Dimension

Here we derive the maximum likelihood estimator (MLE) of the dimension m from i.i.d. observations $X_1, \ldots, X_n$ in $\mathbb{R}^p$. The observations represent an embedding of a lower-dimensional sample, i.e., $X_i = g(Y_i)$, where Y_i are sampled from an unknown smooth density f on $\mathbb{R}^m$, with unknown $m \leq p$, and g is a continuous and sufficiently smooth (but not necessarily globally isometric) mapping. This assumption ensures that close neighbors in $\mathbb{R}^m$ are mapped to close neighbors in the embedding.

The basic idea is to fix a point x, assume $f(x) \approx$ const in a small sphere $S_x(R)$ of radius R around x, and treat the observations as a homogeneous Poisson process in $S_x(R)$. Consider the inhomogeneous process $\{N(t,x),\, 0 \leq t \leq R\}$,

$$N(t,x) = \sum_{i=1}^{n} \mathbf{1}\{X_i \in S_x(t)\} \tag{3}$$

which counts observations within distance t from x. Approximating this binomial (fixed n) process by a Poisson process and suppressing the dependence on x for now, we can write the rate $\lambda(t)$ of the process $N(t)$ as

$$\lambda(t) = f(x)V(m)mt^{m-1} \tag{4}$$

This follows immediately from the Poisson process properties since $V(m)mt^{m-1} = \frac{d}{dt}[V(m)t^m]$ is the surface area of the sphere $S_x(t)$. Letting $\theta = \log f(x)$, we can write the log-likelihood of the observed process $N(t)$ as (see e.g., [16])

$$L(m,\theta) = \int_0^R \log\lambda(t)\, dN(t) - \int_0^R \lambda(t)\, dt$$

This is an exponential family for which MLEs exist with probability $\to 1$ as $n \to \infty$ and are unique. The MLEs must satisfy the likelihood equations

$$\frac{\partial L}{\partial \theta} = \int_0^R dN(t) - \int_0^R \lambda(t)dt = N(R) - e^{\theta}V(m)R^m = 0, \tag{5}$$

$$\frac{\partial L}{\partial m} = \left(\frac{1}{m} + \frac{V'(m)}{V(m)}\right) N(R) + \int_0^R \log t\, dN(t) - \\ -e^{\theta}V(m)R^m\left(\log R + \frac{V'(m)}{V(m)}\right) = 0. \tag{6}$$

Substituting (5) into (6) gives the MLE for m:

$$\hat{m}_R(x) = \left[\frac{1}{N(R,x)} \sum_{j=1}^{N(R,x)} \log \frac{R}{T_j(x)} \right]^{-1} . \tag{7}$$

In practice, it may be more convenient to fix the number of neighbors k rather than the radius of the sphere R. Then the estimate in (7) becomes

$$\hat{m}_k(x) = \left[\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(x)}{T_j(x)} \right]^{-1} . \tag{8}$$

Note that we omit the last (zero) term in the sum in (7). One could divide by $k-2$ rather than $k-1$ to make the estimator asymptotically unbiased, as we show below. Also note that the MLE of θ can be used to obtain an instant estimate of the entropy of f, which was also provided by the method used in [15].

For some applications, one may want to evaluate local dimension estimates at every data point, or average estimated dimensions within data clusters. We will, however, assume that all the data points come from the same "manifold", and therefore average over all observations.

The choice of k clearly affects the estimate. It can be the case that a dataset has different intrinsic dimensions at different scales, e.g., a line with noise added to it can be viewed as either 1-d or 2-d (this is discussed in detail in [14]). In such a case, it is informative to have different estimates at different scales. In general, for our estimator to work well the sphere should be small and contain sufficiently many points, and we have work in progress on choosing such a k automatically. For this paper, though, we simply average over a range of small to moderate values $k = k_1 \dots k_2$ to get the final estimates

$$\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_k(X_i) , \qquad \hat{m} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \hat{m}_k . \tag{9}$$

The choice of k_1 and k_2 and behavior of $\hat{m}_k$ as a function of k are discussed further in Section 4. The only parameters to set for this method are k_1 and k_2, and the computational cost is essentially the cost of finding k_2 nearest neighbors for every point, which has to be done for most manifold projection methods anyway.

3.1 Asymptotic behavior of the estimator for m fixed, $n \to \infty$.

Here we give a sketchy discussion of the asymptotic bias and variance of our estimator, to be elaborated elsewhere. The computations here are under the assumption that m is fixed, $n \to \infty, k \to \infty$, and $k/n \to 0$.

As we remarked, for a given x if $n \to \infty$ and $R \to 0$, the inhomogeneous binomial process $N(t,x)$ in (3) converges weakly to the inhomogeneous Poisson process with rate $\lambda(t)$ given by (4). If we condition on the distance $T_k(x)$ and assume the Poisson approximation is exact, then $\left\{ m^{-1} \log(T_k/T_j) : 1 \le j \le k-1 \right\}$ are distributed as the order statistics of a sample of size $k-1$ from a standard exponential distribution. Hence $U = m^{-1} \sum_{j=1}^{k-1} \log(T_k/T_j)$ has a Gamma$(k-1,1)$ distribution, and $EU^{-1} = 1/(k-2)$. If we use $k-2$ to normalize, then under these assumptions, to a first order approximation

$$E\left(\hat{m}_k(x)\right) = m, \quad \mathrm{Var}\left(\hat{m}_k(x)\right) = \frac{m^2}{k-3} \tag{10}$$

As this analysis is asymptotic in both k and n, the factor $(k-1)/(k-2)$ makes no difference. There are, of course, higher order terms since $N(t,x)$ is in fact a binomial process with $EN(t,x) = \lambda(t)\left(1+O(t^2)\right)$, where $O(t^2)$ depends on m.

With approximations (10), we have $E\hat{m} = E\hat{m}_k = m$, but the computation of $\mathrm{Var}(\hat{m})$ is complicated by the dependence among $\hat{m}_k(X_i)$. We have a heuristic argument (omitted for lack of space) that, by dividing $\hat{m}_k(X_i)$ into n/k roughly independent groups of size k each, the variance can be shown to be of order n^{-1}, as it would if the estimators were independent. Our simulations confirm that this approximation is reasonable – for instance, for m-d Gaussians the ratio of the theoretical SD $= C(k_1,k_2)m/\sqrt{n}$ (where $C(k_1,k_2)$ is calculated as if all the terms in (9) were independent) to the actual SD of $\hat{m}$ was between 0.7 and 1.3 for the range of values of m and n considered in Section 4. The bias, however, behaves worse than the asymptotics predict, as we discuss further in Section 5.

4 Numerical Results

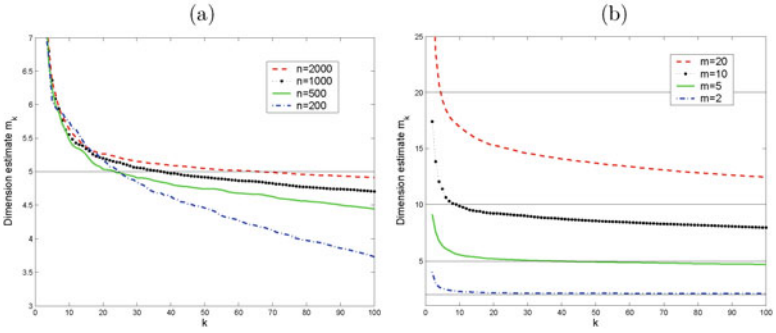

Figure 1: The estimator $\hat{m}_k$ as a function of k. (a) 5-dimensional normal for several sample sizes. (b) Various m-dimensional normals with sample size $n = 1000$.

We first investigate the properties of our estimator in detail by simulations, and then apply it to real datasets. The first issue is the behavior of $\hat{m}_k$ as a function of k. The results shown in Fig. 1 are for m-d Gaussians $N_m(0,I)$, and a similar pattern holds for observations in a unit cube, on a hypersphere, and on the popular "Swiss roll" manifold. Fig. 1(a) shows $\hat{m}_k$ for a 5-d Gaussian as a function of k for several sample sizes n. For very small k the approximation does not work yet and $\hat{m}_k$ is unreasonably high, but for k as small as 10, the estimate is near the true value $m = 5$. The estimate shows some negative bias for large k, which decreases with growing sample size n, and, as Fig. 1(b) shows, increases with dimension. Note, however, that it is the intrinsic dimension m rather than the embedding dimension $p \geq m$ that matters; and as our examples below and many examples elsewhere show, the intrinsic dimension for real data is frequently low.

The plots in Fig. 1 show that the "ideal" range $k_1 \ldots k_2$ is different for every combination of m and n, but the estimator is fairly stable as a function of k, apart from the first few values. While fine-tuning the range $k_1 \ldots k_2$ for different n is possible and would reduce the bias, for simplicity and reproducibility of our results we fix $k_1 = 10$, $k_2 = 20$ throughout this paper. In this range, the estimates are not

affected much by sample size or the positive bias for very small k, at least for the range of m and n under consideration.

Next, we investigate an important and often overlooked issue of what happens when the data are near a manifold as opposed to exactly on a manifold. Fig. 2(a) shows simulation results for a 5-d correlated Gaussian with mean 0, and covariance matrix $[\sigma_{ij}] = [\rho + (1-\rho)\delta_{ij}]$, with $\delta_{ij} = \mathbf{1}\{i = j\}$. As ρ changes from 0 to 1, the dimension changes from 5 (full spherical Gaussian) to 1 (a line in $\mathbb{R}^5$), with intermediate values of ρ providing noisy versions.

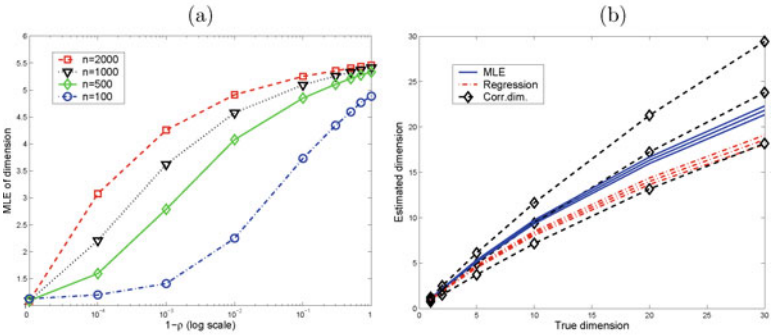

Figure 2: (a) Data near a manifold: estimated dimension for correlated 5-d normal as a function of $1-\rho$. (b) The MLE, regression, and correlation dimension for uniform distributions on spheres with $n = 1000$. The three lines for each method show the mean ± 2 SD (95% confidence intervals) over 1000 replications.

The plots in Fig. 2(a) show that the MLE of dimension does not drop unless ρ is very close to 1, so the estimate is not affected by whether the data cloud is spherical or elongated. For ρ close to 1, when the dimension really drops, the estimate depends significantly on the sample size, which is to be expected: $n = 100$ highly correlated points look like a line, but $n = 2000$ points fill out the space around the line. This highlights the fundamental dependence of intrinsic dimension on the neighborhood scale, particularly when the data may be observed with noise. The MLE of dimension, while reflecting this dependence, behaves reasonably and robustly as a function of both ρ and n.

A comparison of the MLE, the regression estimator (regressing $\log \overline{T}_k$ on $\log k$), and the correlation dimension is shown in Fig. 2(b). The comparison is shown on uniformly distributed points on the surface of an $m-$dimensional sphere, but a similar pattern held in all our simulations. The regression range was held at $k = 10 \ldots 20$ (the same as the MLE) for fair comparison, and the regression for correlation dimension was based on the first $10 \ldots 100$ distinct values of $\log C_n(r)$, to reflect the fact there are many more points for the $\log C_n(r)$ regression than for the $\log \overline{T}_k$ regression. We found in general that the correlation dimension graph can have more than one linear part, and is more sensitive to the choice of range than either the MLE or the regression estimator, but we tried to set the parameters for all methods in a way that does not give an unfair advantage to any and is easily reproducible.

The comparison shows that, while all methods suffer from negative bias for higher dimensions, the correlation dimension has the smallest bias, with the MLE coming

in close second. However, the variance of correlation dimension is much higher than that of the MLE (the SD is at least 10 times higher for *all* dimensions). The regression estimator, on the other hand, has relatively low variance (though always higher than the MLE) but the largest negative bias. On the balance of bias and variance, MLE is clearly the best choice.

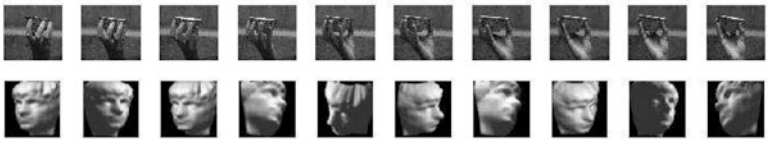

Figure 3: Two image datasets: hand rotation and Isomap faces (example images).

Table 1: Estimated dimensions for popular manifold datasets. For the Swiss roll, the table gives mean(SD) over 1000 uniform samples.

Dataset	Data dim.	Sample size	MLE	Regression	Corr. dim.
Swiss roll	3	1000	2.1(0.02)	1.8(0.03)	2.0(0.24)
Faces	64×64	698	4.3	4.0	3.5
Hands	480×512	481	3.1	2.5	3.9[1]

Finally, we compare the estimators on three popular manifold datasets (Table 1): the Swiss roll, and two image datasets shown on Fig. 3: the Isomap face database[2], and the hand rotation sequence[3] used in [14]. For the Swiss roll, the MLE again provides the best combination of bias and variance.

The face database consists of images of an artificial face under three changing conditions: illumination, and vertical and horizontal orientation. Hence the intrinsic dimension of the dataset should be 3, but only if we had the full 3-d images of the face. All we have, however, are 2-d projections of the face, and it is clear that one needs more than one "basis" image to represent different poses (from casual inspection, front view and profile seem sufficient). The estimated dimension of about 4 is therefore very reasonable.

The hand image data is a real video sequence of a hand rotating along a 1-d curve in space, but again several basis 2-d images are needed to represent different poses (in this case, front, back, and profile seem sufficient). The estimated dimension around 3 therefore seems reasonable. We note that the correlation dimension provides two completely different answers for this dataset, depending on which linear part of the curve is used; this is further evidence of its high variance, which makes it a less reliable estimate that the MLE.

5 Discussion

In this paper, we have derived a maximum likelihood estimator of intrinsic dimension and some asymptotic approximations to its bias and variance. We have shown

[1]This estimate is obtained from the range 500...1000. For this dataset, the correlation dimension curve has two distinct linear parts, with the first part over the range we would normally use, 10...100, producing dimension 19.7, which is clearly unreasonable.

[2]`http://isomap.stanford.edu/datasets.html`

[3]`http://vasc.ri.cmu.edu//idb/html/motion/hand/index.html`

that the MLE produces good results on a range of simulated and real datasets and outperforms two other dimension estimators. It does, however, suffer from a negative bias for high dimensions, which is a problem shared by all dimension estimators. One reason for this is that our approximation is based on sufficiently many observations falling into a small sphere, and that requires very large sample sizes in high dimensions (we shall elaborate and quantify this further elsewhere). For some datasets, such as points in a unit cube, there is also the issue of edge effects, which generally become more severe in high dimensions. One can potentially reduce the negative bias by removing the edge points by some criterion, but we found that the edge effects are small compared to the sample size problem, and we have been unable to achieve significant improvement in this manner. Another option used by [13] is calibration on simulated datasets with known dimension, but since the bias depends on the sampling distribution, and a different curve would be needed for every sample size, calibration does not solve the problem either. One should keep in mind, however, that for most interesting applications intrinsic dimension will not be very high – otherwise there is not much benefit in dimensionality reduction; hence in practice the MLE will provide a good estimate of dimension most of the time.

References

[1] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[2] J. B. Tenenbaum, V. de Silva, and J. C. Landford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in NIPS*, volume 14. MIT Press, 2002.

[4] D. L. Donoho and C. Grimes. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. Technical Report TR 2003-08, Department of Statistics, Stanford University, 2003.

[5] M. Belkin and P. Niyogi. Using manifold structure for partially labelled classification. In *Advances in NIPS*, volume 15. MIT Press, 2003.

[6] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas. Non-linear dimensionality reduction techniques for classification and visualization. In *Proceedings of 8th SIGKDD*, pages 645–651. Edmonton, Canada, 2002.

[7] M. Brand. Charting a manifold. In *Advances in NIPS*, volume 14. MIT Press, 2002.

[8] K.W. Pettis, T.A. Bailey, A.K. Jain, and R.C. Dubes. An intrinsic dimensionality estimator from near-neighbor information. *IEEE Trans. on PAMI*, 1:25–37, 1979.

[9] K. Fukunaga and D.R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Trans. on Computers*, C-20:176–183, 1971.

[10] J. Bruske and G. Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE Trans. on PAMI*, 20(5):572–575, 1998.

[11] P. Verveer and R. Duin. An evaluation of intrinsic dimensionality estimators. *IEEE Trans. on PAMI*, 17(1):81–86, 1995.

[12] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica*, D9:189–208, 1983.

[13] F. Camastra and A. Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based approach. *IEEE Trans. on PAMI*, 24(10):1404–1407, 2002.

[14] B. Kegl. Intrinsic dimension estimation using packing numbers. In *Advances in NIPS*, volume 14. MIT Press, 2002.

[15] J. Costa and A. O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Trans. on Signal Processing*, 2004. To appear.

[16] D. L. Snyder. *Random Point Processes*. Wiley, New York, 1975.

Journal of Machine Learning Research 7 (2006) 705–732 Submitted 7/05; Revised 12/05; Published 5/06

Some Theory for Generalized Boosting Algorithms

Peter J. Bickel BICKEL@STAT.BERKELEY.EDU
Department of Statistics
University of California at Berkeley
Berkeley, CA 94720, USA

Ya'acov Ritov (corresponding author) YAACOV.RITOV@HUJI.AC.IL
Department of Statistics and The Interdisciplinary Center for Neural Computation
The Hebrew University of Jerusalem
91905 Jerusalem, Israel

Alon Zakai ALONZAKA@POB.HUJI.AC.IL
The Interdisciplinary Center for Neural Computation
The Hebrew University of Jerusalem
91904 Jerusalem, Israel

Editor: Bin Yu

Abstract

We give a review of various aspects of boosting, clarifying the issues through a few simple results, and relate our work and that of others to the minimax paradigm of statistics. We consider the population version of the boosting algorithm and prove its convergence to the Bayes classifier as a corollary of a general result about Gauss-Southwell optimization in Hilbert space. We then investigate the algorithmic convergence of the sample version, and give bounds to the time until perfect separation of the sample. We conclude by some results on the statistical optimality of the L_2 boosting.

Keywords: classification, Gauss-Southwell algorithm, AdaBoost, cross-validation, non-parametric convergence rate

1. Introduction

We consider a standard classification problem: Let $(X,Y),(X_1,Y_1),\ldots,(X_n,Y_n)$ be an i.i.d. sample, where $Y_i \in \{-1,1\}$ and $X_i \in \mathcal{X}$. The goal is to find a good classification rule, $\mathcal{X} \to \{-1,1\}$.

The AdaBoost algorithm was originally defined, Schapire (1990), Freund (1995), and Freund and Schapire (1996) as an algorithm to construct a good classifier by a "weighted majority vote" of simple classifiers. To be more exact, let $\mathcal{H}$ be a set of simple classifiers. The AdaBoost classifier is given by $\operatorname{sgn}\left(\sum_{m=1}^{M} \lambda_m h_m(x)\right)$, where $\lambda_m \in \mathbb{R}$, $h_m \in \mathcal{H}$, are found sequentially by the following algorithm:

0. Let $c_1 = c_2 = \cdots = c_n = 1$, and set $m = 1$.

1. Find $h_m = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^{n} c_i h(X_i) Y_i$. Set

$$\lambda_m = \frac{1}{2}\log\left(\frac{\sum_{i=1}^{n} c_i + \sum_{i=1}^{n} c_i h_m(X_i) Y_i}{\sum_{i=1}^{n} c_i - \sum_{i=1}^{n} c_i h_m(X_i) Y_i}\right) = \frac{1}{2}\log\left(\frac{\sum_{h_m(X_i)=Y_i} c_i}{\sum_{h_m(X_i)\neq Y_i} c_i}\right).$$

2. Set $c_i \leftarrow c_i \exp\bigl(-\lambda_m h_m(X_i)Y_i\bigr)$, and $m \leftarrow m+1$, If $m \le M$, return to step 1.

M is unspecified and can be arbitrarily large.

The success of these methods on many data sets and their "resistance to overfitting"—the test set error continues to decrease even after all the training set observations were classified correctly, has led to intensive investigation to which this paper contributes.

Let $\mathcal{F}_\infty$ be the linear span of $\mathcal{H}$. That is,

$$\mathcal{F}_\infty = \bigcup_{k=1}^{\infty} \mathcal{F}_k, \text{ where } \mathcal{F}_k = \Bigl\{ \sum_{j=1}^{k} \lambda_j h_j : \lambda_j \in \mathbb{R},\ h_j \in \mathcal{H},\ 1 \le j \le k \Bigr\}.$$

A number of workers have noted, Breiman (1998,1999), Friedman, Hastie and Tibshirani (2000), Mason, Bartlett, Baxter and Frean (2000), and Schapire and Singer (1999), that the AdaBoost classifier can be viewed as $\operatorname{sgn}\bigl(F(X)\bigr)$, where F is found by a greedy algorithm minimizing

$$n^{-1} \sum_{i=1}^{n} \exp\bigl(-Y_i F(X_i)\bigr)$$

over $\mathcal{F}_\infty$.

¿From this point of view, the algorithm appeared to be justifiable, since as was noted in Breiman (1999) and Friedman, Hastie, and Tibshirani (2000), the corresponding expression $E\exp\bigl(-YF(X)\bigr)$, obtained by replacing the sum by expectation, is minimized by

$$F(X) = \frac{1}{2}\log\Bigl(P(Y=1|X)/P(Y=-1|X)\Bigr),$$

provided the linear span $\mathcal{F}_\infty$ is dense in the space $\mathcal{F}$ of all functions in a suitable way. However, it was also noted that the empirical optimization problem necessarily led to rules which would classify every training set observation correctly and hence not approach the Bayes rule whatever be n, except in very special cases. Jiang (2003) established that, for observation centered stumps, the algorithm converged to nearest neighbor classification, a good but rarely optimal rule.

In another direction, the class of objective functions $W(\cdot)$ that can be considered was extended by Friedman, Hastie, and Tibshirani (2000) to other W, in particular, $W(t) = \log(1+e^{-2t})$, whose empirical version they identified with logistic regression in statistics, and $W(t) = -2t + t^2$, which they referred to as "L_2 Boosting" and has been studied, under the name "matching pursuit", in the signal processing community. For all these objective functions, the population optimization of $EW\bigl(YF(X)\bigr)$ over $\mathcal{F}$ leads to a solution such that $\operatorname{sgn}F(X)$ is the Bayes rule. Friedman et al. also introduced consideration of other algorithms for the empirical optimization problem. Lugosi and Vayatis (2004) added regularization, changing the function whose expectation (both empirically and in the population) is to be minimized from $W\bigl(YF(X)\bigr)$ to $W_n\bigl(YF(X)\bigr)$ where $W_n \to W$ as $n \to \infty$. Bühlmann and Yu (2003) considered L_2 boosting starting from very smooth functions. We shall elaborate on this later.

We consider the behavior of the algorithm as applied to the sample $(Y_1,X_1),\ldots,(Y_n,X_n)$, as well to the "population", that is when means are replaced by expectations and sums by probabilities. The structure of, and the differences between, the population and sample versions of the optimization problem has been explored in various ways by Jiang (2003), Zhang and Yu (2003), Bühlmann (2003), Bartlett, Jordan, and McAuliffe (2003), Bickel and Ritov (2003).

Our goal in this paper is

1. To clarify the issues through a few simple results.

2. To relate our work and that of Bühlmann (2003), Bühlmann and Yu (2003), Lugosi and Vayatis (2004), Zhang (2004), Zhang and Yu (2003) and Bartlett, Jordan, and McAuliffe (2003) to the minimax results of Mammen and Tsybakov (1999), Baraud (2001) and Tsybakov (2001).

In Section 2 we will discuss the population version of the basic boosting algorithms and show how their convergence and that of more general greedy algorithms can be derived from a generalization of Theorem 3 of Mallat and Zhang (1993) with a simple proof. The result can, we believe, also be derived from the even more general theorem of Zhang and Yu (2003), but our method is simpler and the results are transparent.

In Section 3 we show how Bayes consistency of various sample algorithms when suitably stopped or of sample algorithms based on minimization of a regularized W follow readily from population convergence of the algorithms and indicate how test bed validation can be used to do this in a way leading to optimal rates (in Section 4).

In Section 5 we address the issue of bounding the time to perfect separation of the different boosting algorithm (including the standard AdaBoost).

Finally in Section 6 we show how minimax rate results for estimating $E(Y|X)$ may be attained for a "sieve" version of the L_2 boosting algorithm, and relate these to results of Baraud (2001), Lugosi and Vayatis (2004), Bühlmann and Yu (2003), Barron, Birgé, Massart(1999) and Bartlett, Jordan and McAuliffe (2003). We also discuss the relation of these results to classification theory.

2. Boosting "Population" Theorem

We begin with a general theorem on Gauss-Southwell optimization in vector space. It is, in part, a generalization of Theorem 1 of Mallat and Zhang (1993) with a simpler proof. A second part relates to procedures in which the step size is regularized cf. Zhang and Yu (2003) and Bartlett et al. (2003). We make the boosting connection after its statement.

Let w be a real, bounded from below, convex function on a vector space $\mathbb{H}$. Let $\mathcal{H} = \mathcal{H}' \cup (-\mathcal{H}')$, where $\mathcal{H}'$ is a subset of $\mathbb{H}$ whose members are linearly independent, with linear span $\mathcal{F}_\infty = \{\sum_{m=1}^{k} \lambda_m h_m : \lambda_j \in \mathbb{R},\ h_j \in \mathcal{H},\ 1 \le j \le k,\ 1 \le k < \infty\}$. We assume that $\mathcal{F}_\infty$ is dense in $\mathbb{H}$, at least in the sense that $\{w(f) : f \in \mathcal{F}_\infty\}$ is dense in the image of w. We define two relaxed Gauss-Southwell "algorithms".

Algorithm I: For $\alpha \in (0,1]$, and given $f_1 \in \mathbb{H}$, find inductively $f_2, f_3, \ldots, \ldots$ by, $f_{m+1} = f_m + \lambda_m h_m$, $\lambda_m \in \mathbb{R}$, $h_m \in \mathcal{H}$ and

$$w(f_m + \lambda_m h_m) \le \alpha \min_{\lambda \in \mathbb{R},\ h \in \mathcal{H}} w(f_m + \lambda h) + (1-\alpha) w(f_m)\ . \tag{1}$$

Generalize Algorithm I to :

Algorithm II: Like Algorithm I, but replace (1) by

$$w(f_m + \lambda_m h) + \gamma \lambda_m^2 \le \alpha \min_{\lambda \in \mathbb{R},\ h \in \mathcal{H}} (w(f_m + \lambda h) + \gamma \lambda^2) + (1-\alpha) w(f_m)\ .$$

There are not algorithms in the usual sense since they do not specify a unique sequence of iterations but our theorems will apply to any sequence generated in this way. Technically, this scheme

is used in the proof of Theorem 3. The standard boosting algorithms theoretically correspond to $\alpha = 1$, although in practice, since numerical minimization is used, α may equal 1 only approximately. Our generalization makes for a simple proof and covers the possibility that the minimum of $w(f_m + \lambda h)$ over $\mathcal{H}$ and $\mathbb{R}$ is not assumed, or multiply assumed. Let $\omega_0 = \inf_{f \in \mathcal{F}_\infty} w(f) > -\infty$. Let $w'(f;h)$ the linear operator of the Gataux derivative at $f \in \mathcal{F}_\infty$ in the direction $h \in \mathcal{F}_\infty$: $w'(f;h) = \partial w(f+\lambda h)/\partial\lambda|_{\lambda=0}$, and let $w''(f;h)$ be the second derivative of w at f in the direction h: $w''(f,h) \equiv \partial^2 w(f+\lambda h)/\partial\lambda^2|_{\lambda=0}$ (both derivative are assumed to exist). We consider the following conditions.

GS1. For any c_1 and c_2 such that $\omega_0 < c_1 < c_2 < \infty$,

$$0 < \inf\{w''(f,h) : c_1 < w(f) < c_2,\ h \in \mathcal{H}\} \leq \sup\{w''(f,h) : w(f) < c_2,\ h \in \mathcal{H}\} < \infty.$$

GS2. For any $c_2 < \infty$,

$$\sup\{w''(f,h) :\ w(f) < c_2,\ h \in \mathcal{H}\} < \infty\,.$$

Theorem 1 *Under Assumption* GS1*, any sequence of functions generated according to Algorithm I satisfies:*

$$w(f_m) \leq \omega_0 + c_m$$

and if $c_m > 0$*:*

$$w(f_m) - w(f_{m+1}) \geq \xi(w(f_m)) > 0$$

where the sequence $c_m \to 0$ *and the function* $\xi(\cdot)$ *depend only on* α*, the initial points of the iterates, and* $\mathcal{H}$*. The same conclusion holds under Condition* GS2 *for any sequence* f_m *generated according to algorithm II.*

The proof can be found in Appendix A.

Remark:

1. Condition GS2 of Theorem 1 guarantees that $\sum_{m=1}^{\infty} \lambda_m^2 < \infty$. It can be replaced by any other condition that guarantees the same, for example, limiting the step size, replacing the penalty by other penalties, etc.

2. It will be clear from the proof in Appendix A that if w'' is bounded away from 0 and ∞ then c_m is of order $(\log m)^{-\frac{1}{2}}$ so that we, in fact, have an approximation rate – but it is so slow as to be essentially useless. On the other hand, with strong conditions such as orthonormality of the elements of $\mathcal{H}$, and $\mathcal{H}$ a classical approximation class such as trigonometric functions we expect, with L_2 boosting, to obtain rates such as $m^{-1/2}$ or better.

Let $(X,Y) \sim P, X \in \mathcal{X}, Y \in \{-1,1\}$. Let $\mathcal{H} \subset \{h : \mathcal{X} \to [-1,1]\}$ be a symmetric set of functions. In particular, $\mathcal{H}$ can, but need not, be a set of classifiers such as trees with

$$\mathcal{H} = -\mathcal{H}\,. \tag{2}$$

Given a loss function $W : \mathbb{R} \to \mathbb{R}^+$, we consider a greedy sequential procedure for finding a function F that minimizes $EW\big(YF(X)\big)$. That is, given $F_0 \in \mathcal{H}$ fixed, we define for $m \geq 0$:

$$\begin{aligned}
\lambda_m(h) &= \arg\min_{\lambda \in \mathbb{R}} EW\Big(Y\big(F_m(X) + \lambda h(X)\big)\Big) \\
h_m &= \arg\min_{h \in \mathcal{H}} EW\Big(Y\big(F_m(X) + \lambda_m(h) h(X)\big)\Big) \\
F_{m+1} &= F_m + \lambda_m(h_m) h_m.
\end{aligned}$$

Assume, wlog (without loss of generality), by shifting and rescaling, that $W(0) = -W'(0) = 1$. Note that by Bartlett et al. (2003), $W'(0) < 0$ is necessary and sufficient for population consistency defined below. We can suppose again wlog in view of (2), that $\lambda_m \geq 0$. Define $\mathcal{F}_k$ and $\mathcal{F}_\infty$ as in Section 1 and let $\mathcal{F} \equiv \overline{\mathcal{F}}_\infty$ be the closure of $\mathcal{F}_\infty$ in convergence in probability:

$$\begin{aligned}
\mathcal{F} &\equiv \big\{F : \exists F_m \in \mathcal{F}_m,\ F_m(X) \xrightarrow{p} F(X)\big\} \\
F_\infty &\equiv \arg\min_{F \in \mathcal{F}} EW\big(YF(X)\big)
\end{aligned}$$

If $\operatorname{sgn} F_\infty$ is the Bayes rule for 0-1 loss, we say that F_∞ is population consistent for classification, "calibrated" in the Bartlett et al. terminology. Let

$$\begin{aligned}
p(X) &\equiv P(Y = 1|X) \\
\widetilde{W}(x,d) &\equiv p(x)W\big(d\big) + \big(1 - p(x)\big)W\big(-d\big). \\
\widetilde{W}(F) &\equiv \widetilde{W}(X, F(X))
\end{aligned}$$

By the assumptions below F_∞ is the unique function such that $\widetilde{W}'(F_\infty) = 0$ with probability 1, where $\widetilde{W}'(F) = \widetilde{W}'(X, F(X))$ and $\widetilde{W}'(x,d) = \partial W(x,d)/\partial d$. Define $\widetilde{W}''$ similarly.

Here are some conditions.

P1. $P[p(X) = 0 \text{ or } 1] = 0$.

P2. W is twice differentiable and convex on $\mathbb{R}$.

P3. $\mathcal{H}$ is closed and compact in the weak topology. $\mathcal{F}$ is the set of all measurable functions on $\mathcal{X}$.

P4. $\widetilde{W}''(F)$ is bounded above and below on $\{F :\ c_1 < \widetilde{W}(F) < c_2\}$ for all c_1, c_2 such that

$$\inf_{F \in \mathcal{F}} E\widetilde{W}(F) < c_1 < c_2 < E\widetilde{W}(F_0).$$

P5. $F_\infty \in L_2(P)$.

Note that P1 and P2 imply that $\widetilde{W}(x,d) \to \infty$ as $|d| \to \infty$, which ensures that F_∞ is finite almost anywhere. Condition P1, which says that no point can be classified with absolute certainty, is only needed technically to ensure that $\widetilde{W}(x,d) \to \infty$ as $|d| \to \infty$, even if W itself is monotone. It is not needed for L_2 boosting.

Conditions P2 and P4 ensure that along the optimizing path W behaves locally like $W_0(t) = -2t + t^2$ corresponding to L_2 boosting. They are more stringent than we would like and, in particular,

rule out W such as the "hinge" appearing in SVM. More elaborate arguments such as those of Zhang and Yu (2003) and Bartlett et al. (2003) can give somewhat better results.

The functions commonly appearing in boosting such as, $W_1(t) = e^{-t}$, $W_2(t) = -2t + t^2$, $W_3(t) = -\log(1 + e^{-2t})$ satisfy condition P4 if P1 also holds. This is obvious for W_2. For W_1 and W_3, it is clear that P4 holds, if P1 does, since otherwise $E\widetilde{W}\left(YF_m(X)\right) \to \infty$. The conclusions of Theorem 2 continue to hold if $h \in \mathcal{H} \implies |h| \geq \delta > 0$ since then below $w''(F;h) = Eh^2(X)\widetilde{W}\left(F(X)\right) \geq \delta^2 E\widetilde{W}(F(X))$ and P4 follows. Note that if $|h| \not\equiv 1$ the λ optimization step requires multiplying λ^2 by $Eh^2(x)$.

We have,

Theorem 2 *If $\mathcal{H}$ is a set of classifiers, $(h^2 \equiv 1)$ and Assumptions* P2 – P5 *hold, then*

$$F_m(X) \xrightarrow{P} F_\infty(X) \ ,$$

and the misclassification error, $P\left(YF_m(X) \leq 0\right) \to P[YF_\infty(X) \leq 0]$, the Bayes risk.

Proof Identify $w(F) = EW\left(YF(X)\right) = E\widetilde{W}\left(F(X)\right)$. Then,

$$w''(F,h) = Eh^2(X)\widetilde{W}''\left(F(X)\right) = E\widetilde{W}''\left(F(X)\right)$$

and (P4) can be identified with condition GS1 of Theorem 1. Thus,

$$E\widetilde{W}\left(F_m(X)\right) \to E\widetilde{W}(F_\infty(X)) \ .$$

Since,

$$E\widetilde{W}\left(F_m(X)\right) - E\widetilde{W}\left(F_\infty(X)\right) = E\left((F_\infty - F_m)^2 \int_0^1 \widetilde{W}''\left((1-\lambda)(X)F_\infty(X) + \lambda F_m(X)\right)\lambda d\lambda\right) \to 0 \ ,$$

the conclusion of Theorem 2 follows from (P4). The second assertion is immediate. ■

3. Consistency of the Boosting Algorithm

In this section we study the Bayes consistency properties of the sample versions of the boosting algorithms we considered in Section 2. In particular, we shall

(i) Show that under mild additional conditions, there will exist a random sequence $m_n \to \infty$ such that $\hat{F}_{m_n} \xrightarrow{P} F_\infty$, where $\hat{F}_m$ is defined below as the mth sample iterate, and moreover, that such a sequence can be determined using the data.

(ii) Comment on the relationship of this result to optimization for penalized versions of W. The difference is that the penalty forces $m < \infty$ to be optimal while with us, cross-validation (or a test bed sample) determines the stopping point. We shall see that the same dichotomy applies later, when we "boost" using the method of sieves for nonparametric regression studied by Barron, Birge and Massart (1999) and Baraud (2001).

3.1 The Golden Chain Argument

Here is a very general framework. This section is largely based on Bickel and Ritov (2003).

Let $\Theta_1 \subset \Theta_2 \subset \ldots$ be a sequence of sets contained in a separable metric space, $\Theta = \overline{\cup \Theta_m}$ where $\bar{}$ denotes closure. Let $\Pi_m : \Theta_m \to 2^{\Theta_{m+1}}$ be a sequence of point to set mappings. Let K be a target function, and $\vartheta_\infty = \arg\min_{\vartheta \in \Theta} K(\vartheta)$. Finally, let $\hat{K}_n$ be a sample based approximation of K. We assume:

G1. $K : \Theta \to \mathbb{R}$ is strictly convex, with a unique minimizer ϑ_∞.

Our result is applicable to loosely defined algorithms. In particular we want to be able to consider the result of the algorithm applied to the data as if it were generated by a random algorithm applied to the population. We need therefore, the following definitions. Let $\mathcal{S}(\vartheta_0, \alpha)$ be the set of all sequences $\bar{\vartheta}_m \in \Theta_m$, $m = 0, 1, \ldots$ with $\bar{\vartheta}_0 = \vartheta_0$ and satisfying:

$$\bar{\vartheta}_{m+1} \in \Pi_m(\bar{\vartheta}_m)$$
$$K(\bar{\vartheta}_{m+1}) \leq \alpha \inf_{\vartheta \in \Pi_m(\bar{\vartheta}_m)} K(\vartheta) + (1-\alpha) K(\bar{\vartheta}_m).$$

The resemblance to Gauss-Southwell Algorithm I and the boosting procedures is not accidental. Suppose the following uniform convergence criterion is satisifed:

G2. If $\{\bar{\vartheta}_m\} \in \mathcal{S}(\vartheta_0, \alpha)$ with any initial ϑ_0, then $K(\bar{\vartheta}_m) - K(\bar{\vartheta}_{m+1}) \geq \xi\big(K(\bar{\vartheta}_m) - K(\vartheta_\infty)\big)$, for $\xi(\cdot) > 0$ strictly increasing, and $K(\bar{\vartheta}_m) - K(\vartheta_\infty) \leq c_m$ where $c_m \to 0$ uniformly over $\mathcal{S}(\vartheta_0, \alpha)$.

In boosting, given P, $\Theta = \{F(X), F \in \tilde{\mathcal{F}}\}$ with a metric of convergence in probability, $\Theta_m = \{\sum_{j=1}^m \lambda_j h_j, h_j \in \mathcal{H}\}$, $\Pi_m(F) = \Pi(F) = \{F + \lambda h, \lambda \in \mathbb{R}, h \in \mathcal{H}\}$, and $K(F) = \mathrm{E}\, W\big(YF(X)\big)$. Condition G2, follows from the conclusion of Theorem 1.

Now suppose $\hat{K}_n(\cdot)$ is a sequence of random functions on Θ, empirical entities that resemble the population K. Let $\hat{\mathcal{S}}_n(\vartheta_0, \alpha')$ be the set of all sequences $\hat{\vartheta}_{0,n}, \hat{\vartheta}_{1,n} \ldots$, such that $\hat{\vartheta}_{0,n} = \vartheta_0$, and

$$\hat{\vartheta}_{m+1,n} \in \Pi_m(\hat{\vartheta}_{m,n})$$
$$\hat{K}_n(\hat{\vartheta}_{m+1,n}) \leq \alpha' \min\{\hat{K}_n(\vartheta) : \vartheta \in \Pi_m(\hat{\vartheta}_{m,n})\} + (1-\alpha')\hat{K}_n(\hat{\vartheta}_{m,n}).$$

We assume

G3. $\hat{K}_n$ is convex, and for all integer m, $\sup\{|\hat{K}_n(\vartheta) - K(\vartheta)| : \vartheta \in A_m\} \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$, for a sequence $A_m \subset \Theta_m$ such that $P(\hat{\vartheta}_{m,n} \in A_m) \to 1$.

In boosting, $\hat{K}_n(F) = n^{-1} \sum_{i=1}^n W\big(Y_i F(X_i)\big)$, $K(F) = E_p\big(YF(X)\big)$

The sequence $\{\bar{\vartheta}_m\}$ is the golden chain we try to follow using the obscure information in the sample.

We now state and prove,

Theorem 3 *If assumptions* G1– G3 *hold, and* $\alpha' \in (0, 1]$, *then for any sequence* $\{\hat{\vartheta}_{m,n}\} \in \hat{\mathcal{S}}(\vartheta_0, \alpha')$, *there exists a subsequence* $\{\hat{m}_n\}$ *such that* $K(\hat{\vartheta}_{\hat{m}_n,n}) \xrightarrow{\text{p}} K(\vartheta_\infty)$.

Proof

Fix ϑ_0 and α, $\alpha < \alpha'$. Let $M_n \to \infty$ be some sequence, and let $\hat{m}_n = \arg\min_{m \le M_n} K(\hat{\vartheta}_{m,n})$. We need to prove that $K(\hat{\vartheta}_{\hat{m}_n,n}) \xrightarrow{\mathrm{p}} K(\vartheta_\infty)$. We will prove this by contradiction. Suppose otherwise:

$$\inf_{m \le M_n} K(\hat{\vartheta}_{m,n}) - K(\vartheta_\infty) \ge c_1 > 0, \quad n \in \mathcal{N} \tag{3}$$

where $\mathcal{N}$ is unbounded with positive probability. Let $\varepsilon_{m,n} \equiv \sup_{\vartheta \in A_m} |K(\vartheta) - \hat{K}_n(\vartheta)|$. For any fixed m, $\varepsilon_{m,n} \xrightarrow{\text{a.s.}} 0$ by G3. Let

$$m_n = \arg\max\left\{ m' \le M_n : \forall m \le m', \varepsilon_{m-1,n} + 2\varepsilon_{m,n} < (\alpha' - \alpha)\xi(c_1) \ \ \& \ \ \hat{\vartheta}_{m,n} \in A_m \right\}.$$

Clearly, $m_n \xrightarrow{\mathrm{p}} \infty$, and for any $m \le m_n$, assuming (3):

$$\begin{aligned}
K(\hat{\vartheta}_{m,n}) &\le \hat{K}_n(\hat{\vartheta}_{m,n}) + \varepsilon_{m,n} \\
&\le \alpha' \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} \hat{K}_n(\vartheta) + (1-\alpha')\hat{K}_n(\hat{\vartheta}_{m-1,n}) + \varepsilon_{m,n} \\
&\le \alpha' \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha')K(\hat{\vartheta}_{m-1,n}) + \varepsilon_{m-1,n} + 2\varepsilon_{m,n} \\
&= \alpha \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha)K(\hat{\vartheta}_{m-1,n}) \\
&\quad - (\alpha' - \alpha)\big(K(\hat{\vartheta}_{m-1,n}) - \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta)\big) + \varepsilon_{m-1,n} + 2\varepsilon_{m,n} \\
&\le \alpha \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha)K(\hat{\vartheta}_{m-1,n}) \\
&\quad - (\alpha' - \alpha)\xi\Big(K(\hat{\vartheta}_{m,n}) - K(\vartheta_\infty)\Big) + \varepsilon_{m-1,n} + 2\varepsilon_{m,n} \\
&\le \alpha \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha)K(\hat{\vartheta}_{m-1,n}) \\
&\quad - (\alpha' - \alpha)\xi\Big(c_1\Big) + \varepsilon_{m-1,n} + 2\varepsilon_{m,n} \\
&\le \alpha \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha)K(\hat{\vartheta}_{m-1,n}) \text{ for all } m \le m_n \,.
\end{aligned}$$

Thus, there is a sequence $\{\bar{\vartheta}_1^{(n)}, \bar{\vartheta}_2^{(n)}, \ldots\} \in \mathcal{S}(\vartheta_0, \alpha)$, such that $\bar{\vartheta}_m^{(n)} = \hat{\vartheta}_{m,n}$, $m \le m_n$. Hence, by Assumption G2, $K(\vartheta_{m_n,n}) \le K(\vartheta_\infty) + c_{m_n}$, where $\{c_m\}$ is independent of n, and $c_m \to 0$. Therefore, since $m_n \to \infty$, $K(\hat{\vartheta}_{m_n,n}) \to K(\vartheta_\infty)$, contradicting (3). ■

In fact we have proved that sequences m_n can be chosen in the following way involving K.

Corollary 4 *Let M_n be* any *sequence tending to ∞. Let $\tilde{m}_n = \arg\min\{K(\hat{\vartheta}_{m,n}) : 1 \le m \le M_n\}$. Then, under* G1 – G3, $\hat{\vartheta}_{\tilde{m}_n} \xrightarrow{P} \vartheta_\infty$.

To find $\hat{\vartheta}_{\tilde{m}_n,n}$ which are totally determined by the data determining $\hat{K}_n$, we need to add some information about the speed of convergence of $\hat{K}_n$ to K on the "sample" iterates. Specifically, suppose we can determine, in advance, $M_n^* \to \infty$, $\varepsilon_n \to 0$ such that,

$$P[\sup\{|\hat{K}_n(\hat{\vartheta}_{m,n}) - K(\hat{\vartheta}_{m,n})| : 1 \le m \le M_n^*\} \ge \varepsilon_n] \le \varepsilon_n \,.$$

Then $\hat{m}_n = \arg\min\{\hat{K}_n(\hat{\vartheta}_{m,n}) : 1 \le m \le M_n^*\}$ yields an appropriate $\hat{\vartheta}_{\hat{m}_n}$ sequence. We consider this in Section 4. Before that we return to the application of the result of this section to boosting.

3.2 Back to Boosting

We return to boosting, where we consider $\Theta_m = \{\sum_{j=1}^m \lambda_j h_j : \lambda_j \in \mathbb{R}, h_j \in \mathcal{H}\}$, and therefore $\Pi_m \equiv \Pi$, $\Pi(\vartheta) = \{\vartheta + \lambda h, \lambda \in \mathbb{R}, h \in \mathcal{H}\}$. To simplify notation, for any function $a(X,Y)$, let $P_n a(X,Y) = n^{-1}\sum_{i=1}^n a(X_i,Y_i)$ and $Pa(X,Y) = Ea(X,Y)$. Finally, we identify $\hat{\vartheta}_{m,n} = \sum_{j=1}^m \hat{\lambda}_j \hat{h}_j = \sum_{j=1}^m \hat{\lambda}_{j,n}\hat{h}_{j,n}$.

We assume further

GA1. $W(\cdot)$ is of bounded variation on finite intervals.

GA2. $\mathcal{H}$ has finite L_1 bracketing entropy.

GA3. There are finite $a_1, a_2, \ldots$ such that $\sup_n \sum_{j=1}^m |\hat{\lambda}_{j,n}| \le a_m$ with probability 1.

Theorem 5 *Suppose the conclusion of Theorem 1 and Conditions* GA1–GA3 *are satisfied, then conditions* G2, G3 *are satisfied.*

Proof Condition G2 follows from Theorem 1. It remains to prove the uniform convergence in Condition G3. However, GA2 and GA3 imply that $\mathcal{F} \equiv \{F : F = \sum_{j=1}^m \lambda_j h_j, h_j \in \mathcal{H}, |\lambda_j| \le M\}$ has finite L_1 bracketing entropy. Since W can be written as the difference of two monotone functions $\{W(YF) : F \in \mathcal{F}\}$ inherits this property. The result follows from Bickel and Millar (1991), Proposition 2.1. ■

4. Test Bed Stopping

Again we face the issue of data dependent and in some way optimal selection of $\hat{m}_n$. We claim that this can be achieved over a wide range of possible rates of convergence of $EW\big(\hat{F}_{\hat{m}_n}(YX)\big)$ to $EW\big(F_\infty(YX)\big)$ by using a test bed sample to pick the estimator. The following general result plays a key role.

Let $B = B_n \to \infty$, and let $(X,Y),(X_1,Y_1),\ldots,(X_{n+B},Y_{n+B})$ be i.i.d. P, $X \in \mathcal{X}$, $|Y| \le 1$. Let $\hat{\vartheta}_m : \mathcal{X} \to \mathbb{R}$, $1 \le m \le m_n$ be data dependent functions which depend only on $(X_1,Y_1),\ldots,(X_n,Y_n)$ which are predictors of Y. For $g, g_1, g_2 : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$, given P, define

$$\begin{aligned}
\langle g_1, g_2\rangle_* &\equiv \frac{1}{B_n}\sum_{b=1}^{B_n} g_1(X_{b+n},Y_{b+n})g_2(X_{b+n},Y_{b+n}) \\
\langle g_1, g_2\rangle_P &\equiv P\big(g_1(X,Y)g_2(X,Y)\big) = \int g_1(x,y)g_2(x,y)dP(x,y) \\
\|g\|_*^2 &\equiv \langle g_1, g_2\rangle_* \\
\|g\|_P^2 &\equiv \langle g_1, g_2\rangle_P
\end{aligned}$$

Let,

$$\tau = \arg\min\{\|Y - \hat{\vartheta}_m(X)\|_*^2 : 1 \le m \le M_n\}$$

and $\hat{\vartheta}_\tau$ be the selected predictor. Similarly, let

$$o = \arg\min\{\|Y - \hat{\vartheta}_m(X)\|_P^2 : \ 1 \le m \le M_n\}$$

and $\hat{\vartheta}_o$ be the corresponding predictor.

That is, $\hat{\vartheta}_o(X,Y)$ is the predictor an "oracle" knowing P and (X_i,Y_i), $1 \le i \le n$ would pick from $\hat{\vartheta}_1,\ldots,\hat{\vartheta}_{M_n}$ to minimize squared error loss. Let $\vartheta_o(X) \equiv E_P(Y|X)$, the Bayes predictor. Let $\mathcal{P}$ be a set of probabilities and $r_n \equiv \sup\{E_P\|\hat{\vartheta}_o - \vartheta_o\|_P^2 : \ P \in \mathcal{P}\}$.

The following result is due to Györfi et al. (2002) (Theorem 7.1), although there it is stated in the form of an oracle inequality. We need the following condition:

C. $B_n r_n / \log M_n \to \infty$.

Theorem 6 *(Györfi et al.) Suppose condition* **C** *is satisfied, and* $|Y| \le 1$, $\|\hat{\vartheta}_m\|_\infty \le 1$. *Then,*

$$\sup\{\left|E_P(Y - \hat{\vartheta}_\tau)^2 - E_P(Y - \hat{\vartheta}_o)^2\right| : P \in \mathcal{P}\} = o(r_n).$$

Condition **C** very simply asks that the test sample size B_n be large only: (i) In terms of r_n, the minimax rate of convergence; (ii) In terms of the logarithm of the number of procedures being studied. If $|Y| \le 1$, there is no loss in requiring $\|\hat{\vartheta}_m\|_\infty \le 1$, since we could also replace $\hat{\vartheta}_m$ by its truncation at ± 1, minimizing the L_2 cross validated test set risk. Along similar lines, using $\mathrm{sgn}(\hat{\vartheta}_m)$ is equivalent to cross validating the probability of misclassification for these rules, since if $\hat{\vartheta}_m, Y \in \{-1,1\}$, $E(Y - \hat{\vartheta}_m)^2 = 4P(\hat{\vartheta}_m \ne Y)$.

As we shall see in Section 6, typically $r_n = n^{-1+\delta}$, and M_n is at most polynomial in n. If n/B_n is slowly varying, we can check that the conditions hold. Essentially we can only not deal with r_n of order $n^{-1}\log n$.

5. Algorithmic Speed of Convergence

We consider now the time it takes the sample algorithm to convergence. The fact that the algorithm converges follows from Theorem 1. We show in this section that in fact the algorithm perfectly separates the data (*perfect separation* is achieved when $Y_i F_m(x_i) > 0$ for all $i = 1,\ldots,n$) after no more than $c_1 n^2$ steps. Perfect separation is equivalent to empirical misclassification error 0.

The randomness considered in this section comes only from the Y_i, while the design points are considered fixed. We denote them, therefore, by lower case $x_1,\ldots,x_n$. We consider the following assumptions:

O1. W has regular growth in the sense that $W'' < \kappa(W+1)$ for some $\kappa < \infty$. Assume, wlog, that $W(0) = -W'(0) = 1$.

O2. Suppose $x_1,\ldots,x_n$ are all different Then the points can be finitely isolated by $\mathcal{H}$ in the sense that there is k and positive $\alpha_1,\ldots,\alpha_k$ such that for every i there are $h_1,\ldots,h_k \in \mathcal{H}$ such that $\sum_{j=1}^k \alpha_j h_j(x_s) = 1$ if $s = i$, and 0 otherwise. Assume further, as usual, that if $h \in \mathcal{H}$ then $h^2 \equiv 1$ and $-h \in \mathcal{H}$.

Condition O1 is satisfied by all the loss functions mentioned in the introduction. Condition O2 is satisfied, for example by stumps, trees, and any $\mathcal{H}$ whose span includes indicators of small sets with arbitrary location. In particular, if $x_i \in \mathbb{R}$, $x_1 < x_2 < \cdots < x_n$, and $\mathcal{H} = \{\mathrm{sgn}(\cdot - x), x \in \mathbb{R}\}$, we can then take $\alpha_1 = \alpha_2 = 1$, $h_1(\cdot) = \mathrm{sgn}\big(\cdot - (x_{i-1} + x_i)/2\big)$, and $h_2(\cdot) = -\mathrm{sgn}\big(\cdot - (x_i + x_{i+1})/2\big)$

Theorem 7 *Suppose assumptions* O1 *and* O2 *are satisfied and the algorithm starts with* $F_0(0) = 0$. *If* $Y_i F_m(x_i) < 0$ *for at least one i, then*

$$\frac{1}{n}\sum_{i=1}^{n} W\big(Y_i F_m(x_i)\big) - \frac{1}{n}\sum_{i=1}^{n} W\big(Y_i F_{m+1}(x_i)\big) \geq \frac{1}{2\kappa\big(n\sum_{j=1}^{k}\alpha_j\big)^2}.$$

Hence, the boosting algorithm perfectly separates the data after at most $2\kappa(n\sum_{j=1}^{k}|\alpha_j|)^2$ *steps.*

Proof Let, for i such that $Y_i F_m(x_i) < 0$,

$$f_m(\lambda; h) = n^{-1}\sum_{s=1}^{n} W\Big(Y_i\big(F_m(x_s) + \lambda h(x_s)\big)\Big),$$

and $f'_m(0;h) = df_m(\lambda;h)/d\lambda\big|_{\lambda=0}$. Consider $h_1, \ldots, h_k$ as in assumption O2. Replace h_j by $-h_j$ if necessary to ensure that $Y_i\sum_{j=1}^{k}\alpha_j h_j(x_s) = \delta_{si}$. Then

$$\begin{aligned}\sum_{j=1}^{k}\alpha_j f'_m(0;h_j) &= n^{-1}\sum_{j=1}^{k}\alpha_j\sum_{s=1}^{n} W'\big(Y_i F_m(x_s)\big)Y_i h_j(x_s)\\ &= n^{-1}W'\big(Y_i F_m(x_i)\big).\end{aligned}$$

Hence

$$\inf_{h\in\mathcal{H}} f'_m(0;h) \leq \frac{1}{n\sum_{j=1}^{k}\alpha_j}\min_i W'(Y_i F_m(x_i)) \leq \frac{W'(0)}{n\sum_{j=1}^{k}\alpha_j} = \frac{-1}{n\sum_{j=1}^{k}\alpha_j}, \tag{4}$$

since $Y_i F_m(x_i) < 0$ for at least one i.

Let $\bar{h}$ be the minimizer of $f'_m(0,h)$. Note that in particular $f'_m(0;\bar{h}) < 0$. The function $f_m(\cdot;\bar{h})$ is convex, hence it is decreasing in some neighborhood of 0. Denote by $\bar{\lambda}$ its minimizer. Consider the Taylor expansion:

$$\begin{aligned}f_m(\bar{\lambda};\bar{h}) &= f_m(0;\bar{h}) + \bar{\lambda}f'_m(0;\bar{h}) + \frac{\bar{\lambda}^2}{2n}\sum_{s=1}^{n} W''\Big(Y_i\big(F_m(x_s) + \tilde{\lambda}(\lambda)\bar{h}(x_s)\big)\Big)\\ &= f_m(0;\bar{h}) + \inf_{\lambda}\left\{\lambda f'_m(0;\bar{h}) + \frac{\lambda^2}{2n}\sum_{s=1}^{n} W''\Big(Y_i\big(F_m(x_s) + \tilde{\lambda}(\lambda)\bar{h}(x_s)\big)\Big)\right\}\end{aligned}$$

where $\tilde{\lambda}(\lambda)$ lies between 0 and $\bar{\lambda}$. By condition 01,

$$\begin{aligned}&\inf_{\lambda}\left\{\lambda f'_m(0;\bar{h}) + \frac{\lambda^2}{2n}\sum_{s=1}^{n} W''\Big(Y_i\big(F_m(x_s) + \tilde{\lambda}(\lambda)\bar{h}(x_s)\big)\Big)\right\}\\ &\leq \inf_{\lambda}\{\lambda f'_m(0;\bar{h}) + \frac{\lambda^2\kappa}{4n}\sum_{s=1}^{n} W\Big(Y_i\big(F_m(x_s) + \tilde{\lambda}(\lambda)\bar{h}(x_s)\big)\Big) + \frac{\lambda^2\kappa}{4}\right\}\\ &\leq \inf_{\lambda}\{\lambda f'_m(0;\bar{h}) + \frac{\lambda^2\kappa}{2}\}\end{aligned} \tag{5}$$

because $\frac{1}{n}\sum_{s=1}^{n} W(Y_i(F_m(x_s)+\tilde{\lambda}(\lambda)\bar{h}(x_s)) \le \frac{1}{n}\sum_{s=1}^{n} W(Y_i F_m(x_s)) \le W(0) = 1$ since $\bar{\lambda}$ minimizes $f_m(\lambda;\bar{h})$ on $[0,\bar{\lambda}]$, $\tilde{\lambda}$ is an intermediate point, and $F_0 \equiv 0$. Combining (4) and (5) and the minimizing property of $\bar{h}$,

$$f_m(\bar{\lambda};\bar{h}) \le f_m(0;\bar{h}) - \frac{\left(f'_m(0;\bar{h})\right)^2}{2\kappa}$$
$$\le f_m(0;\bar{h}) - \frac{1}{2\kappa(n\sum_{j=1}^{k}\alpha_j)^2} \quad .$$

The second statement of the theorem follows because the initial value of $n^{-1}\sum_{i=1}^{n} W\left(Y_i F_0(x_i)\right)$ is 1, and the value would fall below 0 after at most $m = 2\kappa(n\sum_{j=1}^{k}\alpha_j)^2$ steps in which at least one observation is not classified correctly. Since the value is necessarily positive, we conclude that all observations would be classified correctly before the mth step.

■

6. Achieving Rates with Sieve Boosting

We propose a regularization of L_2 boosting which we view as being in the spirit of the original proposal, but, unlike it, can be shown for, suitable $\mathcal{H}$, to achieve minimax rates for estimation of $E(Y|X)$ under quadratic loss for $\mathcal{P}$ for which $E(Y|X)$ is assumed to belong to a compact set of functions such as a ball in Besov space if $X \in \mathbb{R}$ or to appropriate such subsets of spaces of smooth functions in $X \in \mathbb{R}^d$—see, for example, the classes $\mathcal{F}$ of Györfi et al. (2003). In fact, they are adaptive in the sense of Donoho et al (1995) for scales of such spaces. We note that Bühlmann and Yu (2003) have introduced a version of L_2 boosting which achieves minimax rates for Sobolev classes on $\mathbb{R}$ adaptively already. However, their construction is in a different spirit than that of most boosting papers. They start out with $\mathcal{H}$ consisting of one extremely smooth and complex function and show that boosting reduces bias (roughness of the function) while necessarily increasing variance. Early stopping is still necessary and they show it can achieve minimax rates.

It follows, using a result of Yang (1999) that our rule is adaptive minimax for classification loss for some of the classes we have mentioned as well. Unfortunately, as pointed out by Tsybakov (2001), the sets $\{x: |F_B(x)| \le \varepsilon\}$ can behave very badly as $\varepsilon \downarrow 0$, no matter how smooth F_B, the misclassification Bayes rule, is, so that these results are not as indicative as we would like them to be. In a recent paper, Bartlett, Jordan, and McAuliffe (2003) considered minimization of the W empirical risk $n^{-1}\sum_{i=1}^{n} W(Y_i F(X_i))$, for fairly general convex W, over sets of the form $\mathcal{F} = \{F = \sum_{j=1}^{m}\alpha_j h_j,\ h_j \in \mathcal{H},\ \sum_{j=1}^{m}|\alpha_j| \le \alpha_n,\ (\text{for some representation of } F)\}$. They obtained oracle inequalities relating $EW(Y\hat{F}(X))$ for $\hat{F}_j$ the empirical minimizer over $\mathcal{F}_j$ to the empirical W risk minimum. They then proceeded to show using conditions related to Tsybakov's (A1) above how to relate the misclassification regret of $\hat{f}_j$, given by $\langle P[Y\hat{F}_j(X) < 0] - P[YF_B(X) < 0]\rangle$ to $\langle E_p W(Y\hat{F}_j) - E_p W(YF_B^*)\rangle$, the W regret where F_B^* is the Bayes rule for W. Using these results (Theorems 3 and 10) they were able to establish oracle inequalities for $\hat{F}_j$ under misclassification loss. Manor, Meir, and Zhang (2004) considered the same problem, but focused their analysis mainly on L_2 boosting. They obtained an oracle inequality similar to that of Bartlett et al. regularizing by permitting step sizes which are only a fraction $\beta < 1$ of the step size declared optimal by Gauss-Southwell. They went further by obtaining near minimax results on suitable sets.

We also limit our results to L_2 boosting, although we believe this limitation is primarily due to the lack of minimax theorems for prediction when other losses than L_2 are considered. We use yet a different regularization method in what follows. We show in Theorem 8 our variant of L_2 boosting achieves minimax rates for estimating $E(Y|X)$ in a wide class of situations. Boosting up to a simple data-determined cutoff in each sieve level of a model, and then cross-validating to choose between sieve levels, we can obtain results equivalent to those in which full optimization using penalties are used, such as Theorem 2.1 of Baraud (2000) and results of Baron, Birgé, Massart (1999). Then, in Theorem 9, we show, using inequalities related to ones of Tsybakov (2001), Zhang (2004) and Bartlett et al. (2003), that the rules we propose are also minimax for 0–1 loss in suitable spaces.

6.1 The Rule

Our regularization requires that $\mathcal{H} \equiv \mathcal{H}^{(\infty)} = \overline{\cup_{m\geq 1}\mathcal{H}^{(m)}}$ where $\mathcal{H}^{(m)}$ are finite sets with certain properties. For instance, if $\mathcal{H}$ consists of the stumps in $[0,1]$, $\mathcal{H} = \{F_y(\cdot) : F_y(x) = \mathrm{sgn}(x-y),\ x,y \in [0,1]\}$ we can take $\mathcal{H}^{(m)} = \{F_y(\cdot) : y \text{ a dyadic number of order } k,\ y = \frac{j}{2^k},\ 0 \leq j \leq 2^k\}$. Essentially, we construct a sieve approximating $\mathcal{H}$. Let $\mathcal{F}^{(m)}$ be the linear span of $\mathcal{H}^{(m)}$. Evidently $\mathcal{F} = \overline{\cup_{m\geq 1}\mathcal{F}^{(m)}}$. Let $|\mathcal{H}^{(m)}| \equiv D_m$. Then, $\dim(\mathcal{F}^{(m)}) = D_m$. We now describe our proposed regularization of L_2 boosting.

We use the following notation of Section 4, and begin with a glossary and conditions. Let $(X_1,Y_1),\ldots,(X_n,Y_n),(X,Y)$ i.i.d. with

$$\begin{aligned}
(X,Y) &\sim P << \mu, \quad P \in \mathcal{P}, \quad \mathbf{X} \equiv (X_1,\ldots,X_n),\ \mathbf{Y} \equiv (Y_1,\ldots,Y_n)\,. \\
Y &\in \{-1,1\} \\
\|f\|_\mu^2 &\equiv \int f^2\, d\mu \\
\|f\|_n^2 &\equiv \frac{1}{n}\sum_{i=1}^n f^2(X_i,Y_i) \\
\|f\|_\infty &= \sup_{x,y} |f(x,y)| \\
F_P(X) &\equiv E_P(Y|X) \\
\hat{F}_m(X) &= \mathrm{argmin}\{\|t(X)-Y\|_n^2 :\ t \in \mathcal{F}^{(m)}\} \\
F_m(X) &= \mathrm{argmin}\{\|t(X)-Y\|_P^2 :\ t \in \mathcal{F}^{(m)}\} \\
E_{\mathbf{X}} &\equiv \text{ Conditional expectation given } X_1,\ldots,X_n
\end{aligned}$$

Note that we will often suppress $\mathbf{X},\mathbf{Y}$ in $v(\mathbf{X},\mathbf{Y},X,Y)$ and drop subscript to P.

Let $\hat{F}_{m,k}$, the kth iterate in $\mathcal{F}_m$, be defined as follows

$$\begin{aligned}
\hat{F}_{1,0} &\equiv F_0 \\
\hat{F}_{m+1,0} &= \hat{F}_{m,\hat{k}(m)} \\
\hat{F}_{m,k+1} &= \hat{F}_{m,k} + \hat{\lambda}_{m,k}\hat{h}_{m,km}
\end{aligned}$$

where

$$\begin{aligned}
(\hat{\lambda}_{m,k},\hat{h}_{m,k}) &\equiv \underset{\lambda\in\mathbb{R},h\in\mathcal{H}^{(m)}}{\mathrm{argmin}}\ \{-2\lambda P_n(Y-\hat{F}_{m,k})h + \lambda^2 P_n(h^2)\} \\
\hat{k}(m) &= \text{ First } k \text{ such that } \hat{\lambda}_{m,k}^2 \leq \Delta_{m,n},
\end{aligned}$$

where $\Delta_{m,n}$ are constants. Let

$$\tilde{F}_m = H(\hat{F}_{m,\hat{k}(m)})$$

where

$$H(x) = \begin{cases} x & \text{if } |x| \le 1 \\ \text{sgn}(x) & \text{if } |x| > 1 \end{cases} \tag{6}$$

Note that we have suppressed dependence on n here, indicating it only by the "hats". Let,

$$\hat{m} = \arg\min\{\|Y - \tilde{F}_m(x)\|_* : \ m \le M_n\}$$

where

$$\|f\|_*^2 = \frac{1}{B}\sum_{i=n+1}^{n+B} f^2(X_i, Y_i), \text{ and we take } B = B_n = \frac{n}{\log n} .$$

The rule we propose is: $\hat{\delta} = \text{sgn}(\hat{\tilde{F}})$, where

$$\hat{\tilde{F}} \equiv H(F_{\hat{m},\hat{k}(\hat{m})}) . \tag{7}$$

Note: We show at the end of the Appendix (Proof of Lemma 10) that for wavelet $\mathcal{H}$ we take at most $Cn\log n$ steps total in this algorithm.

6.2 Conditions and Results

We use C as a generic constant throughout, possibly changing from line to line but not depending on m, n, or P. Lemma 6.3 and the condition we give are essentially due to Baraud (2001). Let μ be a sigma finite measure on $\mathcal{H}$ and $\|f\|_\mu$ be the $L_2(\mu)$ norm.

R1. If $\mathcal{H}^{(m)} = \{h_{m,1}, \ldots, h_{m,D_m}\}$ and $f_{m,j} \equiv h_{m,j}/\|h_{m,j}\|_\mu$, then $\{f_{m,j}\}, j \ge 1$ is an orthonormal basis of $\mathcal{F}^{(m)}$ in $L_2(\mu)$ such that:

(i) $\|f_{m,j}\|_\infty \le C_\infty D_m^{\frac{1}{2}}$ for all j, where $\|f\|_\infty = \sup_x |f(x)|$.

(ii) There exists an L such that for all m, j, j',
$f_{m,j} f_{m,j'} = 0$ if $|j - j'| \ge L$.

R2. There exists $\varepsilon = \varepsilon(P) > 0$ such that, $\varepsilon \le \frac{dP}{d\mu} \le \varepsilon^{-1}$ for all $P \in \mathcal{P}$.

R3. $\sup_{P\in\mathcal{P}} \|F_P - F_m\|_P^2 \le CD_m^{-\beta}$ for all m, $\beta > 1$.

R4. $M_n \le D_{M_n} \le \frac{n}{(\log n))^p}$ for some $p > 1$.

Condition R1 is needed to conclude that we can bound the behavior of the L_∞ norm on $\mathcal{F}^{(m)}$ by that of the L_2 norm for μ. Condition R2 simply ensures that we can do so for $P \in \mathcal{P}$ as well. The members $f_{m,j}$ of the basis of $\mathcal{F}^{(m)}$ must have compact support. It is well known that if $\mathcal{H}_m$ consists of scaled wavelets (in any dimension) then R1 holds. Clearly, if say μ is Lebesgue measure on an hypercube then to satisfy R2 $\mathcal{P}$ can consist only of densities bounded from above and away from 0. Condition R3 gives the minimum approximation error incurred by using an estimate F based

on $\mathcal{F}^{(m)}$, and thus limits our choice of $\mathcal{H}$. Finally, R4 links the oracle error for these sequences of procedures to the number of candidate procedures.

Let

$$r_n(P) = \inf\{E_P\|\hat{F}_m - F_P\|_P^2 : \ 1 \le m \le M_n\}, \qquad r_n \equiv \sup_{P\in\mathcal{P}} r_n(P).$$

Thus, r_n is the minimax regret for an oracle knowing P but restricted to $\hat{F}_m$. We use the notation $a_n \asymp b_n$ for a shortcut for $a_n = O(b_n)$ and $b_n = O(a_n)$, We have

Theorem 8 *Suppose that $\mathcal{P}$ and $\mathcal{F}$ satisfy* R1–R4 *and that $\mathcal{H}$ is a VC class. If $\Delta_{m,n} = O(D_m/n)$, then,*

$$\sup_{\mathcal{P}} E_P\|\hat{\tilde{F}}(X) - F_P(X)\|_P^2 \asymp r_n \ . \tag{8}$$

Thus, $\hat{\tilde{F}}$ given by (7) is rate minimax.

Theorem 9 *Suppose the assumptions of Theorem 8 hold and $\mathcal{P}_0 = \mathcal{P} \cap \{P : \ P\big(\ |F_P(X)| \le t\big) \le ct^\alpha\}$, $\alpha \ge 0$. Let $\Delta_n(F,P)$ be the Bayes classification regret for P,*

$$\Delta_n(F,P) \equiv P\big(YF(X) < 0\big) - P\big(YF_P(X) < 0\big) \ . \tag{9}$$

Then,

$$\sup_{\mathcal{P}_0} \Delta_n(\hat{\tilde{F}}, P) \asymp r_n^{\frac{\alpha+1}{\alpha+2}} \ . \tag{10}$$

The condition $P[|F_P(x)| \le t] \le ct^\alpha$, some $\alpha \ge 0$, t sufficiently small appears in Proposition 1 of Tsybakov (2001) as sufficient for his condition (A1) which is studied by both Bartlett et al. (2003) and Mammen and Tsybakov (1999).

The proof of Theorem 9 uses 2 lemmas of interest which we now state. Their proofs are in the Appendix.

We study the algorithm on $\mathcal{F}_m$. For any positive definite matrix Σ define the condition number $\gamma(\Sigma) \equiv \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$, where $\lambda_{\max}$, $\lambda_{\min}$ are the largest and smallest eigenvalues of Σ. Let $G_m(P) = \|E_P f_{m,i}(X) f_{m,j}(X)\|$ be the $D_m \times D_m$ Gram matrix of the basis $\{f_{m,1}, \ldots, f_{m,D_m}\}$.

Lemma 10 *Under* R1 *and* R2,

a) $\gamma(G_m(P)) \le \varepsilon^{-2}$, where ε is as in R2.

b) Let $G_m(P_n)$ be the empirical Gram matrix $\hat{\gamma}_m \equiv \gamma(G_m(P_n))$. Then, if in addition to R1 *and* R2, *$\mathcal{H}$ is a VC class, $P[\gamma(\hat{G}_m) \ge C_1] \le C_2 \exp\{-C_3 n/L^2 D_m\}$ for all $m \le M_n$ for such that $D_m \le n/(\log n)^p$ for $p > 1$.*

c) If $\mathcal{H}$ is a VC class, $P[\|\hat{F}_{m,\hat{k}(m)} - \hat{F}_m\|_k \le C\frac{D_m}{n}] = 1 - O(\frac{1}{n})$ The C and 0 terms are determed solely by the constants appearing in the R conditions.

Lemma 11 *Suppose* R1, R2, *and* R4 *hold. Then,*

$$E_P(\tilde{F}_m - F_P)^2 \le C\{E_P(F_m - F_P)^2 + \frac{D_m}{n} + E_P(\tilde{F}_m - \hat{F}_m)^2\}.$$

This "oracle inequality" is key for what follows.

Proof of Theorem 9

$$P\big(YF(x)<0\big)=\frac{1}{2}E_P\Big(1\big(F(X)>0\big)\big(1-F_P(X)\big)\Big)+\frac{1}{2}E_P\Big(1\big(F(X)<0\big)\big(1+F_P(X)\big)\Big).$$

Hence for all $\varepsilon>0$,

$$\begin{aligned}\Delta_n(F,P)&=E_P\Big(1\big(F(X)<0,F_P(X)>0\Big)F_P(X)-1\big(F(X)>0,F_P(X)<0\Big)F_P(X)\Big)\\&=E_P\Big(|F_P(X)|1(F_P(X)F(X)<0)\Big)\\&\le E_P\Big(\big|F(X)-F_P(X)\big|1(F_PF(X)<0,|F_P(X)|>\varepsilon)\Big)+\varepsilon P\big(|F_P(X)|\le\varepsilon\big)\\&\le\frac{1}{\varepsilon}E_P\big(F(X)-F_P(X)\big)^2+c\varepsilon^{\alpha+1}\end{aligned}$$

by assumption. The theorem follows. ■

6.3 Discussion

1) If $X\in\mathbb{R}$ and $\mathcal{H}^{(m)}$ consists of stumps with the discontinuity at a dyadic rational $j/2^m$, then $\mathcal{F}^{(m)}$ is the linear space of Haar wavelets of order m. This is also true if $\mathcal{H}_m$ is the space of differences of two such dyadic stumps. More generally, if $\mathcal{H}$ consists of suitably scaled wavelets, so that $|h|\le 1$, based on the dyadic rationals of order m, then $\mathcal{F}^{(m)}$ is the linear space spanned by the first 2^m elements of the wavelet series. A slight extension of results of Baraud (2001) yields that if we run the algorithm to the limit $k=\infty$ for each m rather than stopping as we indicate, the resulting $\hat{F}_m$ obey the oracle inequality of Lemma 11 with $\Delta_{m,n}=0$.

   Suppose that $X\in\mathbb{R}$ and F_∞ ranges over a ball in an approximation space such as Sobolev or, more generally, Besov. Then, if $\mathcal{F}^{(m)}$ has the appropriate approximation properties, e.g., wavelets as smooth as the functions in the specified space, it follows from Baraud (2001) that we can use penalties not dependent on the data to pick $\hat{F}_{\hat{m}}$ such that,

$$\begin{aligned}\max_{\hat{F}}E_P\Big(\hat{F}_{\hat{m}}(X)-E_P(Y|X)\Big)^2&\asymp\min_{\hat{F}}\max\Big\{E_P\big(\hat{F}(X)-E_P(Y|X)\big)^2:E_P(Y|X)\in\mathcal{F}\Big\}\\&\asymp n^{-1+\varepsilon}\Omega(n)\end{aligned}$$

   where $\Omega(n)$ is slowly varying and $0<\varepsilon<1$. Here $\hat{F}$ ranges over all estimators based only on the data and not on P. The same type of result has been established for more specialized models with $X\in\mathbb{R}^d$ by Baron, Birgé, Massart (1999), and others, see Györfi et al. (2003).

   The resulting minimax risk,

$$\min_{\hat{F}}\max\{E_P\big(\hat{F}(X)-E_P(Y|X)\big)^2:\ E_P(Y|X)\in\mathcal{F}\}$$

   is always of order $n^{-1+\varepsilon}\Omega(n)$ where $\Omega(n)$ is typically constant and $0<\varepsilon<1$.

   What we show in Theorem 8 is that if, rather than optimizing all the way for each m, we stop in a natural fashion and cross validate as we have indicated, then we can achieve the optimal order as well.

2) "Stumps" unfortunately do not satisfy condition R1 with μ Lebesgue measure. Their Gram matrices are too close to being singular. But differences of stumps work.

3) It follows from the results of Yang (1999) that the rate of Theorem 9 for $\alpha = 0$, that is, if $\mathcal{P}_0 = \mathcal{P}$, is best possible for Sobolev balls and the other spaces we have mentioned.

   Tsybakov implicitly defines a class of F_P for which he is able to specify classification minimax rates. Specifically let $X \in [0,1]^d$ and let $b(x_1,\ldots,x_{d-1})$ be a function having continuous partial derivatives up to order ℓ. Let $p_{b,x}(\cdot)$ be the Taylor polynomial or order ℓ obtained from expanding b at x. Then, he defines $\Sigma(l,L)$ to be the class of all such b for which, $|b(y) - p_{b,x}(y)| \le L|y-x|^\ell$ for all $x,y \in [0,1]^{d-1}$. Evidently if b has bounded partial derivatives of order $\ell+1$, $b \in \Sigma(\ell,L)$, for some L. Now let

$$\begin{aligned}\mathcal{P}_\ell = \quad & \{P : F_P(x) = x_d - b(x_1,\ldots,x_{d-1}),\\ & P[|F_P(x)| \le t] \le Ct, \text{ for all } 0 \le t \le 1, b \in \Sigma(\ell,L)\}\end{aligned}$$

   Tsybakov following Mammen and Tsybakov (1999) shows that the classification minimax regret for $\mathcal{P}$ (Theorem 2 of Tsybakov (2001) for $K = 2$) is $\frac{2\ell}{3\ell+(d-1)}$. On the other hand, if we assume that $Y = F_P(x) + \varepsilon$ where ε is independent of X, bounded and $E(\varepsilon) = 0$, then the L_2 minimax regret rate is $2\ell/(2\ell+(d-1))$ – see Birgé and Massart (1999) Sections 4.1.1 and Theorem 9. Our theorem 9 now yields a classification minimax regret rate of

$$\frac{2}{3} \cdot \frac{2\ell}{2\ell+(d-1)} = \frac{2\ell}{3\ell+\frac{3}{2}(d-1)}$$

   which is slightly worse than what can be achieved using Tsybakov's not as readily computable procedures. However, note that as $\ell \to \infty$ so that F_P and the boundary become arbitrarily smooth, L_2 boosting approaches the best possible rate for $\mathcal{P}_\ell$ of $\frac{2}{3}$. Similar remarks can be made about $0 < \alpha \le 1$.

7. Conclusions

In this paper we presented different mathematical aspects of boosting. We consider the observations as an i.i.d. sample from a population (i.e., a distribution). The boosting algorithm is a Gauss-Southwell minimization of a classification loss function (which typically dominates the 0-1 misclassification loss). We show that the output of the boosting algorithm follows the theoretical path as if it were applied to the true distribution of the population. Since early stopping is possible as argued, the algorithm, supplied with an appropriate stopping rule, is consistent.

However, there are no simple rate results other than those of Bühlmann and Yu (2003), which we discuss, for the convergence of the boosting classifier to the Bayes classifier. We showed that rate results can be obtained when the boosting algorithm is modified to a cautious version, in which at each step the boosting is done only over a small set of permitted directions.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF grant DMS-0104075), and the Israel Science Foundation (ISF grant 793/03).

Appendix A. Proof of Theorem 1:

Let $w_0 = \inf_{f\in\mathcal{F}_\infty} w(f)$. Let $f_k^* = \sum_m \alpha_{km} h_{km}$, $h_{k,m} \in \mathcal{H}$, $\sum_m |\alpha_{km}| < \infty$, $k = 0,1,2,\ldots$ be any member of $\mathcal{F}_\infty$ such that (i) $f_0^* = f_0$; (ii) $w(f_k^*) \searrow w_0$ is strictly decreasing sequence; (iii) The following condition is satisfied:

$$w(f_k^*) \geq \alpha w_0 + (1-\alpha) w(f_{k-1}^*) + (1-\alpha)(\nu_{k-1} - \nu_k), \tag{11}$$

where $\nu_k \searrow 0$ is a strictly decreasing real sequence. The construction of the sequence $\{f_k^*\}$ is possible since, by assumption, $\mathcal{F}_\infty$ is dense in the image of $w(\cdot)$. That is, we can start with the sequence $\{w(f_k^*)\}$, and then look for suitable $\{f_k^*\}$. Here is a possible construction. Let c and η be suitable small number. Let $\gamma = (1-\alpha)(1+2\eta)/(1-\eta)$, $\nu_k = c\eta\gamma^k/(1-\gamma)$. Select now f_k^* such $w_0 + c(1-\eta)\gamma^k \leq w(f_k^*) \leq w_0 + c(1+\eta)\gamma^k$. ($\eta$ should be small enough such that $\gamma < 1$ and c should selected such that $w(f_1^*) < w(f_0)$.) Our argument rests on the following,

Lemma 12 *There is a sequence $m_k \to \infty$ such that $w(f_m) \leq w(f_k^*) + \nu_k$ for $m \geq m_k$, $k = 1,2,\ldots$, and $m_k \leq \zeta_k(m_{k-1}) < \infty$, where $\zeta_k(\cdot)$ is a monotone non-decreasing functions which depends only on the sequences $\{\nu_k\}$ and $\{f_k^*\}$.*

Proof of Lemma 12:

We will use the following notation. For $f \in \mathcal{F}_\infty$ let $\|f\|_* = \inf\{\sum|\gamma_i|,\ f = \sum\gamma_i h_i,\ h_i \in \mathcal{H}\}$.

Recall that by definition $w(f_0) = w(f_0^*)$. Our argument proceeds as follows, We will inductively define m_k satisfying the conclusion of the lemma, and make, if $\varepsilon_{k,m} \equiv w(f_m) - w(f_k^*)$,

$$\varepsilon_{k,m} \leq c_{k,m} \equiv \max\Big\{\nu_k,\ \frac{\sqrt{512}B}{\alpha^2\beta_k}\frac{w(f_{k-1}^*) - w_0}{\Big(\log\Big(1 + \frac{8(w(f_{k-1}^*)-w_0)}{\alpha\beta_k(\tau_k+\rho_k m_{k-1})}(m - m_{k-1} + 1)\Big)\Big)^{1/2}}\Big\}, \tag{12}$$

where

$$\begin{aligned} \beta_k = &\ \inf\{w''(f;h) : w_0 + \nu_k \leq w(f) \leq w(f_0),\ h \in \mathcal{H}\} \\ B = &\ \sup\{w''(f;h) : w(f) \leq w(f_o),\ h \in \mathcal{H}\} < \infty. \end{aligned} \tag{13}$$

and

$$\begin{aligned} \tau_k &= 2\|f_0 - f_k^*\|_*^2 \\ \rho_k &= \frac{16}{\alpha\beta_k}\big(w(f_0) - w_0\big). \end{aligned} \tag{14}$$

Having defined m_k we establish (12) as part of our induction hypothesis for $m_{k-1} < m \leq m_k$. We begin by choosing $m = m_1 = 1$ so that (12) holds for $m = M - 1 = 1$. We do do this by choosing $\nu_0 > 0$, sufficiently small. Having established the induction for $m \leq m_{k-1}$ we define m_k as follows. Write now the RHS of (12) as $g(m_{k-1})$, where

$$g(\nu) \equiv \max\Big\{\nu_k,\ \frac{\sqrt{512}B}{\alpha^2\beta_k}\frac{w(f_{k-1}^*) - w_0}{\Big(\log\Big(1 + \frac{8(w(f_{k-1}^*)-w_0)}{\alpha\beta_k(\tau_k+\rho_k\nu)}(m - \nu + 1)\Big)\Big)^{1/2}}\Big\},$$

We can now pick $\zeta_k(\nu) \equiv \max\{\nu+1, \min\{m:\ g(\nu) \le \nu_k\}\}$, and define $m_k = \zeta_k(\nu_{k-1})$.

Note that $\{\beta_k\}$, $\{\tau_k\}$, $\{\rho_k\}$, and B depend only the sequences $\{f_k^*\}$ and $\{\nu_k\}$. We now proceed to establish (12). for $m_{k-1} < m \le m_k$. Note first that since $\varepsilon_{k,m}$ as a function of m is non-increasing, (12) holds trivially for $m' > m$ if $\varepsilon_{k,m} \le 0$. By induction (12) holds for $m \le m_{k-1}$, and my hold for some $m > mk-1$. Recall that the definition of the algorithm relates the actual gain at the mth to the maximal gain achieved in this step given the previous steps, see its definition (1). Suppose

$$\inf_{\lambda} w(f_m + \lambda h_m) \le w_0 + \nu_k. \tag{15}$$

Then

$$\begin{aligned}
w(f_{m+1}) &\le \alpha \inf_{\lambda} w(f_m + \lambda h_m) + (1-\alpha) w(f_m), \quad \text{by (1)} \\
&\le \alpha(w_0 + \nu_k) + (1-\alpha) w(f_m), \quad \text{by (15)} \\
&\le \alpha(w_0 + \nu_k) + (1-\alpha)\big(w(f_{k-1}^*) + \nu_{k-1}\big), \quad \text{by the outer induction, since } m \ge m_{k-1} \\
&\le \alpha(w_0 + \nu_k) + \big(w(f_k^*) - \alpha w_0 + (1-\alpha)\nu_k\big), \quad \text{by (11)} \\
&= w(f_k^*) + \nu_k,
\end{aligned}$$

so that $\varepsilon_{k,m+1} \le \nu_k$. Therefore, m_k' is not larger than $m+1$, that is $\varepsilon_{k,m'} \le \nu_k$ for $m' > m$ then (12) holds trivially for $m' > m$, and hence, by the second induction assumption for all m. We have established (12) save for m such that,

$$\inf_{\lambda} w(f_m + \lambda h_m) > w_0 + \nu_k \text{ and } \varepsilon_{k,m} \ge 0. \tag{16}$$

We now deal with this case.

Note first that by convexity,

$$|w'(f_m; f_m - f_k^*)| \ge w(f_m) - w(f_k^*) \equiv \varepsilon_{k,m}. \tag{17}$$

We obtain from (17) and the linearity of the derivative that, if $f_m - f_k^* = \sum \gamma_i \widetilde{h}_i \in \mathcal{F}_\infty$,

$$\varepsilon_{k,m} \le \left| \sum -\gamma_i w'(f_m; \tilde{h}_i) \right| \le \sup_{h \in \mathcal{H}} |w'(f_m; h)| \sum |\gamma_i| .$$

Hence

$$\sup_{h \in \mathcal{H}} |w'(f_m; h)| \ge \frac{\varepsilon_{k,m}}{\|f_m - f_k^*\|_*}. \tag{18}$$

Now, if $f_{m+1} = f_m + \lambda_m h_m$ then,

$$w(f_m + \lambda_m h_m) = w(f_m) + \lambda_m w'(f_m; h_m) + \frac{1}{2}\lambda_m^2 w''(\tilde{f}_m; h_m), \quad \lambda \in [0, \lambda_m]. \tag{19}$$

where $\tilde{f}_m = f_m + \tilde{\lambda}_m h_m$ and $0 \le \widetilde{\lambda}_m \le \lambda_m$. By convexity, for $0 \le \lambda \le \lambda_m$,

$$w(f_m + \lambda h_m) = w(f_m(1 - \frac{\lambda}{\lambda_m}) + \frac{\lambda}{\lambda_m} f_{m+1}) \le \max\{w(f_m), w(f_{m+1})\} = w(f_m) \le w(f_1).$$

We obtain from Assumption GS1 that $w''(\tilde{f}_m;h) \in (\beta_k, B)$ given in (13). But then we conclude from (19) that,

$$\begin{aligned} w(f_m+\lambda_m h_m) &\geq w(f_m) + \inf_{\lambda\in\mathbb{R}}\left(\lambda w'(f_m;h_m) + \frac{1}{2}\lambda^2\beta_k\right) \\ &= w(f_m) - \frac{|w'(f_m;h_m)|^2}{2\beta_k} \quad . \end{aligned} \tag{20}$$

Note that $w(f_m+\lambda h) = w(f_m)+\lambda w'(f_m,h)+\lambda^2 w''(f_m+\lambda' h,h)/2$ for some $\lambda' \in [0,\lambda]$, and if $w(f_m+\lambda h)$ is close to $\inf_{\lambda,h} w(f_m+\lambda,h)$ then by convexity, $w(f_m+\lambda' h) \leq w(f_m) \leq w(f_0)$. We obtain from the upper bound on w'' we obtain:

$$\begin{aligned} w(f_m+\lambda_m h_m) &\leq \alpha \inf_{\lambda\in\mathbb{R},\, h\in\mathcal{H}} w(f_m+\lambda h) + (1-\alpha)w(f_m), \qquad \text{by definition,} \\ &\leq \alpha \inf_{\lambda\in\mathbb{R},\, h\in\mathcal{H}} (w(f_m)+\lambda w'(f_m;h) + \frac{1}{2}\lambda^2 B) + (1-\alpha)w(f_m) \\ &= w(f_m) - \frac{\alpha \sup_{h\in\mathcal{H}} |w'(f_m;h)|^2}{2B}, \end{aligned} \tag{21}$$

by minimizing over λ. Hence combining (20) and (21) we obtain,

$$|w'(f_m;h_m)| \geq \alpha \sup_{h\in\mathcal{H}} |w'(f_m;h)| \sqrt{\frac{\beta_k}{B}} \tag{22}$$

By (21) for the LHS and convexity for the RHS:

$$\frac{\alpha \sup_{h\in\mathcal{H}} |w'(f_m;h)|^2}{2B} \leq w(f_m) - w(f_{m+1}) \leq -\lambda_m w'(f_m;h_m)$$

Hence

$$|\lambda_m| \geq \frac{\alpha \sup_{h\in\mathcal{H}} |w'(f_m;h)|}{2B}.$$

Applying (18) we obtain:

$$|\lambda_m| \geq \frac{\alpha}{2B}\frac{\varepsilon_{k,m}}{l_{k,m}}, \tag{23}$$

where $l_{k,m} \equiv \|f_m - f_k^*\|_*$.

Let λ_m^0 be the minimal point of $w(f_m+\lambda h_m)$. Taylor expansion around that point and using the lower bound on the curvature:

$$w(f_m+\lambda h_m) \geq w(f_m+\lambda_m^0 h_m) + \frac{1}{2}\beta_k(\lambda-\lambda_m^0)^2 \tag{24}$$

Hence

$$\begin{aligned} {\lambda_m^0}^2 &\leq \frac{2}{\beta_k}\left(w(f_m) - w(f_m+\lambda_m^0 h_m)\right) \\ &\leq \frac{2}{\alpha\beta_k}\left(w(f_m) - w(f_{m+1})\right), \end{aligned} \tag{25}$$

where the RHS follows (1). Similarly

$$\begin{aligned}(\lambda_m - \lambda_m^0)^2 &\le \frac{2}{\beta_k}\big(w(f_{m+1}) - w(f_m + \lambda_m^0 h_m)\big) \\ &\le \frac{2(1-\alpha)}{\alpha\beta_k}\big(w(f_m) - w(f_{m+1})\big)\end{aligned} \tag{26}$$

Combining (25) and (26):

$$\lambda_m^2 \le \frac{8}{\alpha\beta_k}\big(w(f_m) - w(f_{m+1})\big). \tag{27}$$

Since $\varepsilon_{k,m} \ge 0$ by assumption (16), we conclude from (27) that,

$$\sum_{i=m_{k-1}}^{m} \lambda_i^2 \le \frac{8}{\alpha\beta_k}(w(f_{k-1}^*) - w_0). \tag{28}$$

However, by definition,

$$\begin{aligned}l_{k,m+1} &\le l_{k,m} + |\lambda_m| \\ &\le l_k + \sum_{i=m_{k-1}}^{m} |\lambda_i| \\ &\le l_k + (m+1-m_{k-1})^{1/2}\Big(\sum_{i=m_{k-1}}^{m} \lambda_i^2\Big)^{1/2}\end{aligned} \tag{29}$$

by Cauchy-Schwarz, where, similarly,

$$\begin{aligned}l_k = l_{k,m_{k-1}} &= \|f_{m_{k-1}} - f_k^*\|_* \\ &\le \|f_0 - f_k^*\|_* + \|f_{m_{k-1}} - f_0\|_* \\ &\le \|f_0 - f_k^*\|_* + \sum_{m=0}^{m_{k-1}-1} |\lambda_m| \\ &\le \|f_0 - f_k^*\|_* + m_{k-1}^{1/2}\sqrt{\sum_{m=0}^{m_{k-1}-1} \lambda_m^2} \\ &\le \|f_0 - f_k^*\|_* + \sqrt{\frac{8m_{k-1}}{\alpha\beta_k}}\sqrt{w(f_0) - w(f_{m_{k-1}})}, \quad \text{by (27)} \\ &\le \|f_0 - f_k^*\|_* + \sqrt{\frac{8m_{k-1}}{\alpha\beta_k}}\sqrt{w(f_0) - w_0} \\ &\le \sqrt{\tau_k + \rho_k m_{k-1}}, \quad \text{as defined in (14).}\end{aligned} \tag{30}$$

Together, (23), (28), and (29) yield:

$$\begin{aligned}\frac{8}{\alpha\beta_k}(w(f^*_{k-1})-w_0) &\geq \sum_{i=m_{k-1}}^{m}\lambda_i^2 \\ &\geq \frac{\alpha^2}{4B^2}\sum_{i=m_{k-1}}^{m}\frac{\varepsilon_{k,i}^2}{l_{k,i}^2} \\ &\geq \frac{\alpha^2}{4B^2}\sum_{i=m_{k-1}}^{m}\frac{\varepsilon_{k,i}^2}{(l_k+(8(w(f^*_{k-1})-w_0)/\alpha\beta_k)^{1/2}(i-m_{k-1})^{1/2})^2}\end{aligned} \tag{31}$$

Further, since $\varepsilon_{k,m}$ are decreasing by construction and positive by assumption (16), we can simplify the sum on the RHS of (31):

$$\begin{aligned}&\sum_{i=m_{k-1}}^{m}\frac{\varepsilon_{k,i}^2}{(l_k+(8(w(f^*_{k-1})-w_0)/\alpha\beta_k)^{1/2}(i-m_{k-1})^{1/2})^2} \\ &\qquad\geq \frac{\varepsilon_{k,m}^2}{2}\sum_{i=0}^{m-m_{k-1}}\frac{1}{l_k^2+8i(w(f^*_{k-1})-w_0)/\alpha\beta_k}.\end{aligned} \tag{32}$$

Using the inequality,

$$\sum_{i=0}^{m-m_{k-1}}\frac{1}{a+bi}\geq\int_0^{m-m_{k-1}+1}\frac{1}{a+bt}\,dt=\frac{1}{b}\log\Big(1+\frac{b}{a}(m-m_{k-1}+1)\Big)$$

on the RHS of (32), we obtain from (31) and (32) that (12) holds, for the case (16). This establishes (16) for all k and m.

■

Proof of Theorem 1: Since the lemma established the existence of monotone ζ_k's, it followed from the definition of these function that $w(f_m)\leq w(f^*_{k(m)})$ where $k(m)=\sup\{k:\zeta^{(k)}(f_0^*)\leq m\}$ and $\zeta^{(k)}=\zeta_k\circ\cdots\circ\zeta_1$ is the kth iterate of the ζs. Since $\zeta^{(k)}(f_0^*)<\infty$ for all k, we have established the uniform rate of convergence and can define the sequence $\{c_m\}$, where $c_m=w(f^*_{k(m)})-w_0$.

We now prove the uniform step improvement claim of the theorem and identify a suitable function $\xi(\cdot)$. From (26) and (23) if $\varepsilon_{k,m}\geq 0$

$$w(f_m)-w(f_{m+1})\geq\frac{\alpha\beta_k}{2}\lambda_m^2\geq\frac{\alpha\beta_k}{2}\left(\frac{\alpha}{2B}\frac{\varepsilon_{k,m}}{l_{k,m}}\right)^2, \tag{33}$$

Bound $l_{k,m}$ similarly to (30) by

$$l_{k,m}\leq l_{k,1}+m^{1/2}\Big(\sum_{i=1}^{m}\lambda_i^2\Big)^2\leq l_{k,1}+\sqrt{\frac{8m}{\alpha\beta_k}(w(f_0)-w_0)}. \tag{34}$$

Let $m^*(v)=\inf\{m':c_{m'}\leq v-w_0\}$, which is well defined since $c_m\to 0$. Thus, any realization of the algorithm will cross the v line on or before step number $m^*(v)$. In particular, $m\leq m^*\big(w(f_m)\big)$ for

any m and any realization of the algorithm. We obtain therefore by plugging-in (34) in (33), using the m^* as a bound on m and the identity $(a+b)^2 \leq 2a^2+2b^2$ that:

$$w(f_m) - w(f_{m+1}) \geq \frac{\alpha^3 \beta_k}{16B^2} \frac{w(f_m) - w(f_k^*)}{l_{k,1}^2 + 8m^*\big(w(f_m)\big)\big(w(f_0) - w_o\big)/\alpha\beta_k},$$

as long as $\varepsilon_{k,m} \geq 0$. Taking the maximum of the RHS over the permitted range, yields a candidate for the ξ function:

$$\xi(w) \equiv \sup_{k:\, w(f_k^*) \leq w} \Big\{ \frac{\alpha^3 \beta_k}{16B^2} \frac{w - w(f_k^*)}{l_{k,1}^2 + m^*\big(w\big)\big(w(f_0) - w_o\big)/\alpha\beta_k} \Big\}.$$

This proves the theorem under GS1. Under GS2, the only inequality which we need to replace is (20) since now $\beta_k = 0$ is possible. However the definition of Algorithm 2 ensures that we have a coefficient of at least γ on λ^2 in (20). The theorem is proved.

■

Appendix B. Proof of Lemmas 10 and 11 and Theorem 8

Proof of Lemma 10 Since by (R2)

$$\begin{aligned} \lambda_{\max}(G_m(P)) &= \sup_{\|x\|=1} x' G_m(P) x \\ &= \sup_{\|x\|=1} \sum\sum x_i x_j \int f_{m,i} f_{m,d} dP \\ &= \sup_{\|x\|=1} \int (\sum x_i f_{m,i})^2 dP \qquad (35) \\ &\leq \varepsilon^{-1} \sup_{\|x\|=1} \int (\sum x_i f_{m,i})^2 d\mu = \varepsilon^{-1} \end{aligned}$$

$\lambda_{\max}(G_m(P)) \geq \varepsilon$, similarly.

Part a) follows.

For any symmetric matrix M define its operator norm $\|\cdot\|_T$ by $\lambda_{\max}(M)$. For simplicity let $G_m = G_m(P)$ and $\hat{G}_m = G_m(P_n)$. Recall that for any symmetric matrices A and and M:

$$\begin{aligned} |\lambda_{\max}(A) - \lambda_{\max}(M)| &\leq \|A - M\|_T \\ |\lambda_{\min}(A) - \lambda_{\min}(M)| &\leq \|A - M\|_T. \end{aligned}$$

Now,

$$\begin{aligned} &P\left[\left| \frac{\lambda_{\max}\big(\hat{G}_m\big)}{\lambda_{\min}\big(\hat{G}_m\big)} - \frac{\lambda_{\max}\big(G_m\big)}{\lambda_{\min}\big(G_m\big)} \right| \geq t \right) \qquad (36) \\ &\quad \leq P\Big(\|\hat{G}_m - G_m\|_T > \frac{\varepsilon}{2} \Big) + P\Big(\|\hat{G}_m - G_m\|_T \geq t/(\frac{1}{\varepsilon} + \frac{2}{\varepsilon^3}) \Big) \end{aligned}$$

Recall that for a banded matrix M of with band of width $2L$,

$$\begin{aligned}\|M\|_T^2 &= \sup_{\|x\|=1}\|Mx\|^2 \\ &= \sup_{\|x\|=1}\sum_a\Big(\sum_b M_{ab}x_b\Big)^2 \\ &\le \sup_{\|x\|=1}\sum_a\sum_{|b-a|<L} x_b^2 M_\infty^2 \\ &\le 2LM_\infty^2 \sup_{\|x\|=1}\sum_a x_a^2 = 2LM_\infty^2,\end{aligned}$$

where $\|M\|_\infty \equiv \max_{a,b}|M_{ab}|$. Since both $\hat{G}_m$ and $G_m(P)$ are banded of width d, say,

$$\|\hat{G}_m - G_m\|_T \le 2L\max\Big\{\Big|\frac{1}{n}\sum_{i=1}^n (f_{m,a}f_{m,b})(X_i) - E_P f_{m,a}f_{m,b}(X_i))\Big| : |a-b|<L\Big\}. \tag{37}$$

If $\mathcal{H}$ is a VC class, we can conclude from (35)–(37) that,

$$P[\gamma(\hat{G}_m) \ge C_1] \le C_2\exp\{-C_3 n/L^2 D_m\} \tag{38}$$

since by R1 (i), $\|f_m\|_\infty \le C_\infty D_m^{\frac{1}{2}}$. The constants ε, C_1, C_2 and C_3 depend on the constants of the R conditions only. This is a consequence of Theorem 2.14.16 p. 246 of van der Vaart and Wellner (1996). This complete the proof of part b).

By a standard result for the Gauss-Southwell method, Luenberger (1984), page 229:

$$\|\hat{F}_{m,k+1} - \hat{F}_m\|_n^2 \le \Big(1 - \frac{1}{\hat{\gamma}_m D_m}\Big)\|\hat{F}_{m,k} - \hat{F}_m\|_n^2 \tag{39}$$

Hence

$$\|\hat{F}_{m,k} - \hat{F}_m\|_n^2 - \|\hat{F}_{m,k+1} - \hat{F}_m\|_n^2 \ge \frac{1}{\hat{\gamma}_m D_m}\|\hat{F}_{m,k} - \hat{F}_m\|_n^2$$

Thus, if

$$\frac{1}{n} \ge \|\hat{F}_{m,k} - \hat{F}_m\|_n^2 - \|\hat{F}_{m,k+1} - \hat{F}_m\|_n^2$$

we obtain

$$\|\hat{F}_{m,k} - \hat{F}_m\|_n^2 \le D_m\hat{\gamma}_m/n. \tag{40}$$

¿From (40) part (c) follows. ■

Note: Since

$$\|\hat{F}_{m,k-1} - \hat{F}_m\|_n^2 - \|\hat{F}_{m,k} - \hat{F}_m\|_n^2 \;\ge\; \frac{C}{n}$$

(39) implies that

$$\Big(1 - \frac{1}{\hat{\gamma}_m D_m}\Big)^{\hat{k}(m)} \;\ge\; \frac{1}{n}.$$

Therefore:

$$\hat{k}(m) \le \log n\,\hat{\gamma}_m D_m\,.$$

If, for instance, as with wavelets $D_m = 2^m, m \le \log_2 n$ we take at most $Cn\log n$ steps total.

Lemma 13 *:*
If $E_{\mathbf{x}}$ denotes conditional expectation give n $X_1,\ldots,X_n$, under R1 *and $F \equiv F_p$,*

$$E_{\mathbf{x}}\|\hat{F}_m - F_m\|_n^2 \leq C(\frac{D_m}{n} + \|F_m - F\|_p^2) \tag{41}$$

This is a standard type of result – see Barron, Birgé, Massart (1999). We include the proof for completeness. Note that,

$$\|\hat{F}_m(X) - Y\|_n^2 = \frac{1}{n}\mathbf{Y}^T(I-P)\mathbf{Y}$$

where $\mathbf{Y} \equiv (Y_1,\ldots,Y_n)^T$ and P is the projection matrix of dimension D_m onto the L space spanned by $\big(h_j(X_1),\ldots,h_j(X_n)\big)$, $1 \leq j \leq D_m$. Then, $(I-P)v = 0$ for all $v \in L$. Hence,

$$E_{\mathbf{X}}\|\hat{F}_m(X) - Y\|_n^2 = \frac{1}{n}E_{\mathbf{X}}\big(\mathbf{Y} - \mathbf{F}_m(\mathbf{X})\big)^T(I-P)\big(\mathbf{Y} - \mathbf{F}_m(\mathbf{X})\big)$$

where $\mathbf{F}_m(\mathbf{X}) = \big(F_m(X_1),\ldots,F_m(X_n)\big)^T$ is the projection of $(F(X_1),\ldots,F(X_n))^T$ onto L. Note also that,

$$\|\hat{F}_m - F_m\|_n^2 = \|\mathbf{Y} - \mathbf{F}_m(\mathbf{X})\|_n^2 - \|\mathbf{Y} - \hat{\mathbf{F}}_m(\mathbf{X})\|_n^2$$

where $\hat{\mathbf{F}}(X) = (\hat{F}_m(X_1),\ldots,\hat{F}_m(X_n))^T$. Hence,

$$\begin{aligned}
E_{\mathbf{X}}\|\hat{F}_m - F_m\|_n^2 &= \tfrac{1}{n}E_{\mathbf{X}}(\mathbf{Y} - \mathbf{F}_m(X))^T P(\mathbf{Y} - \mathbf{F}_m(\mathbf{X})) \\
&= \tfrac{1}{n}E_{\mathbf{X}}(\mathbf{Y} - \mathbf{F}(\mathbf{X}))^T P(\mathbf{Y} - \mathbf{F}(\mathbf{X})) + \tfrac{2}{n}E_{\mathbf{X}}(\mathbf{F}_m - \mathbf{F})^T P(\mathbf{Y} - \mathbf{F}_m(\mathbf{X})) \\
&= \tfrac{1}{n}E_{\mathbf{X}}\,\mathrm{trace}[P(\mathbf{Y} - \mathbf{F}(\mathbf{X}))(\mathbf{Y} - \mathbf{F}(\mathbf{X}))] \\
&\quad + \tfrac{2}{n}E_{\mathbf{X}}(\mathbf{F}_m - \mathbf{F})^T P(\mathbf{F}_m - \mathbf{F})(\mathbf{X})
\end{aligned}$$

But

$$E_{\mathbf{X}}\,\mathrm{trace}[P(\mathbf{Y} - \mathbf{F}(\mathbf{X})))(\mathbf{Y} - \mathbf{F}(\mathbf{X}))^T] = \frac{1}{n}\sum_{i=1}^{n} Var(Y_i|X_i)p_{ii}(X) \leq \max_i Var(Y_i|X_i)\frac{D_m}{n}$$

since

$$\sum_{i=1}^{n} p_{ii}(X) = \mathrm{trace}\,P = D_m$$

Also, since P is a projection matrix

$$(\mathbf{F}_m - \mathbf{F})^T P(\mathbf{F}_m - \mathbf{F})(\mathbf{X}) \leq \|F_m - F\|_n^2$$

and (41) *follows.*

Proof of Lemma 11:
Take $\Delta_{m,n} = 0$. Let $\tilde{\rho}_m = \sup\left\{\frac{\|t(X)\|_p}{\|t(X)\|_n} \;:\; t \in \mathcal{T}_m\right\}$. By Proposition 5.2 of Baraud (2001), if $\rho_0 > h_0^{-1}$,

$$P[\tilde{\rho}_m > \rho_0] \leq D_m^2 \exp\{-\frac{(h_0 - \rho_0^{-1})^2}{4h_1}\, c_n \log n\}$$

where $c_n = \frac{n}{CD_m \log n}$. Here $h_0, h_1 C$ are generic constants. Baraud gives a proof for the case $Var(Y|X) =$ constant, but this is immaterial since only functions of $\underset{\sim}{X}$ are involved in $\tilde{\rho}_m$. Therefore,

$$\begin{aligned} & E_P(\hat{F}_m - F_P)^2 1(\rho_m \le \rho_0) \\ \le \; & 2\rho_0^2 E_P\{E_n(\hat{F}_m - F_m)^2 + E_n(F_m - F_P)^2\} \\ \le \; & C(\frac{D_m}{n} + \|F_m - F_P\|^2) \end{aligned} \tag{42}$$

On the other hand,

$$\begin{aligned} E_P(\hat{F}_m - F_P)^2 1(\rho_m > \rho_0) &\le 2P[\rho_m > \rho_0] \\ &= CD_m^2 \exp\{-AC_n \log n\} \end{aligned} \tag{43}$$

Combining (42) and (43) we obtain Lemma 11 for $\Delta_{m,n} = 0$, $\hat{F}_m = \widetilde{F}_m$. Putting in $\widetilde{F}_m$ we add a term $CE_P(\hat{F}_m - \widetilde{F}_m)^2$. We now apply Lemma 10 c) and the argument we used to obtain (42) and (43). ■

Proof of Theorem 8: Note that we are limited to rates of convergence which are slower than $n^{-\frac{1}{2}}$. This comes from the combination of R1(i) and bounding the operator by the l_∞ norm of the Gram matrix. It is not clear how either of these conditions can be relaxed.

We need only check that if the $\{\widetilde{F}_m\}$ are the θ_m of Theorem 6 then the conditions of that theorem are satisfied. By construction, $\|\widetilde{F}_m\|_\infty \le 1$, $B_n = \frac{n}{\log n}$. By Lemma 11 and (R3),

$$r_n \le C_1 \frac{D_m}{n} + C_2 D_m^{-\beta} \tag{44}$$

and the right hand side of (44) is bounded by $n^{-(\frac{\beta}{\beta+1})}$. ■

References

P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: A large sample study. *Ann. Stat.* 10:1100–1120, 1982.

Y. Baraud. Model selection for regression on a random design.*Tech. Report*, U. Paris Sud, 2001.

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection under penalization. *Prob. Theory and Related Fields*, 113:301–413, 1999.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Tech. Report* 638, Department of Statistics, University of California at Berkeley, 2003.

P. J. Bickel and P. W. Millar. Uniform convergence of probability measures on classes of functions. Statistica Sinica 2:1-15, 1992.

P. J. Bickel and Y. Ritov. The golden chain. A comment. *Ann. Statist.*, 32:91–96, 2003.

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

L. Breiman. Prediction games and arcing algorithms. *Neural Computation* 11:1493-1517, 1999.

L. Breiman. Some infinity theory for predictor ensembles *Technical Report* U.C. Berkeley, 2000.

P. Bühlmann. Consistency for L_2 boosting and matching pursuit with trees and tree type base functions. *Technical Report* ETH Zürich, 2002.

P. Bühlmann and B. Yu. Boosting the L_2 loss: regression and classification. *J. of Amer. Statist. Assoc.*, 98:324–339, 2003

D. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia (with discussion). *J. Roy. Statist. Soc. Ser. B* 57:371–394, 1995.

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

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

Y, Freund and R. E. Schapire. Experiments with a new boosting algorithm. *Machine Learning: Proc. 13th International Conference*, 148–156. Morgan Kauffman, San Francisco, 1996.

G. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution Free Theory of Nonparametric Regression*. Springer, New York, 2002.

W. Jiang. Process consistency for ADABOOST. Technical Report 00-05, Dept. of Statistics, Northwestern University, 2002.

Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data . *J. of Amer. Statist. Assoc.*, 99:67–81, 2002.

D. G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley Publishing Company, Reading, 1984.

G. Lugosi and N. Vayatis. On the Bayes-risk consistency of boosting methods. *Ann. Statist.* 32:30–55, 2004.

S. Mallat and Z. Zhang. Matching pursuit with time frequency dictionaries. *IEEE Transactions on Signal Processing* 41:3397–3415, 1993.

E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Ann. Statist.* 27:1808–1829, 1999.

S. Mannor, R. Meir, and T. Zhang. Greedy algorithms for classification—consistency, convergence rates and adaptivity. J. of Machine Learning Research 4:713–742, 2004.

L. Mason, P. Bartlett, J. Baxter, and M. Frean. Functional gradient techniques for combining hypotheses. In Schölkopg, Smola, A., Bartlett, P., and Schurmans, D. (edts.) *Advances in Large Margin Classifiers*, MIT Press, Boston, 2000.

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

R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence related predictions. *Machine Learning*, 37:297–336, 1999.

A, Tsybakov. Optimal aggregation of classifiers in statistical learning. *Technical Report*, U. of Paris IV, 2001.

A. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

Y. Yang. Minimax nonparametric classification – Part I Rates of convergence, Part II Model selection, *IEEE Trans. Inf. Theory* 45:2271–2292, 1999.

T. Zhang and B. Yu. Boosting with early stopping: convergence and consistency. Tech Report 635, Stat Dept, UCB, 2003.

T. Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32:56–134, 2004.

The Annals of Statistics
2009, Vol. 37, No. 4, 1705–1732
DOI: 10.1214/08-AOS620

SIMULTANEOUS ANALYSIS OF LASSO AND DANTZIG SELECTOR[1]

BY PETER J. BICKEL, YA'ACOV RITOV AND ALEXANDRE B. TSYBAKOV

University of California at Berkeley, The Hebrew University and Université Paris VI and CREST

We show that, under a sparsity scenario, the Lasso estimator and the Dantzig selector exhibit similar behavior. For both methods, we derive, in parallel, oracle inequalities for the prediction risk in the general nonparametric regression model, as well as bounds on the ℓ_p estimation loss for $1 \le p \le 2$ in the linear model when the number of variables can be much larger than the sample size.

1. Introduction. During the last few years, a great deal of attention has been focused on the ℓ_1 penalized least squares (Lasso) estimator of parameters in high-dimensional linear regression when the number of variables can be much larger than the sample size [8, 9, 11, 17, 18, 20–22, 26] and [27]. Quite recently, Candes and Tao [7] have proposed a new estimate for such linear models, the Dantzig selector, for which they establish optimal ℓ_2 rate properties under a sparsity scenario; that is, when the number of nonzero components of the true vector of parameters is small.

Lasso estimators have also been studied in the nonparametric regression setup [2–4, 12, 13, 19] and [5]. In particular, Bunea, Tsybakov and Wegkamp [2–5] obtain sparsity oracle inequalities for the prediction loss in this context and point out the implications for minimax estimation in classical nonparametric regression settings, as well as for the problem of aggregation of estimators. An analog of Lasso for density estimation with similar properties (SPADES) is proposed in [6]. Modified versions of Lasso estimators (nonquadratic terms and/or penalties slightly different from ℓ_1) for nonparametric regression with random design are suggested and studied under prediction loss in [14] and [25]. Sparsity oracle inequalities for the Dantzig selector with random design are obtained in [15]. In linear fixed design regression, Meinshausen and Yu [18] establish a bound on the ℓ_2 loss for the coefficients of Lasso that is quite different from the bound on the same loss for the Dantzig selector proven in [7].

The main message of this paper is that, under a sparsity scenario, the Lasso and the Dantzig selector exhibit similar behavior, both for linear regression and

Received August 2007; revised April 2008.

[1]Supported in part by NSF Grant DMS-06-05236, ISF grant, France-Berkeley Fund, the Grant ANR-06-BLAN-0194 and the European Network of Excellence PASCAL.

AMS 2000 subject classifications. Primary 60K35, 62G08; secondary 62C20, 62G05, 62G20.

Key words and phrases. Linear models, model selection, nonparametric statistics.

for nonparametric regression models, for ℓ_2 prediction loss and for ℓ_p loss in the coefficients for $1 \le p \le 2$. All the results of the paper are nonasymptotic.

Let us specialize to the case of linear regression with many covariates, $\mathbf{y} = X\beta + \mathbf{w}$, where X is the $n \times M$ deterministic design matrix, with M possibly much larger than n, and $\mathbf{w}$ is a vector of i.i.d. standard normal random variables. This is the situation considered most recently by Candes and Tao [7] and Meinshausen and Yu [18]. Here, sparsity specifies that the high-dimensional vector β has coefficients that are mostly 0.

We develop general tools to study these two estimators in parallel. For the fixed design Gaussian regression model, we recover, as particular cases, sparsity oracle inequalities for the Lasso, as in Bunea, Tsybakov and Wegkamp [4], and ℓ_2 bounds for the coefficients of Dantzig selector, as in Candes and Tao [7]. This is obtained as a consequence of our more general results, which are the following:

- In the nonparametric regression model, we prove sparsity oracle inequalities for the Dantzig selector; that is, bounds on the prediction loss in terms of the best possible (oracle) approximation under the sparsity constraint.
- Similar sparsity oracle inequalities are proved for the Lasso in the nonparametric regression model, and this is done under more general assumptions on the design matrix than in [4].
- We prove that, for nonparametric regression, the Lasso and the Dantzig selector are approximately equivalent in terms of the prediction loss.
- We develop geometrical assumptions that are considerably weaker than those of Candes and Tao [7] for the Dantzig selector and Bunea, Tsybakov and Wegkamp [4] for the Lasso. In the context of linear regression where the number of variables is possibly much larger than the sample size, these assumptions imply the result of [7] for the ℓ_2 loss and generalize it to ℓ_p loss $1 \le p \le 2$ and to prediction loss. Our bounds for the Lasso differ from those for Dantzig selector only in numerical constants.

We begin, in the next section, by defining the Lasso and Dantzig procedures and the notation. In Section 3, we present our key geometric assumptions. Some sufficient conditions for these assumptions are given in Section 4, where they are also compared to those of [7] and [18], as well as to ones appearing in [4] and [5]. We note a weakness of our assumptions, and, hence, of those in the papers we cited, and we discuss a way of slightly remedying them. Sections 5 and 6 give some equivalence results and sparsity oracle inequalities for the Lasso and Dantzig estimators in the general nonparametric regression model. Section 7 focuses on the linear regression model and includes a final discussion. Two important technical lemmas are given in Appendix B as well as most of the proofs.

2. Definitions and notation. Let $(Z_1, Y_1), \dots, (Z_n, Y_n)$ be a sample of independent random pairs with

$$Y_i = f(Z_i) + W_i, \qquad i = 1, \dots, n,$$

where $f : \mathcal{Z} \to \mathbb{R}$ is an unknown regression function to be estimated, $\mathcal{Z}$ is a Borel subset of $\mathbb{R}^d$, the Z_i's are fixed elements in $\mathcal{Z}$ and the regression errors W_i are Gaussian. Let $\mathcal{F}_M = \{f_1, \ldots, f_M\}$ be a finite dictionary of functions $f_j : \mathcal{Z} \to \mathbb{R}$, $j = 1, \ldots, M$. We assume throughout that $M \geq 2$.

Depending on the statistical targets, the dictionary $\mathcal{F}_M$ can contain qualitatively different parts. For instance, it can be a collection of basis functions used to approximate f in the nonparametric regression model (e.g., wavelets, splines with fixed knots, step functions). Another example is related to the aggregation problem, where the f_j are estimators arising from M different methods. They can also correspond to M different values of the tuning parameter of the same method. Without much loss of generality, these estimators f_j are treated as fixed functions. The results are viewed as being conditioned on the sample that the f_j are based on.

The selection of the dictionary can be very important to make the estimation of f possible. We assume implicitly that f can be well approximated by a member of the span of $\mathcal{F}_M$. However, this is not enough. In this paper, we have in mind the situation where $M \gg n$, and f can be estimated reasonably only because it can approximated by a linear combination of a small number of members of $\mathcal{F}_M$, or, in other words, it has a sparse approximation in the span of $\mathcal{F}_M$. But, when sparsity is an issue, equivalent bases can have different properties. A function that has a sparse representation in one basis may not have it in another, even if both of them span the same linear space.

Consider the matrix $X = (f_j(Z_i))_{i,j}$, $i = 1, \ldots, n$, $j = 1, \ldots, M$ and the vectors $\mathbf{y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\mathbf{f} = (f(Z_1), \ldots, f(Z_n))^{\mathrm{T}}$, $\mathbf{w} = (W_1, \ldots, W_n)^{\mathrm{T}}$. With the notation

$$\mathbf{y} = \mathbf{f} + \mathbf{w},$$

we will write $|x|_p$ for the ℓ_p norm of $x \in \mathbb{R}^M$, $1 \leq p \leq \infty$. The notation $\|\cdot\|_n$ stands for the empirical norm

$$\|g\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} g^2(Z_i)}$$

for any $g : \mathcal{Z} \to \mathbb{R}$. We suppose that $\|f_j\|_n \neq 0$, $j = 1, \ldots, M$. Set

$$f_{\max} = \max_{1 \leq j \leq M} \|f_j\|_n, \qquad f_{\min} = \min_{1 \leq j \leq M} \|f_j\|_n.$$

For any $\beta = (\beta_1, \ldots, \beta_M) \in \mathbb{R}^M$, define $f_\beta = \sum_{j=1}^M \beta_j f_j$ or, explicitly, $f_\beta(z) = \sum_{j=1}^M \beta_j f_j(z)$ and $\mathbf{f}_\beta = X\beta$. The estimates we consider are all of the form $f_{\tilde{\beta}}(\cdot)$, where $\tilde{\beta}$ is data determined. Since we consider mainly sparse vectors $\tilde{\beta}$, it will be convenient to define the following. Let

$$\mathcal{M}(\beta) = \sum_{j=1}^{M} I_{\{\beta_j \neq 0\}} = |J(\beta)|$$

denote the number of nonzero coordinates of β, where $I_{\{\cdot\}}$ denotes the indicator function $J(\beta) = \{j \in \{1, \dots, M\} : \beta_j \neq 0\}$ and $|J|$ denotes the cardinality of J. The value $\mathcal{M}(\beta)$ characterizes the *sparsity* of the vector β. The smaller $\mathcal{M}(\beta)$, the "sparser" β. For a vector $\boldsymbol{\delta} \in \mathbb{R}^M$ and a subset $J \subset \{1, \dots, M\}$, we denote by $\boldsymbol{\delta}_J$ the vector in $\mathbb{R}^M$ that has the same coordinates as $\boldsymbol{\delta}$ on J and zero coordinates on the complement J^c of J.

Introduce the residual sum of squares

$$\widehat{S}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - f_\beta(Z_i)\}^2$$

for all $\beta \in \mathbb{R}^M$. Define the Lasso solution $\widehat{\beta}_L = (\widehat{\beta}_{1,L}, \dots, \widehat{\beta}_{M,L})$ by

$$\widehat{\beta}_L = \arg\min_{\beta \in \mathbb{R}^M} \left\{ \widehat{S}(\beta) + 2r \sum_{j=1}^{M} \|f_j\|_n |\beta_j| \right\}, \tag{2.1}$$

where $r > 0$ is some tuning constant, and introduce the corresponding Lasso estimator

$$\widehat{f}_L(x) = f_{\widehat{\beta}_L}(x) = \sum_{j=1}^{M} \widehat{\beta}_{j,L} f_j(z). \tag{2.2}$$

The criterion in (2.1) is convex in β, so that standard convex optimization procedures can be used to compute $\widehat{\beta}_L$. We refer to [9, 10, 20, 21, 24] and [16] for detailed discussion of these optimization problems and fast algorithms.

A necessary and sufficient condition of the minimizer in (2.1) is that 0 belongs to the subdifferential of the convex function $\beta \mapsto n^{-1}|y - X\beta|_2^2 + 2r|D^{1/2}\beta|_1$. This implies that the Lasso selector $\widehat{\beta}_L$ satisfies the constraint

$$\left| \frac{1}{n} D^{-1/2} X^{\mathrm{T}} (y - X\widehat{\beta}_L) \right|_\infty \leq r, \tag{2.3}$$

where D is the diagonal matrix

$$D = \operatorname{diag}\{\|f_1\|_n^2, \dots, \|f_M\|_n^2\}.$$

More generally, we will say that $\beta \in \mathbb{R}^M$ satisfies the Dantzig constraint if β belongs to the set

$$\left\{ \beta \in \mathbb{R}^M : \left| \frac{1}{n} D^{-1/2} X^{\mathrm{T}} (y - X\beta) \right|_\infty \leq r \right\}.$$

The Dantzig estimator of the regression function f is based on a particular solution of (2.3), the Dantzig selector $\widehat{\beta}_D$, which is defined as a vector having the smallest ℓ_1 norm among all β satisfying the Dantzig constraint

$$\widehat{\beta}_D = \arg\min \left\{ |\beta|_1 : \left| \frac{1}{n} D^{-1/2} X^{\mathrm{T}} (y - X\beta) \right|_\infty \leq r \right\}. \tag{2.4}$$

The Dantzig estimator is defined by

$$\widehat{f}_D(z) = f_{\widehat{\beta}_D}(z) = \sum_{j=1}^{M} \widehat{\beta}_{j,D} f_j(z), \tag{2.5}$$

where $\widehat{\beta}_D = (\widehat{\beta}_{1,D}, \ldots, \widehat{\beta}_{M,D})$ is the Dantzig selector. By the definition of Dantzig selector, we have $|\widehat{\beta}_D|_1 \le |\widehat{\beta}_L|_1$.

The Dantzig selector is computationally feasible, since it reduces to a linear programming problem [7].

Finally, for any $n \ge 1$, $M \ge 2$, we consider the Gram matrix

$$\Psi_n = \frac{1}{n} X^{\mathrm{T}} X = \left(\frac{1}{n} \sum_{i=1}^{n} f_j(Z_i) f_{j'}(Z_i) \right)_{1 \le j, j' \le M},$$

and let $\phi_{\max}$ denote the maximal eigenvalue of Ψ_n.

3. Restricted eigenvalue assumptions. We now introduce the key assumptions on the Gram matrix that are needed to guarantee nice statistical properties of the Lasso and Dantzig selectors. Under the sparsity scenario, we are typically interested in the case where $M > n$, and even $M \gg n$. Then, the matrix Ψ_n is degenerate, which can be written as

$$\min_{\delta \in \mathbb{R}^M : \delta \ne 0} \frac{(\delta^{\mathrm{T}} \Psi_n \delta)^{1/2}}{|\delta|_2} \equiv \min_{\delta \in \mathbb{R}^M : \delta \ne 0} \frac{|X\delta|_2}{\sqrt{n}|\delta|_2} = 0.$$

Clearly, ordinary least squares does not work in this case, since it requires positive definiteness of Ψ_n; that is,

$$\min_{\delta \in \mathbb{R}^M : \delta \ne 0} \frac{|X\delta|_2}{\sqrt{n}|\delta|_2} > 0. \tag{3.1}$$

It turns out that the Lasso and Dantzig selector require much weaker assumptions. The minimum in (3.1) can be replaced by the minimum over a restricted set of vectors, and the norm $|\delta|_2$ in the denominator of the condition can be replaced by the ℓ_2 norm of only a part of δ.

One of the properties of both the Lasso and the Dantzig selectors is that, for the linear regression model, the residuals $\delta = \hat{\beta}_L - \beta$ and $\delta = \hat{\beta}_D - \beta$ satisfy, with probability close to 1,

$$|\delta_{J_0^c}|_1 \le c_0 |\delta_{J_0}|_1, \tag{3.2}$$

where $J_0 = J(\beta)$ is the set of nonzero coefficients of the true parameter β of the model. For the linear regression model, the vector of Dantzig residuals δ satisfies (3.2) with probability close 1 if $c_0 = 1$ and M is large [cf. (B.9) and the fact that β of the model satisfies the Dantzig constraint with probability close to 1 if M is

large]. A similar inequality holds for the vector of Lasso residuals $\boldsymbol{\delta} = \widehat{\beta}_L - \beta$, but this time with $c_0 = 3$ [cf. Corollary B.2].

Now, for example, consider the case where the elements of the Gram matrix Ψ_n are close to those of a positive definite $(M \times M)$-matrix Ψ. Denote, by $\varepsilon_n \stackrel{\triangle}{=} \max_{i,j} |(\Psi_n - \Psi)_{i,j}|$, the maximal difference between the elements of the two matrices. Then, for any $\boldsymbol{\delta}$ satisfying (3.2), we get

$$
\begin{aligned}
\frac{\boldsymbol{\delta}^{\mathrm{T}} \Psi_n \boldsymbol{\delta}}{|\boldsymbol{\delta}|_2^2} &= \frac{\boldsymbol{\delta}^{\mathrm{T}} \Psi \boldsymbol{\delta} + \boldsymbol{\delta}^{\mathrm{T}} (\Psi_n - \Psi) \boldsymbol{\delta}}{|\boldsymbol{\delta}|_2^2} \\
&\geq \frac{\boldsymbol{\delta}^{\mathrm{T}} \Psi \boldsymbol{\delta}}{|\boldsymbol{\delta}|_2^2} - \frac{\varepsilon_n |\boldsymbol{\delta}|_1^2}{|\boldsymbol{\delta}|_2^2} \\
&\geq \frac{\boldsymbol{\delta}^{\mathrm{T}} \Psi \boldsymbol{\delta}}{|\boldsymbol{\delta}|_2^2} - \varepsilon_n \biggl(\frac{(1 + c_0) |\boldsymbol{\delta}_{J_0}|_1}{|\boldsymbol{\delta}_{J_0}|_2} \biggr)^2 \\
&\geq \frac{\boldsymbol{\delta}^{\mathrm{T}} \Psi \boldsymbol{\delta}}{|\boldsymbol{\delta}|_2^2} - \varepsilon_n (1 + c_0)^2 |J_0|.
\end{aligned} \tag{3.3}
$$

Thus, for $\boldsymbol{\delta}$ satisfying (3.2), which are the vectors that we have in mind, and for $\varepsilon_n |J_0|$ small enough, the LHS of (3.3) is bounded away from 0. This means that we have a kind of "restricted" positive definiteness, which is valid only for the vectors satisfying (3.2). This suggests the following conditions, which will suffice for the main argument of the paper. We refer to these conditions as *restricted eigenvalue* (RE) assumptions.

ASSUMPTION RE(s, c_0). For some integer s such that $1 \leq s \leq M$ and a positive number c_0, the following condition holds:

$$
\kappa(s, c_0) \stackrel{\triangle}{=} \min_{\substack{J_0 \subseteq \{1, \ldots, M\}, \\ |J_0| \leq s}} \min_{\substack{\boldsymbol{\delta} \neq 0, \\ |\boldsymbol{\delta}_{J_0^c}|_1 \leq c_0 |\boldsymbol{\delta}_{J_0}|_1}} \frac{|X\boldsymbol{\delta}|_2}{\sqrt{n} |\boldsymbol{\delta}_{J_0}|_2} > 0.
$$

The integer s here plays the role of an upper bound on the sparsity $\mathcal{M}(\beta)$ of a vector of coefficients β.

Note that, if Assumption RE(s, c_0) is satisfied with $c_0 \geq 1$, then

$$
\min\{|X\boldsymbol{\delta}|_2 : \mathcal{M}(\boldsymbol{\delta}) \leq 2s, \boldsymbol{\delta} \neq 0\} > 0.
$$

In other words, the square submatrices of size $\leq 2s$ of the Gram matrix are necessarily positive definite. Indeed, suppose that, for some $\boldsymbol{\delta} \neq 0$, we have simultaneously $\mathcal{M}(\boldsymbol{\delta}) \leq 2s$ and $X\boldsymbol{\delta} = 0$. Partition $J(\boldsymbol{\delta})$ in two sets $J(\boldsymbol{\delta}) = I_0 \cup I_1$, such that $|I_i| \leq s$, $i = 0, 1$. Without loss of generality, suppose that $|\boldsymbol{\delta}_{I_1}|_1 \leq |\boldsymbol{\delta}_{I_0}|_1$. Since, clearly, $|\boldsymbol{\delta}_{I_1}|_1 = |\boldsymbol{\delta}_{I_0^c}|_1$ and $c_0 \geq 1$, we have $|\boldsymbol{\delta}_{I_0^c}|_1 \leq c_0 |\boldsymbol{\delta}_{I_0}|_1$. Hence, $\kappa(s, c_0) = 0$, a contradiction.

To introduce the second assumption, we need some notation. For integers s, m such that $1 \le s \le M/2$ and $m \ge s$, $s+m \le M$, a vector $\boldsymbol{\delta} \in \mathbb{R}^M$ and a set of indices $J_0 \subseteq \{1, \dots, M\}$ with $|J_0| \le s$; denote by J_1 the subset of $\{1, \dots, M\}$ corresponding to the m largest in absolute value coordinates of $\boldsymbol{\delta}$ outside of J_0, and define $J_{01} \stackrel{\triangle}{=} J_0 \cup J_1$. Clearly, J_1 and J_{01} depend on m, but we do not indicate this in our notation for the sake of brevity.

ASSUMPTION RE(s, m, c_0).

$$\kappa(s, m, c_0) \stackrel{\triangle}{=} \min_{\substack{J_0 \subseteq \{1,\dots,M\}, \\ |J_0| \le s}} \min_{\substack{\boldsymbol{\delta} \ne 0, \\ |\boldsymbol{\delta}_{J_0^c}|_1 \le c_0 |\boldsymbol{\delta}_{J_0}|_1}} \frac{|X\boldsymbol{\delta}|_2}{\sqrt{n}|\boldsymbol{\delta}_{J_{01}}|_2} > 0.$$

Note that the only difference between the two assumptions is in the denominators, and $\kappa(s, m, c_0) \le \kappa(s, c_0)$. As written, for fixed n, the two assumptions are equivalent. However, asymptotically for large n, Assumption RE(s, c_0) is less restrictive than RE(s, m, c_0), since the ratio $\kappa(s, m, c_0)/\kappa(s, c_0)$ may tend to 0 if s and m depend on n. For our bounds on the prediction loss and on the ℓ_1 loss of the Lasso and Dantzig estimators, we will only need Assumption RE(s, c_0). Assumption RE(s, m, c_0) will be required exclusively for the bounds on the ℓ_p loss with $1 < p \le 2$.

Note also that Assumptions RE(s', c_0) and RE(s', m, c_0) imply Assumptions RE(s, c_0) and RE(s, m, c_0), respectively, if $s' > s$.

4. Discussion of the RE assumptions. There exist several simple sufficient conditions for Assumptions RE(s, c_0) and RE(s, m, c_0) to hold. Here, we discuss some of them.

For a real number $1 \le u \le M$, we introduce the following quantities that we will call *restricted eigenvalues*:

$$\phi_{\min}(u) = \min_{x \in \mathbb{R}^M : 1 \le \mathcal{M}(x) \le u} \frac{x^{\mathrm{T}} \Psi_n x}{|x|_2^2},$$

$$\phi_{\max}(u) = \max_{x \in \mathbb{R}^M : 1 \le \mathcal{M}(x) \le u} \frac{x^{\mathrm{T}} \Psi_n x}{|x|_2^2}.$$

Denote by X_J the $n \times |J|$ submatrix of X obtained by removing from X the columns that do not correspond to the indices in J, and, for $1 \le m_1, m_2 \le M$, introduce the following quantities called *restricted correlations*:

$$\theta_{m_1,m_2} = \max\left\{ \frac{c_1^{\mathrm{T}} X_{I_1}^{\mathrm{T}} X_{I_2} c_2}{n|c_1|_2|c_2|_2} : I_1 \cap I_2 = \varnothing, |I_i| \le m_i, c_i \in \mathbb{R}^{I_i} \setminus \{0\}, i = 1, 2 \right\}.$$

In Lemma 4.1, below, we show that a sufficient condition for RE(s, c_0) and RE(s, s, c_0) to hold is given, for example, by the following assumption on the Gram matrix.

ASSUMPTION 1. Assume that

$$\phi_{\min}(2s) > c_0\theta_{s,2s}$$

for some integer $1 \le s \le M/2$ and a constant $c_0 > 0$.

This condition with $c_0 = 1$ appeared in [7], in connection with the Dantzig selector. Assumption 1 is more general, in that we can have an arbitrary constant $c_0 > 0$ that will allow us to cover not only the Dantzig selector but also the Lasso estimators and to prove oracle inequalities for the prediction loss when the model is nonparametric.

Our second sufficient condition for $\mathrm{RE}(s, c_0)$ and $\mathrm{RE}(s, m, c_0)$ does not need bounds on correlations. Only bounds on the minimal and maximal eigenvalues of "small" submatrices of the Gram matrix Ψ_n are involved.

ASSUMPTION 2. Assume that

$$m\phi_{\min}(s+m) > c_0^2 s\phi_{\max}(m)$$

for some integers s, m, such that $1 \le s \le M/2$, $m \ge s$ and $s + m \le M$, and a constant $c_0 > 0$.

Assumption 2 can be viewed as a weakening of the condition on $\phi_{\min}$ in [18]. Indeed, taking $s + m = s\log n$ (we assume, without loss of generality, that $s\log n$ is an integer and $n > 3$) and assuming that $\phi_{\max}(\cdot)$ is uniformly bounded by a constant, we get that Assumption 2 is equivalent to

$$\phi_{\min}(s\log n) > c/\log n,$$

where $c > 0$ is a constant. The corresponding, slightly stronger, assumption in [18] is stated in asymptotic form, for $s = s_n \to \infty$, as

$$\liminf_n \phi_{\min}(s_n \log n) > 0.$$

The following two constants are useful when Assumptions 1 and 2 are considered:

$$\kappa_1(s, c_0) = \sqrt{\phi_{\min}(2s)}\left(1 - \frac{c_0\theta_{s,2s}}{\phi_{\min}(2s)}\right)$$

and

$$\kappa_2(s, m, c_0) = \sqrt{\phi_{\min}(s+m)}\left(1 - c_0\sqrt{\frac{s\phi_{\max}(m)}{m\phi_{\min}(s+m)}}\right).$$

The next lemma shows that if Assumptions 1 or 2 are satisfied, then the quadratic form $x^{\mathrm{T}}\Psi_n x$ is positive definite on some restricted sets of vectors x. The construction of the lemma is inspired by Candes and Tao [7] and covers, in particular, the corresponding result in [7].

LEMMA 4.1. *Fix an integer $1 \le s \le M/2$ and a constant $c_0 > 0$.*

(i) *Let Assumption* 1 *be satisfied. Then, Assumptions* RE(s, c_0) *and* RE(s, s, c_0) *hold with* $\kappa(s, c_0) = \kappa(s, s, c_0) = \kappa_1(s, c_0)$. *Moreover, for any subset* J_0 *of* $\{1, \dots, M\}$, *with cardinality* $|J_0| \le s$, *and any* $\boldsymbol{\delta} \in \mathbb{R}^M$ *such that*

$$|\boldsymbol{\delta}_{J_0^c}|_1 \le c_0 |\boldsymbol{\delta}_{J_0}|_1, \tag{4.1}$$

we have

$$\frac{1}{\sqrt{n}} |P_{01} X \boldsymbol{\delta}|_2 \ge \kappa_1(s, c_0) |\boldsymbol{\delta}_{J_{01}}|_2,$$

where P_{01} *is the projector in* $\mathbb{R}^M$ *on the linear span of the columns of* $X_{J_{01}}$.

(ii) *Let Assumption* 2 *be satisfied. Then, Assumptions* RE(s, c_0) *and* RE(s, m, c_0) *hold with* $\kappa(s, c_0) = \kappa(s, m, c_0) = \kappa_2(s, m, c_0)$. *Moreover, for any subset* J_0 *of* $\{1, \dots, M\}$, *with cardinality* $|J_0| \le s$, *and any* $\boldsymbol{\delta} \in \mathbb{R}^M$ *such that* (4.1) *holds, we have*

$$\frac{1}{\sqrt{n}} |P_{01} X \boldsymbol{\delta}|_2 \ge \kappa_2(s, m, c_0) |\boldsymbol{\delta}_{J_{01}}|_2.$$

The proof of the lemma is given in Appendix A.

There exist other sufficient conditions for Assumptions RE(s, c_0) and RE(s, m, c_0) to hold. We mention here three of them implying Assumption RE(s, c_0). The first one is the following [1].

ASSUMPTION 3. For an integer s such that $1 \le s \le M$, we have

$$\phi_{\min}(s) > 2 c_0 \theta_{s,1} \sqrt{s},$$

where $c_0 > 0$ is a constant.

To argue that Assumption 3 implies RE(s, c_0), it suffices to remark that

$$\begin{aligned} \frac{1}{n} |X \boldsymbol{\delta}|_2^2 &\ge \frac{1}{n} \boldsymbol{\delta}_{J_0}^{\mathrm{T}} X^{\mathrm{T}} X \boldsymbol{\delta}_{J_0} - \frac{2}{n} |\boldsymbol{\delta}_{J_0}^{\mathrm{T}} X^{\mathrm{T}} X \boldsymbol{\delta}_{J_0^c}| \\ &\ge \phi_{\min}(s) |\boldsymbol{\delta}_{J_0}|_2^2 - \frac{2}{n} |\boldsymbol{\delta}_{J_0}^{\mathrm{T}} X^{\mathrm{T}} X \boldsymbol{\delta}_{J_0^c}| \end{aligned}$$

and, if (4.1) holds,

$$\begin{aligned} |\boldsymbol{\delta}_{J_0}^{\mathrm{T}} X^{\mathrm{T}} X \boldsymbol{\delta}_{J_0^c}| / n &\le |\boldsymbol{\delta}_{J_0^c}|_1 \max_{j \in J_0^c} |\boldsymbol{\delta}_{J_0}^{\mathrm{T}} X^{\mathrm{T}} \mathbf{x}_{(j)}| / n \\ &\le \theta_{s,1} |\boldsymbol{\delta}_{J_0^c}|_1 |\boldsymbol{\delta}_{J_0}|_2 \\ &\le c_0 \theta_{s,1} \sqrt{s} |\boldsymbol{\delta}_{J_0}|_2^2. \end{aligned}$$

Another type of assumption related to "mutual coherence" [8] is discussed in connection to Lasso in [4, 5]. We state it in two different forms, which are given below.

ASSUMPTION 4. For an integer s such that $1 \le s \le M$, we have

$$\phi_{\min}(s) > 2c_0\theta_{1,1}s,$$

where $c_0 > 0$ is a constant.

It is easy to see that Assumption 4 implies RE(s, c_0). Indeed, if (4.1) holds,

$$\begin{aligned}\frac{1}{n}|X\boldsymbol{\delta}|_2^2 &\ge \frac{1}{n}\boldsymbol{\delta}_{J_0}^{\mathrm{T}}X^{\mathrm{T}}X\boldsymbol{\delta}_{J_0} - 2\theta_{1,1}|\boldsymbol{\delta}_{J_0^c}|_1|\boldsymbol{\delta}_{J_0}|_1\\ &\ge \phi_{\min}(s)|\boldsymbol{\delta}_{J_0}|_2^2 - 2c_0\theta_{1,1}|\boldsymbol{\delta}_{J_0}|_1^2 \qquad (4.2)\\ &\ge \bigl(\phi_{\min}(s) - 2c_0\theta_{1,1}s\bigr)|\boldsymbol{\delta}_{J_0}|_2^2.\end{aligned}$$

If all the diagonal elements of matrix $X^{\mathrm{T}}X/n$ are equal to 1 (and thus $\theta_{1,1}$ coincides with the mutual coherence [8]), then a simple sufficient condition for Assumption RE(s, c_0) to hold is stated as follows.

ASSUMPTION 5. All the diagonal elements of the Gram matrix Ψ_n are equal to 1, and for an integer s, such that $1 \le s \le M$, we have

$$\theta_{1,1} < \frac{1}{(1+2c_0)s}, \qquad (4.3)$$

where $c_0 > 0$ is a constant.

In fact, separating the diagonal and off-diagonal terms of the quadratic form, we get

$$\boldsymbol{\delta}_{J_0}^{\mathrm{T}}X^{\mathrm{T}}X\boldsymbol{\delta}_{J_0}/n \ge |\boldsymbol{\delta}_{J_0}|_2^2 - \theta_{1,1}|\boldsymbol{\delta}_{J_0}|_1^2 \ge |\boldsymbol{\delta}_{J_0}|_2^2(1-\theta_{1,1}s).$$

Combining this inequality with (4.2), we see that Assumption RE(s, c_0) is satisfied whenever (4.3) holds.

Unfortunately, Assumption RE(s, c_0) has some weakness. Let, for example, f_j, $j = 1, \dots, 2^m - 1$, be the Haar wavelet basis on $[0, 1]$ ($M = 2^m$), and consider $Z_i = i/n$, $i = 1, \dots, n$. If $M \gg n$, then it is clear that $\phi_{\min}(1) = 0$, since there are functions f_j on the highest resolution level whose supports (of length M^{-1}) contain no points Z_i. So, none of Assumptions 1–4 hold. A less severe, although similar, situation is when we consider step functions $f_j(t) = I_{\{t<j/M\}}$ for $t \in [0, 1]$. It is clear that $\phi_{\min}(2) = O(1/M)$, although sparse representation in this basis is very natural. Intuitively, the problem arises only because we include very high resolution components. Therefore, we may try to restrict the set J_0 in RE(s, c_0) to low resolution components, which is quite reasonable, because the "true" or "interesting" vectors of parameters β are often characterized by such J_0. This idea is formalized in Section 6 (cf. Corollary 6.2, see also a remark after Theorem 7.2 in Section 7).

5. Approximate equivalence. In this section, we prove a type of approximate equivalence between the Lasso and the Dantzig selector. It is expressed as closeness of the prediction losses $\|\widehat{f}_D - f\|_n^2$ and $\|\widehat{f}_L - f\|_n^2$ when the number of nonzero components of the Lasso or the Dantzig selector is small as compared to the sample size.

THEOREM 5.1. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Fix $n \geq 1$, $M \geq 2$. Let Assumption* RE$(s, 1)$ *be satisfied with $1 \leq s \leq M$. Consider the Dantzig estimator $\widehat{f}_D$ defined by* (2.5)–(2.4) *with*

$$r = A\sigma\sqrt{\frac{\log M}{n}},$$

where $A > 2\sqrt{2}$, and consider the Lasso estimator $\widehat{f}_L$ defined by (2.1)–(2.2) *with the same r.*

If $\mathcal{M}(\widehat{\beta}_L) \leq s$, then, with probability at least $1 - M^{1-A^2/8}$, we have

$$\bigl|\|\widehat{f}_D - f\|_n^2 - \|\widehat{f}_L - f\|_n^2\bigr| \leq 16A^2 \frac{\mathcal{M}(\widehat{\beta}_L)\sigma^2}{n} \frac{f_{\max}^2}{\kappa^2(s, 1)} \log M. \tag{5.1}$$

Note that the RHS of (5.1) is bounded by a product of three factors (and a numerical constant which, unfortunately, equals at least 128). The first factor $\mathcal{M}(\widehat{\beta}_L)\sigma^2/n \leq s\sigma^2/n$ corresponds to the error rate for prediction in regression with s parameters. The two other factors, $\log M$ and $f_{\max}^2/\kappa^2(s, 1)$, can be regarded as a price to pay for the large number of regressors. If the Gram matrix Ψ_n equals the identity matrix (the white noise model), then there is only the $\log M$ factor. In the general case, there is another factor $f_{\max}^2/\kappa^2(s, 1)$ representing the extent to which the Gram matrix is ill-posed for estimation of sparse vectors.

We also have the following result that we state, for simplicity, under the assumption that $\|f_j\|_n = 1$, $j = 1, \ldots, M$. It gives a bound in the spirit of Theorem 5.1 but with $\mathcal{M}(\widehat{\beta}_D)$ rather than $\mathcal{M}(\widehat{\beta}_L)$ on the right-hand side.

THEOREM 5.2. *Let the assumptions of Theorem* 5.1 *hold, but with* RE$(s, 5)$ *in place of* RE$(s, 1)$, *and let $\|f_j\|_n = 1$, $j = 1, \ldots, M$. If $\mathcal{M}(\widehat{\beta}_D) \leq s$, then, with probability at least $1 - M^{1-A^2/8}$, we have*

$$\|\widehat{f}_L - f\|_n^2 \leq 10\|\widehat{f}_D - f\|_n^2 + 81A^2 \frac{\mathcal{M}(\widehat{\beta}_D)\sigma^2}{n} \frac{\log M}{\kappa^2(s, 5)}. \tag{5.2}$$

REMARK. The approximate equivalence is essentially that of the rates as Theorem 5.1 exhibits. A statement free of $\mathcal{M}(\beta)$ holds for linear regression, see discussion after Theorems 7.2 and 7.3 below.

6. Oracle inequalities for prediction loss. Here, we prove sparsity oracle inequalities for the prediction loss of the Lasso and Dantzig estimators. These inequalities allow us to bound the difference between the prediction errors of the estimators and the best sparse approximation of the regression function (by an oracle that knows the truth but is constrained by sparsity). The results of this section, together with those of Section 5, show that the distance between the prediction losses of the Dantzig and Lasso estimators is of the same order as the distances between them and their oracle approximations.

A general discussion of sparsity oracle inequalities can be found in [23]. Such inequalities have been recently obtained for the Lasso type estimators in a number of settings [2–6, 14] and [25]. In particular, the regression model with fixed design that we study here is considered in [2–4]. The assumptions on the Gram matrix Ψ_n in [2–4] are more restrictive than ours. In those papers, either Ψ_n is positive definite, or a mutual coherence condition similar to (4.3) is imposed.

THEOREM 6.1. *Let W_i be independent $\mathcal{N}(0,\sigma^2)$ random variables with $\sigma^2 > 0$. Fix some $\varepsilon > 0$ and integers $n \geq 1$, $M \geq 2$, $1 \leq s \leq M$. Let Assumption* RE$(s, 3+4/\varepsilon)$ *be satisfied. Consider the Lasso estimator $\widehat{\widehat{f}}_L$ defined by* (2.1)–(2.2) *with*

$$r = A\sigma\sqrt{\frac{\log M}{n}}$$

for some $A > 2\sqrt{2}$. Then, with probability at least $1 - M^{1-A^2/8}$, we have

$$\begin{aligned} &\|\widehat{f}_L - f\|_n^2 \\ &\quad\leq (1+\varepsilon) \inf_{\substack{\beta\in\mathbb{R}^M:\\ \mathcal{M}(\beta)\leq s}} \left\{ \|f_\beta - f\|_n^2 + \frac{C(\varepsilon) f_{\max}^2 A^2\sigma^2}{\kappa^2(s, 3+4/\varepsilon)} \frac{\mathcal{M}(\beta)\log M}{n} \right\}, \end{aligned} \tag{6.1}$$

where $C(\varepsilon) > 0$ is a constant depending only on ε.

We now state, as a corollary, a softer version of Theorem 6.1 that can be used to eliminate the pathologies mentioned at the end of Section 4. For this purpose, we define

$$\mathcal{J}_{s,\gamma,c_0} = \left\{ J_0 \subset \{1,\dots,M\} : |J_0| \leq s \text{ and } \min_{\substack{\delta\neq 0,\\ |\delta_{J_0^c}|_1 \leq c_0|\delta_{J_0}|_1}} \frac{|X\delta|_2}{\sqrt{n}|\delta_{J_0}|_2} \geq \gamma \right\},$$

where $\gamma > 0$ is a constant, and set

$$\Lambda_{s,\gamma,c_0} = \{\beta : J(\beta) \in \mathcal{J}_{s,\gamma,c_0}\}.$$

In similar way, we define $\mathcal{J}_{s,\gamma,m,c_0}$ and Λ_{s,γ,m,c_0} corresponding to Assumption RE(s, m, c_0).

COROLLARY 6.2. *Let W_i, s and the Lasso estimator $\widehat{f}_L$ be the same as in Theorem* 6.1. *Then, for all $n \geq 1$, $\varepsilon > 0$, and $\gamma > 0$, with probability at least $1 - M^{1-A^2/8}$ we have*

$$\|\widehat{f}_L - f\|_n^2 \leq (1+\varepsilon) \inf_{\beta \in \bar{\Lambda}_{s,\gamma,\varepsilon}} \left\{ \|f_\beta - f\|_n^2 + \frac{C(\varepsilon) f_{\max}^2 A^2 \sigma^2}{\gamma^2} \left(\frac{\mathcal{M}(\beta) \log M}{n} \right) \right\},$$

where $\bar{\Lambda}_{s,\gamma,\varepsilon} = \{\beta \in \Lambda_{s,\gamma,3+4/\varepsilon} : \mathcal{M}(\beta) \leq s\}$.

To obtain this corollary, it suffices to observe that the proof of Theorem 6.1 goes through if we drop Assumption RE$(s, 3+4/\varepsilon)$, but we assume instead that $\beta \in \Lambda_{s,\gamma,3+4/\varepsilon}$, and we replace $\kappa(s, 3+4/\varepsilon)$ by γ.

We would like now to get a sparsity oracle inequality similar to that of Theorem 6.1 for the Dantzig estimator $\widehat{f}_D$. We will need a mild additional assumption on f. This is due to the fact that not every $\beta \in \mathbb{R}^M$ obeys the Dantzig constraint; thus, we cannot assure the key relation (B.9) for all $\beta \in \mathbb{R}^M$. One possibility would be to prove inequality as (6.1), where the infimum on the right hand side is taken over β satisfying not only $\mathcal{M}(\beta) \leq s$ but also the Dantzig constraint. However, this seems not to be very intuitive, since we cannot guarantee that the corresponding f_β gives a good approximation of the unknown function f. Therefore, we choose another approach (cf. [5]), in which we consider f satisfying the *weak sparsity* property relative to the dictionary $f_1, \ldots, f_M$. That is, we assume that there exist an integer s and constant $C_0 < \infty$ such that the set

$$\Lambda_s = \left\{ \beta \in \mathbb{R}^M : \mathcal{M}(\beta) \leq s, \|f_\beta - f\|_n^2 \leq \frac{C_0 f_{\max}^2 r^2}{\kappa^2(s, 3+4/\varepsilon)} \mathcal{M}(\beta) \right\} \tag{6.2}$$

is nonempty. The second inequality in (6.2) says that the "bias" term $\|f_\beta - f\|_n^2$ cannot be much larger than the "variance term" $\sim f_{\max}^2 r^2 \kappa^{-2} \mathcal{M}(\beta)$ [cf. (6.1)]. Weak sparsity is milder than the sparsity property in the usual sense. The latter means that f admits the exact representation $f = f_{\beta^*}$, for some $\beta^* \in \mathbb{R}^M$, with hopefully small $\mathcal{M}(\beta^*) = s$.

PROPOSITION 6.3. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Fix some $\varepsilon > 0$ and integers $n \geq 1$, $M \geq 2$. Let f obey the weak sparsity assumption for some $C_0 < \infty$ and some s such that $1 \leq s \max\{C_1(\varepsilon), 1\} \leq M$, where*

$$C_1(\varepsilon) = 4[(1+\varepsilon)C_0 + C(\varepsilon)] \frac{\phi_{\max} f_{\max}^2}{\kappa^2 f_{\min}^2}$$

and $C(\varepsilon)$ is the constant in Theorem 6.1. *Suppose, further, that Assumption* RE$(s \max\{C_1(\varepsilon), 1\}, 3 + 4/\varepsilon)$ *is satisfied. Consider the Dantzig estimator $\widehat{f}_D$ defined by* (2.5)–(2.4) *with*

$$r = A\sigma \sqrt{\frac{\log M}{n}}$$

and $A > 2\sqrt{2}$. *Then, with probability at least* $1 - M^{1-A^2/8}$, *we have*

$$\|\widehat{f}_D - f\|_n^2 \le (1+\varepsilon) \inf_{\beta \in \mathbb{R}^M : \mathcal{M}(\beta) = s} \|f_\beta - f\|_n^2 + C_2(\varepsilon) \frac{f_{\max}^2 A^2 \sigma^2}{\kappa_0^2} \left(\frac{s \log M}{n} \right). \tag{6.3}$$

Here, $C_2(\varepsilon) = 16C_1(\varepsilon) + C(\varepsilon)$ *and* $\kappa_0 = \kappa(\max(C_1(\varepsilon), 1)s, 3 + 4/\varepsilon)$.

Note that the sparsity oracle inequality (6.3) is slightly weaker than the analogous inequality (6.1) for the Lasso. Here, we have $\inf_{\beta \in \mathbb{R}^M : \mathcal{M}(\beta) = s}$ instead of $\inf_{\beta \in \mathbb{R}^M : \mathcal{M}(\beta) \le s}$ in (6.1).

7. Special case. Parametric estimation in linear regression. In this section, we assume that the vector of observations $\mathbf{y} = (Y_1, \dots, Y_n)^{\mathrm{T}}$ is of the form

$$\mathbf{y} = X\beta^* + \mathbf{w}, \tag{7.1}$$

where X is an $n \times M$ deterministic matrix $\beta^* \in \mathbb{R}^M$ and $\mathbf{w} = (W_1, \dots, W_n)^{\mathrm{T}}$.

We consider dimension M that can be of order n and even much larger. Then, β^* is, in general, not uniquely defined. For $M > n$, if (7.1) is satisfied for $\beta^* = \beta_0$, then there exists an affine space $\mathcal{U} = \{\beta^* : X\beta^* = X\beta_0\}$ of vectors satisfying (7.1). The results of this section are valid for any β^* such that (7.1) holds. However, we will suppose that Assumption RE(s, c_0) holds with $c_0 \ge 1$ and that $\mathcal{M}(\beta^*) \le s$. Then, the set $\mathcal{U} \cap \{\beta^* : \mathcal{M}(\beta^*) \le s\}$ reduces to a single element (cf. Remark 2 at the end of this section). In this sense, there is a unique sparse solution of (7.1).

Our goal in this section, unlike that of the previous ones, is to estimate both $X\beta^*$ for the purpose of prediction and β^* itself for purpose of model selection. We will see that meaningful results are obtained when the sparsity index $\mathcal{M}(\beta^*)$ is small.

It will be assumed throughout this section that the diagonal elements of the Gram matrix $\Psi_n = X^{\mathrm{T}}X/n$ are all equal to 1 (this is equivalent to the condition $\|f_j\|_n = 1$, $j = 1, \dots, M$, in the notation of previous sections). Then, the Lasso estimator of β^* in (7.1) is defined by

$$\widehat{\beta}_L = \arg\min_{\beta \in \mathbb{R}^M} \left\{ \frac{1}{n} |\mathbf{y} - X\beta|_2^2 + 2r|\beta|_1 \right\}. \tag{7.2}$$

The correspondence between the notation here and that of the previous sections is

$$\|f_\beta\|_n^2 = |X\beta|_2^2/n, \qquad \|f_\beta - f\|_n^2 = |X(\beta - \beta^*)|_2^2/n,$$
$$\|\widehat{f}_L - f\|_n^2 = |X(\widehat{\beta}_L - \beta^*)|_2^2/n.$$

The Dantzig selector for linear model (7.1) is defined by

$$\widehat{\beta}_D = \arg\min_{\beta \in \Lambda} |\beta|_1, \tag{7.3}$$

where

$$\Lambda = \left\{ \beta \in \mathbb{R}^M : \left| \frac{1}{n} X^{\mathrm{T}} (\mathbf{y} - X\beta) \right|_\infty \le r \right\}$$

is the set of all β satisfying the Dantzig constraint.

We first get bounds on the rate of convergence of Dantzig selector.

THEOREM 7.1. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$, let all the diagonal elements of the matrix $X^{\mathrm{T}}X/n$ be equal to 1 and $\mathcal{M}(\beta^*) \le s$, where $1 \le s \le M$, $n \ge 1$, $M \ge 2$. Let Assumption* RE$(s, 1)$ *be satisfied. Consider the Dantzig selector $\widehat{\beta}_D$ defined by* (7.3) *with*

$$r = A\sigma \sqrt{\frac{\log M}{n}}$$

and $A > \sqrt{2}$. Then, with probability at least $1 - M^{1-A^2/2}$, we have

$$|\widehat{\beta}_D - \beta^*|_1 \le \frac{8A}{\kappa^2(s, 1)} \sigma s \sqrt{\frac{\log M}{n}}, \tag{7.4}$$

$$|X(\widehat{\beta}_D - \beta^*)|_2^2 \le \frac{16A^2}{\kappa^2(s, 1)} \sigma^2 s \log M. \tag{7.5}$$

If Assumption RE$(s, m, 1)$ *is satisfied, then, with the same probability as above, simultaneously for all $1 < p \le 2$, we have*

$$|\widehat{\beta}_D - \beta^*|_p^p \le 2^{p-1} 8 \left\{ 1 + \sqrt{\frac{s}{m}} \right\}^{2(p-1)} s \left(\frac{A\sigma}{\kappa^2(s, m, 1)} \sqrt{\frac{\log M}{n}} \right)^p. \tag{7.6}$$

Note that, since $s \le m$, the factor in curly brackets in (7.6) is bounded by a constant independent of s and m. Under Assumption 1 in Section 4, with $c_0 = 1$ [which is less general than RE$(s, s, 1)$, cf. Lemma 4.1(i)], a bound of the form (7.6) for the case $p = 2$ is established by Candes and Tao [7].

Bounds on the rate of convergence of the Lasso selector are quite similar to those obtained in Theorem 7.1. They are given by the following result.

THEOREM 7.2. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Let all the diagonal elements of the matrix $X^{\mathrm{T}}X/n$ be equal to 1, and let $\mathcal{M}(\beta^*) \le s$, where $1 \le s \le M$, $n \ge 1$, $M \ge 2$. Let Assumption* RE$(s, 3)$ *be satisfied. Consider the Lasso estimator $\widehat{\beta}_L$ defined by* (7.2) *with*

$$r = A\sigma \sqrt{\frac{\log M}{n}}$$

and $A > 2\sqrt{2}$. *Then, with probability at least* $1 - M^{1-A^2/8}$, *we have*

$$|\widehat{\beta}_L - \beta^*|_1 \le \frac{16A}{\kappa^2(s,3)}\sigma s\sqrt{\frac{\log M}{n}}, \tag{7.7}$$

$$|X(\widehat{\beta}_L - \beta^*)|_2^2 \le \frac{16A^2}{\kappa^2(s,3)}\sigma^2 s \log M, \tag{7.8}$$

$$\mathcal{M}(\widehat{\beta}_L) \le \frac{64\phi_{\max}}{\kappa^2(s,3)}s. \tag{7.9}$$

If Assumption RE$(s,m,3)$ *is satisfied, then, with the same probability as above, simultaneously for all* $1 < p \le 2$, *we have*

$$|\widehat{\beta}_L - \beta^*|_p^p \le 16\left\{1 + 3\sqrt{\frac{s}{m}}\right\}^{2(p-1)} s\left(\frac{A\sigma}{\kappa^2(s,m,3)}\sqrt{\frac{\log M}{n}}\right)^p. \tag{7.10}$$

Inequalities of the form similar to (7.7) and (7.8) can be deduced from the results of [3] under more restrictive conditions on the Gram matrix (the mutual coherence assumption, cf. Assumption 5 of Section 4).

Assumptions RE$(s,1)$ and RE$(s,3)$, respectively, can be dropped in Theorems 7.1 and 7.2 if we assume $\beta^* \in \Lambda_{s,\gamma,c_0}$ with $c_0 = 1$ or $c_0 = 3$ as appropriate. Then, (7.4) and (7.5) or, respectively, (7.7) and (7.8) hold with $\kappa = \gamma$. This is analogous to Corollary 6.2. Similarly, (7.6) and (7.10) hold with $\kappa = \gamma$ if $\beta^* \in \Lambda_{s,\gamma,m,c_0}$ with $c_0 = 1$ or $c_0 = 3$ as appropriate.

Observe that, combining Theorems 7.1 and 7.2, we can immediately get bounds for the differences between Lasso and Dantzig selector $|\widehat{\beta}_L - \widehat{\beta}_D|_p^p$ and $|X(\widehat{\beta}_L - \widehat{\beta}_D)|_2^2$. Such bounds have the same form as those of Theorems 7.1 and 7.2, up to numerical constants. Another way of estimating these differences follows directly from the proof of Theorem 7.1. It suffices to observe that the only property of β^* used in that proof is the fact that β^* satisfies the Dantzig constraint on the event of given probability, which is also true for the Lasso solution $\widehat{\beta}_L$. So, we can replace β^* by $\widehat{\beta}_L$ and s by $\mathcal{M}(\widehat{\beta}_L)$ everywhere in Theorem 7.1. Generalizing a bit more, we easily derive the following fact.

THEOREM 7.3. *The result of Theorem* 7.1 *remains valid if we replace* $|\widehat{\beta}_D - \beta^*|_p^p$ *by* $\sup\{|\widehat{\beta}_D - \beta|_p^p : \beta \in \Lambda, \mathcal{M}(\beta) \le s\}$ *for* $1 \le p \le 2$ *and* $|X(\widehat{\beta}_D - \beta^*)|_2^2$ *by* $\sup\{|X(\widehat{\beta}_D - \beta)|_2^2 : \beta \in \Lambda, \mathcal{M}(\beta) \le s\}$, *respectively. Here,* Λ *is the set of all vectors satisfying the Dantzig constraint.*

REMARKS.

1. Theorems 7.1 and 7.2 only give nonasymptotic upper bounds on the loss, with some probability and under some conditions. The probability depends on M and the conditions depend on n and M. Recall that Assumptions RE(s,c_0) and RE(s,m,c_0) are imposed on the $n \times M$ matrix X. To deduce asymptotic conver-

gence (as $n \to \infty$ and/or as $M \to \infty$) from Theorems 7.1 and 7.2, we would need some very strong additional properties, such as simultaneous validity of Assumption RE(s, c_0) or RE(s, m, c_0) (with one and the same constant κ) for infinitely many n and M.

2. Note that neither Assumption RE(s, c_0) or RE(s, m, c_0) implies identifiability of β^* in the linear model (7.1). However, the vector β^* appearing in the statements of Theorems 7.1 and 7.2 is uniquely defined, because we additionally suppose that $\mathcal{M}(\beta^*) \le s$ and $c_0 \ge 1$. Indeed, if there exists a β' such that $X\beta' = X\beta^*$, and $\mathcal{M}(\beta') \le s$, then, in view of assumption RE(s, c_0) with $c_0 \ge 1$, we necessarily have $\beta^* = \beta'$ [cf. discussion following the definition of RE(s, c_0)]. On the other hand, Theorem 7.3 applies to certain values of β that do not come from the model (7.1) at all.

3. For the smallest value of A (which is $A = 2\sqrt{2}$) the constants in the bound of Theorem 7.2 for the Lasso are larger than the corresponding numerical constants for the Dantzig selector given in Theorem 7.1, again, for the smallest admissible value $A = \sqrt{2}$. On the contrary, the Dantzig selector has certain defects as compared to Lasso when the model is nonparametric, as discussed in Section 6. In particular, to obtain sparsity oracle inequalities for the Dantzig selector, we need some restrictions on f, for example, the weak sparsity property. On the other hand, the sparsity oracle inequality (6.1) for the Lasso is valid with no restriction on f.

4. The proofs of Theorems 7.1 and 7.2 differ mainly in the value of the tuning constant, which is $c_0 = 1$ in Theorem 7.1 and $c_0 = 3$ in Theorem 7.2. Note that, since the Lasso solution satisfies the Dantzig constraint, we could have obtained a result similar to Theorem 7.2, but with less accurate numerical constants, by simply conducting the proof of Theorem 7.1 with $c_0 = 3$. However, we act differently, and we deduce (B.30) directly from (B.1) and not from (B.25). This is done only for the sake of improving the constants. In fact, using (B.25) with $c_0 = 3$ would yield (B.30) with the doubled constant on the right-hand side.

5. For the Dantzig selector in the linear regression model and under Assumptions 1 or 2, some further improvement of constants in the ℓ_p bounds for the coefficients can be achieved by applying the general version of Lemma 4.1 with the projector P_{01} inside. We do not pursue this issue here.

6. All of our results are stated with probabilities at least $1 - M^{1-A^2/2}$ or $1 - M^{1-A^2/8}$. These are reasonable (but not the most accurate) lower bounds on the probabilities $\mathbb{P}(\mathcal{B})$ and $\mathbb{P}(\mathcal{A})$, respectively. We have chosen them for readability. Inspection of (B.4) shows that they can be refined to $1 - 2M\Phi(A\sqrt{\log M})$ and $1 - 2M\Phi(A\sqrt{\log M}/2)$, respectively, where $\Phi(\cdot)$ is the standard normal c.d.f.

APPENDIX A

PROOF OF LEMMA 4.1. Consider a partition J_0^c into subsets of size m, with the last subset of size $\le m$: $J_0^c = \bigcup_{k=1}^{K} J_k$, where $K \ge 1$, $|J_k| = m$ for

$k = 1, \dots, K-1$ and $|J_K| \le m$, such that J_k is the set of indices corresponding to m largest in absolute value coordinates of $\boldsymbol{\delta}$ outside $\bigcup_{j=1}^{k-1} J_j$ (for $k < K$) and J_K is the remaining subset. We have

$$
\begin{aligned}
|P_{01} X\boldsymbol{\delta}|_2 &\ge |P_{01} X\boldsymbol{\delta}_{J_{01}}|_2 - \left|\sum_{k=2}^{K} P_{01} X\boldsymbol{\delta}_{J_k}\right|_2 \\
&= |X\boldsymbol{\delta}_{J_{01}}|_2 - \left|\sum_{k=2}^{K} P_{01} X\boldsymbol{\delta}_{J_k}\right|_2 \qquad \text{(A.1)} \\
&\ge |X\boldsymbol{\delta}_{J_{01}}|_2 - \sum_{k=2}^{K} |P_{01} X\boldsymbol{\delta}_{J_k}|_2.
\end{aligned}
$$

We will prove first part (ii) of the lemma. Since for $k \ge 1$ the vector $\boldsymbol{\delta}_{J_k}$ has only m nonzero components, we obtain

$$
\frac{1}{\sqrt{n}} |P_{01} X\boldsymbol{\delta}_{J_k}|_2 \le \frac{1}{\sqrt{n}} |X\boldsymbol{\delta}_{J_k}|_2 \le \sqrt{\phi_{\max}(m)}\, |\boldsymbol{\delta}_{J_k}|_2. \qquad \text{(A.2)}
$$

Next, as in [7], we observe that $|\boldsymbol{\delta}_{J_{k+1}}|_2 \le |\boldsymbol{\delta}_{J_k}|_1/\sqrt{m}$, $k = 1, \dots, K-1$. Therefore,

$$
\sum_{k=2}^{K} |\boldsymbol{\delta}_{J_k}|_2 \le \frac{|\boldsymbol{\delta}_{J_0^c}|_1}{\sqrt{m}} \le \frac{c_0 |\boldsymbol{\delta}_{J_0}|_1}{\sqrt{m}} \le c_0 \sqrt{\frac{s}{m}}\, |\boldsymbol{\delta}_{J_0}|_2 \le c_0 \sqrt{\frac{s}{m}}\, |\boldsymbol{\delta}_{J_{01}}|_2, \qquad \text{(A.3)}
$$

where we used (4.1). From (A.1)–(A.3), we find

$$
\begin{aligned}
\frac{1}{\sqrt{n}} |X\boldsymbol{\delta}|_2 &\ge \frac{1}{\sqrt{n}} |X\boldsymbol{\delta}_{J_{01}}|_2 - c_0 \sqrt{\phi_{\max}(m)} \sqrt{\frac{s}{m}}\, |\boldsymbol{\delta}_{J_{01}}|_2 \\
&\ge \left(\sqrt{\phi_{\min}(s+m)} - c_0 \sqrt{\phi_{\max}(m)} \sqrt{\frac{s}{m}} \right) |\boldsymbol{\delta}_{J_{01}}|_2,
\end{aligned}
$$

which proves part (ii) of the lemma.

The proof of part (i) is analogous. The only difference is that we replace, in the above argument, m by s, and instead of (A.2), we use the bound (cf. [7])

$$
\frac{1}{\sqrt{n}} |P_{01} X\boldsymbol{\delta}_{J_k}|_2 \le \frac{\theta_{s,2s}}{\sqrt{\phi_{\min}(2s)}} |\boldsymbol{\delta}_{J_k}|_2. \qquad \square
$$

APPENDIX B: TWO LEMMAS AND THE PROOFS OF THE RESULTS

LEMMA B.1. *Fix $M \ge 2$ and $n \ge 1$. Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$, and let $\widehat{f}_L$ be the Lasso estimator defined by* (2.2) *with*

$$
r = A\sigma \sqrt{\frac{\log M}{n}}
$$

for some $A > 2\sqrt{2}$. Then, with probability at least $1 - M^{1-A^2/8}$, we have, simultaneously for all $\beta \in \mathbb{R}^M$,

$$\begin{aligned}
&\|\widehat{f}_L - f\|_n^2 + r\sum_{j=1}^M \|f_j\|_n|\widehat{\beta}_{j,L} - \beta_j| \\
&\qquad \le \|f_\beta - f\|_n^2 + 4r\sum_{j\in J(\beta)} \|f_j\|_n|\widehat{\beta}_{j,L} - \beta_j| \\
&\qquad \le \|f_\beta - f\|_n^2 + 4r\sqrt{\mathcal{M}(\beta)}\sqrt{\sum_{j\in J(\beta)} \|f_j\|_n^2|\widehat{\beta}_{j,L} - \beta_j|^2},
\end{aligned} \tag{B.1}$$

and

$$\left|\frac{1}{n}X^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_L)\right|_\infty \le 3rf_{\max}/2. \tag{B.2}$$

Furthermore, with the same probability,

$$\mathcal{M}(\widehat{\beta}_L) \le 4\phi_{\max} f_{\min}^{-2}(\|\widehat{f}_L - f\|_n^2/r^2), \tag{B.3}$$

where $\phi_{\max}$ denotes the maximal eigenvalue of the matrix $X^{\mathrm{T}}X/n$.

PROOF OF LEMMA B.1. The result (B.1) is essentially Lemma 1 from [5]. For completeness, we give its proof. Set $r_{n,j} = r\|f_j\|_n$. By definition,

$$\widehat{S}(\widehat{\beta}_L) + 2\sum_{j=1}^M r_{n,j}|\widehat{\beta}_{j,L}| \le \widehat{S}(\beta) + 2\sum_{j=1}^M r_{n,j}|\beta_j|$$

for all $\beta \in \mathbb{R}^M$, which is equivalent to

$$\begin{aligned}
&\|\widehat{f}_L - f\|_n^2 + 2\sum_{j=1}^M r_{n,j}|\widehat{\beta}_{j,L}| \\
&\qquad \le \|f_\beta - f\|_n^2 + 2\sum_{j=1}^M r_{n,j}|\beta_j| + \frac{2}{n}\sum_{i=1}^n W_i(\widehat{f}_L - f_\beta)(Z_i).
\end{aligned}$$

Define the random variables $V_j = n^{-1}\sum_{i=1}^n f_j(Z_i)W_i$, $1 \le j \le M$, and the event

$$\mathcal{A} = \bigcap_{j=1}^M \{2|V_j| \le r_{n,j}\}.$$

Using an elementary bound on the tails of Gaussian distribution, we find that the probability of the complementary event $\mathcal{A}^c$ satisfies

$$\begin{aligned}
\mathbb{P}\{\mathcal{A}^c\} &\le \sum_{j=1}^M \mathbb{P}\{\sqrt{n}|V_j| > \sqrt{n}r_{n,j}/2\} \le M\mathbb{P}\{|\eta| \ge r\sqrt{n}/(2\sigma)\} \\
&\le M\exp\left(-\frac{nr^2}{8\sigma^2}\right) = M\exp\left(-\frac{A^2\log M}{8}\right) = M^{1-A^2/8},
\end{aligned} \tag{B.4}$$

where $\eta \sim \mathcal{N}(0,1)$. On the event $\mathcal{A}$ we have

$$\|\widehat{f}_L - f\|_n^2 \le \|f_\beta - f\|_n^2 + \sum_{j=1}^M r_{n,j}|\widehat{\beta}_{j,L} - \beta_j| + \sum_{j=1}^M 2r_{n,j}|\beta_j| - \sum_{j=1}^M 2r_{n,j}|\widehat{\beta}_{j,L}|.$$

Adding the term $\sum_{j=1}^M r_{n,j}|\widehat{\beta}_{j,L} - \beta_j|$ to both sides of this inequality yields, on $\mathcal{A}$,

$$\begin{aligned}&\|\widehat{f}_L - f\|_n^2 + \sum_{j=1}^M r_{n,j}|\widehat{\beta}_{j,L} - \beta_j| \\ &\qquad \le \|f_\beta - f\|_n^2 + 2\sum_{j=1}^M r_{n,j}(|\widehat{\beta}_{j,L} - \beta_j| + |\beta_j| - |\widehat{\beta}_{j,L}|).\end{aligned}$$

Now, $|\widehat{\beta}_{j,L} - \beta_j| + |\beta_j| - |\widehat{\beta}_{j,L}| = 0$ for $j \notin J(\beta)$, so that, on $\mathcal{A}$, we get (B.1).

To prove (B.2) it suffices to note that, on $\mathcal{A}$, we have

$$\left|\frac{1}{n} D^{-1/2} X^{\mathrm{T}} W\right|_\infty \le r/2. \tag{B.5}$$

Now, $\mathbf{y} = \mathbf{f} + \mathbf{w}$, and (B.2) follows from (2.3) and (B.5).

We finally prove (B.3). The necessary and sufficient condition for $\widehat{\beta}_L$ to be the Lasso solution can be written in the form

$$\begin{aligned}&\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}(y - X\widehat{\beta}_L) = r\|f_j\|_n \operatorname{sign}(\widehat{\beta}_{j,L}) \qquad \text{if } \widehat{\beta}_{j,L} \ne 0, \\ &\left|\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}(y - X\widehat{\beta}_L)\right| \le r\|f_j\|_n \qquad \text{if } \widehat{\beta}_{j,L} = 0,\end{aligned} \tag{B.6}$$

where $\mathbf{x}_{(j)}$ denotes the jth column of X, $j = 1, \dots, M$. Next, (B.5) yields that, on $\mathcal{A}$, we have

$$\left|\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}} W\right| \le r\|f_j\|_n/2, \qquad j = 1, \dots, M. \tag{B.7}$$

Combining (B.6) and (B.7), we get

$$\left|\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_L)\right| \ge r\|f_j\|_n/2 \qquad \text{if } \widehat{\beta}_{j,L} \ne 0. \tag{B.8}$$

Therefore,

$$\begin{aligned}\frac{1}{n^2}(\mathbf{f} - X\widehat{\beta}_L)^{\mathrm{T}} X X^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_L) &= \frac{1}{n^2}\sum_{j=1}^M \big(\mathbf{x}_{(j)}^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_L)\big)^2 \\ &\ge \frac{1}{n^2}\sum_{j:\widehat{\beta}_{j,L}\ne 0} \big(\mathbf{x}_{(j)}^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_L)\big)^2 \\ &= \mathcal{M}(\widehat{\beta}_L) r^2 \|f_j\|_n^2/4 \ge f_{\min}^2 \mathcal{M}(\widehat{\beta}_L) r^2/4.\end{aligned}$$

Since the matrices $X^{\mathrm{T}}X/n$ and XX^{T}/n have the same maximal eigenvalues,

$$\frac{1}{n^2}(\mathbf{f}-X\widehat{\beta}_L)^{\mathrm{T}}XX^{\mathrm{T}}(\mathbf{f}-X\widehat{\beta}_L)\le\frac{\phi_{\max}}{n}|\mathbf{f}-X\widehat{\beta}_L|_2^2=\phi_{\max}\|f-\widehat{f}_L\|_n^2,$$

and we deduce (B.3) from the last two displays. □

COROLLARY B.2. *Let the assumptions of Lemma* B.1 *be satisfied and* $\|f_j\|_n=1$, $j=1,\dots,M$. *Consider the linear regression model* $\mathbf{y}=X\beta+\mathbf{w}$. *Then, with probability at least* $1-M^{1-A^2/8}$, *we have*

$$|\boldsymbol{\delta}_{J_0^c}|_1\le 3|\boldsymbol{\delta}_{J_0}|_1,$$

where $J_0=J(\beta)$ *is the set of nonzero coefficients of* β *and* $\boldsymbol{\delta}=\widehat{\beta}_L-\beta$.

PROOF. Use the first inequality in (B.1) and the fact that $f=f_\beta$ for the linear regression model. □

LEMMA B.3. *Let* $\beta\in\mathbb{R}^M$ *satisfy the Dantzig constraint*

$$\left|\frac{1}{n}D^{-1/2}X^{\mathrm{T}}(y-X\beta)\right|_\infty\le r$$

and set $\boldsymbol{\delta}=\widehat{\beta}_D-\beta$, $J_0=J(\beta)$. *Then,*

$$|\boldsymbol{\delta}_{J_0^c}|_1\le|\boldsymbol{\delta}_{J_0}|_1. \tag{B.9}$$

Further, let the assumptions of Lemma B.1 *be satisfied with* $A>\sqrt{2}$. *Then, with probability of at least* $1-M^{1-A^2/2}$, *we have*

$$\left|\frac{1}{n}X^{\mathrm{T}}(\mathbf{f}-X\widehat{\beta}_D)\right|_\infty\le 2rf_{\max}. \tag{B.10}$$

PROOF OF LEMMA B.3. Inequality (B.9) follows immediately from the definition of Dantzig selector (cf. [7]). To prove (B.10), consider the event

$$\mathcal{B}=\left\{\left|\frac{1}{n}D^{-1/2}X^{\mathrm{T}}W\right|_\infty\le r\right\}=\bigcap_{j=1}^M\{|V_j|\le r_{n,j}\}.$$

Analogously to (B.4), $\mathbb{P}\{\mathcal{B}^c\}\le M^{1-A^2/2}$. On the other hand, $\mathbf{y}=\mathbf{f}+\mathbf{w}$, and, using the definition of Dantzig selector, it is easy to see that (B.10) is satisfied on $\mathcal{B}$. □

PROOF OF THEOREM 5.1. Set $\boldsymbol{\delta}=\widehat{\beta}_L-\widehat{\beta}_D$. We have

$$\frac{1}{n}|\mathbf{f}-X\widehat{\beta}_L|_2^2=\frac{1}{n}|\mathbf{f}-X\widehat{\beta}_D|_2^2-\frac{2}{n}\boldsymbol{\delta}^{\mathrm{T}}X^{\mathrm{T}}(\mathbf{f}-X\widehat{\beta}_D)+\frac{1}{n}|X\boldsymbol{\delta}|_2^2.$$

This and (B.10) yield

$$\begin{aligned}\|\widehat{f}_D - f\|_n^2 &\le \|\widehat{f}_L - f\|_n^2 + 2|\boldsymbol{\delta}|_1 \left|\frac{1}{n} X^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_D)\right|_\infty - \frac{1}{n}|X\boldsymbol{\delta}|_2^2 \\ &\le \|\widehat{f}_L - f\|_n^2 + 4 f_{\max} r |\boldsymbol{\delta}|_1 - \frac{1}{n}|X\boldsymbol{\delta}|_2^2,\end{aligned} \tag{B.11}$$

where the last inequality holds with probability at least $1 - M^{1-A^2/2}$. Since the Lasso solution $\widehat{\beta}_L$ satisfies the Dantzig constraint, we can apply Lemma B.3 with $\beta = \widehat{\beta}_L$, which yields

$$|\boldsymbol{\delta}_{J_0^c}|_1 \le |\boldsymbol{\delta}_{J_0}|_1 \tag{B.12}$$

with $J_0 = J(\widehat{\beta}_L)$. By Assumption RE$(s, 1)$, we get

$$\frac{1}{\sqrt{n}}|X\boldsymbol{\delta}|_2 \ge \kappa |\boldsymbol{\delta}_{J_0}|_2, \tag{B.13}$$

where $\kappa = \kappa(s, 1)$. Using (B.12) and (B.13), we obtain

$$|\boldsymbol{\delta}|_1 \le 2|\boldsymbol{\delta}_{J_0}|_1 \le 2\mathcal{M}^{1/2}(\widehat{\beta}_L)|\boldsymbol{\delta}_{J_0}|_2 \le \frac{2\mathcal{M}^{1/2}(\widehat{\beta}_L)}{\kappa\sqrt{n}}|X\boldsymbol{\delta}|_2. \tag{B.14}$$

Finally, from (B.11) and (B.14), we get that, with probability at least $1 - M^{1-A^2/2}$,

$$\begin{aligned}\|\widehat{f}_D - f\|_n^2 &\le \|\widehat{f}_L - f\|_n^2 + \frac{8 f_{\max} r \mathcal{M}^{1/2}(\widehat{\beta}_L)}{\kappa\sqrt{n}}|X\boldsymbol{\delta}|_2 - \frac{1}{n}|X\boldsymbol{\delta}|_2^2 \\ &\le \|\widehat{f}_L - f\|_n^2 + \frac{16 f_{\max}^2 r^2 \mathcal{M}(\widehat{\beta}_L)}{\kappa^2},\end{aligned} \tag{B.15}$$

where the RHS follows (B.2), (B.10) and another application of (B.14). This proves one side of the inequality.

To show the other side of the bound on the difference, we act as in (B.11), up to the inversion of roles of $\widehat{\beta}_L$ and $\widehat{\beta}_D$, and we use (B.2). This yields that, with probability at least $1 - M^{1-A^2/8}$,

$$\begin{aligned}\|\widehat{f}_L - f\|_n^2 &\le \|\widehat{f}_D - f\|_n^2 + 2|\boldsymbol{\delta}|_1 \left|\frac{1}{n} X^{\mathrm{T}}(\mathbf{f} - X\widehat{\beta}_L)\right|_\infty - \frac{1}{n}|X\boldsymbol{\delta}|_2^2 \\ &\le \|\widehat{f}_D - f\|_n^2 + 3 f_{\max} r |\boldsymbol{\delta}|_1 - \frac{1}{n}|X\boldsymbol{\delta}|_2^2.\end{aligned} \tag{B.16}$$

This is analogous to (B.11). Now, paralleling the proof leading to (B.15), we obtain

$$\|\widehat{f}_L - f\|_n^2 \le \|\widehat{f}_D - f\|_n^2 + \frac{9 f_{\max}^2 r^2 \mathcal{M}(\widehat{\beta}_L)}{\kappa^2}. \tag{B.17}$$

The theorem now follows from (B.15) and (B.17). □

PROOF OF THEOREM 5.2. Set, again, $\delta = \widehat{\beta}_L - \widehat{\beta}_D$. We apply (B.1) with $\beta = \widehat{\beta}_D$, which yields that, with probability at least $1 - M^{1-A^2/8}$,

$$|\delta|_1 \le 4|\delta_{J_0}|_1 + \|\widehat{f}_D - f\|_n^2/r, \tag{B.18}$$

where, now, $J_0 = J(\widehat{\beta}_D)$. Consider the following two cases: (i) $\|\widehat{f}_D - f\|_n^2 > 2r|\delta_{J_0}|_1$ and (ii) $\|\widehat{f}_D - f\|_n^2 \le 2r|\delta_{J_0}|_1$. In case (i), inequality (B.16) with $f_{\max} = 1$ immediately implies

$$\|\widehat{f}_L - f\|_n^2 \le 10\|\widehat{f}_D - f\|_n^2,$$

and the theorem follows. In case (ii), we get, from (B.18), that

$$|\delta|_1 \le 6|\delta_{J_0}|_1$$

and thus $|\delta_{J_0^c}|_1 \le 5|\delta_{J_0}|_1$. We can therefore apply Assumption RE$(s, 5)$, which yields, similarly to (B.14),

$$|\delta|_1 \le 6\mathcal{M}^{1/2}(\widehat{\beta}_D)|\delta_{J_0}|_2 \le \frac{6\mathcal{M}^{1/2}(\widehat{\beta}_D)}{\kappa\sqrt{n}}|X\delta|_2, \tag{B.19}$$

where $\kappa = \kappa(s, 5)$. Plugging (B.19) into (B.16) we finally get that, in case (ii),

$$\begin{aligned}\|\widehat{f}_L - f\|_n^2 &\le \|\widehat{f}_D - f\|_n^2 + \frac{18r\mathcal{M}^{1/2}(\widehat{\beta}_D)}{\kappa\sqrt{n}}|X\delta|_2 - \frac{1}{n}|X\delta|_2^2 \\ &\le \|\widehat{f}_D - f\|_n^2 + \frac{81r^2\mathcal{M}(\widehat{\beta}_D)}{\kappa^2}.\end{aligned} \tag{B.20}$$

□

PROOF OF THEOREM 6.1. Fix an arbitrary $\beta \in \mathbb{R}^M$ with $\mathcal{M}(\beta) \le s$. Set $\delta = D^{1/2}(\widehat{\beta}_L - \beta)$, $J_0 = J(\beta)$. On the event $\mathcal{A}$, we get, from the first line in (B.1), that

$$\begin{aligned}\|\widehat{f}_L - f\|_n^2 + r|\delta|_1 &\le \|f_\beta - f\|_n^2 + 4r\sum_{j\in J_0}\|f_j\|_n|\widehat{\beta}_{j,L} - \beta_j| \\ &= \|f_\beta - f\|_n^2 + 4r|\delta_{J_0}|_1,\end{aligned} \tag{B.21}$$

and from the second line in (B.1) that

$$\|\widehat{f}_L - f\|_n^2 \le \|f_\beta - f\|_n^2 + 4r\sqrt{\mathcal{M}(\beta)}|\delta_{J_0}|_2. \tag{B.22}$$

Consider, separately, the cases where

$$4r|\delta_{J_0}|_1 \le \varepsilon\|f_\beta - f\|_n^2 \tag{B.23}$$

and

$$\varepsilon\|f_\beta - f\|_n^2 < 4r|\delta_{J_0}|_1. \tag{B.24}$$

In case (B.23), the result of the theorem trivially follows from (B.21). So, we will only consider the case (B.24). All of the subsequent inequalities are valid on the

event $\mathcal{A} \cap \mathcal{A}_1$, where $\mathcal{A}_1$ is defined by (B.24). On this event, we get, from (B.21), that

$$|\boldsymbol{\delta}|_1 \le 4(1 + 1/\varepsilon)|\boldsymbol{\delta}_{J_0}|_1,$$

which implies $|\boldsymbol{\delta}_{J_0^c}|_1 \le (3 + 4/\varepsilon)|\boldsymbol{\delta}_{J_0}|_1$. We now use Assumption RE$(s, 3 + 4/\varepsilon)$. This yields

$$\begin{aligned}\kappa^2|\boldsymbol{\delta}_{J_0}|_2^2 &\le \frac{1}{n}|X\boldsymbol{\delta}|_2^2 = \frac{1}{n}(\widehat{\beta}_K - \beta)^{\mathrm{T}} D^{1/2} X^{\mathrm{T}} X D^{1/2}(\widehat{\beta}_L - \beta) \\ &\le \frac{f_{\max}^2}{n}(\widehat{\beta}_L - \beta)^{\mathrm{T}} X^{\mathrm{T}} X(\widehat{\beta}_L - \beta) = f_{\max}^2 \|\widehat{f}_L - f_\beta\|_n^2,\end{aligned}$$

where $\kappa = \kappa(s, 3 + 4/\varepsilon)$. Combining this with (B.22), we find

$$\begin{aligned}\|\widehat{f}_L - f\|_n^2 &\le \|f_\beta - f\|_n^2 + 4r f_{\max}\kappa^{-1}\sqrt{\mathcal{M}(\beta)}\,\|\widehat{f}_L - f_\beta\|_n \\ &\le \|f_\beta - f\|_n^2 + 4r f_{\max}\kappa^{-1}\sqrt{\mathcal{M}(\beta)}(\|\widehat{f}_L - f\|_n + \|f_\beta - f\|_n).\end{aligned}$$

This inequality is of the same form as (A.4) in [4]. A standard decoupling argument as in [4], using inequality $2xy \le x^2/b + by^2$ with $b > 1$, $x = r\kappa^{-1}\sqrt{\mathcal{M}(\beta)}$ and y being either $\|\widehat{f}_L - f\|_n$ or $\|f_\beta - f\|_n$, yields that

$$\|\widehat{f}_L - f\|_n^2 \le \frac{b+1}{b-1}\|f_\beta - f\|_n^2 + \frac{8b^2 f_{\max}^2}{(b-1)\kappa^2} r^2 \mathcal{M}(\beta) \qquad \forall b > 1.$$

Taking $b = 1 + 2/\varepsilon$ in the last display finishes the proof of the theorem. $\square$

PROOF OF PROPOSITION 6.3. Due to the weak sparsity assumption, there exists $\bar{\beta} \in \mathbb{R}^M$ with $\mathcal{M}(\bar{\beta}) \le s$ such that $\|f_{\bar{\beta}} - f\|_n^2 \le C_0 f_{\max}^2 r^2 \kappa^{-2} \mathcal{M}(\bar{\beta})$, where $\kappa = \kappa(s, 3 + 4/\varepsilon)$ is the same as in Theorem 6.1. Using this together with Theorem 6.1 and (B.3), we obtain that, with probability at least $1 - M^{1-A^2/8}$,

$$\mathcal{M}(\widehat{\beta}_L) \le C_1(\varepsilon)\mathcal{M}(\bar{\beta}) \le C_1(\varepsilon)s.$$

This and Theorem 5.1 imply

$$\|\widehat{f}_D - f\|_n^2 \le \|\widehat{f}_L - f\|_n^2 + \frac{16C_1(\varepsilon) f_{\max}^2 A^2 \sigma^2}{\kappa_0^2}\left(\frac{s \log M}{n}\right),$$

where $\kappa_0 = \kappa(\max(C_1(\varepsilon), 1)s, 3 + 4/\varepsilon)$. Once Again, applying Theorem 6.1, we get the result. $\square$

PROOF OF THEOREM 7.1. Set $\boldsymbol{\delta} = \widehat{\beta}_D - \beta^*$ and $J_0 = J(\beta^*)$. Using Lemma B.3 with $\beta = \beta^*$, we get that, on the event $\mathcal{B}$ (i.e., with probability at least

$1 - M^{1-A^2/2}$), the following are true: (i) $\frac{1}{n}|X^{\mathrm{T}}X\boldsymbol{\delta}|_\infty \le 2r$, and (ii) inequality (4.1) holds with $c_0 = 1$. Therefore, on $\mathcal{B}$ we have

$$\begin{aligned}\frac{1}{n}|X\boldsymbol{\delta}|_2^2 &= \frac{1}{n}\boldsymbol{\delta}^{\mathrm{T}}X^{\mathrm{T}}X\boldsymbol{\delta}\\ &\le \frac{1}{n}|X^{\mathrm{T}}X\boldsymbol{\delta}|_\infty|\boldsymbol{\delta}|_1\\ &\le 2r(|\boldsymbol{\delta}_{J_0}|_1 + |\boldsymbol{\delta}_{J_0^c}|_1) \qquad \text{(B.25)}\\ &\le 2(1+c_0)r|\boldsymbol{\delta}_{J_0}|_1\\ &\le 2(1+c_0)r\sqrt{s}|\boldsymbol{\delta}_{J_0}|_2 = 4r\sqrt{s}|\boldsymbol{\delta}_{J_0}|_2\end{aligned}$$

since $c_0 = 1$. From Assumption RE$(s, 1)$, we get that

$$\frac{1}{n}|X\boldsymbol{\delta}|_2^2 \ge \kappa^2|\boldsymbol{\delta}_{J_0}|_2^2,$$

where $\kappa = \kappa(s, 1)$. This and (B.25) yield that, on $\mathcal{B}$,

$$\frac{1}{n}|X\boldsymbol{\delta}|_2^2 \le 16r^2s/\kappa^2, \qquad |\boldsymbol{\delta}_{J_0}|_2 \le 4r\sqrt{s}/\kappa^2. \qquad \text{(B.26)}$$

The first inequality in (B.26) implies (7.5). Next, (7.4) is straightforward in view of the second inequality in (B.26) and of the relations (with $c_0 = 1$)

$$|\boldsymbol{\delta}|_1 = |\boldsymbol{\delta}_{J_0}|_1 + |\boldsymbol{\delta}_{J_0^c}|_1 \le (1+c_0)|\boldsymbol{\delta}_{J_0}|_1 \le (1+c_0)\sqrt{s}|\boldsymbol{\delta}_{J_0}|_2 \qquad \text{(B.27)}$$

that hold on $\mathcal{B}$. It remains to prove (7.6). It is easy to see that the kth largest in absolute value element of $\boldsymbol{\delta}_{J_0^c}$ satisfies $|\boldsymbol{\delta}_{J_0^c}|_{(k)} \le |\boldsymbol{\delta}_{J_0^c}|_1/k$. Thus,

$$|\boldsymbol{\delta}_{J_{01}^c}|_2^2 \le |\boldsymbol{\delta}_{J_0^c}|_1^2 \sum_{k \ge m+1} \frac{1}{k^2} \le \frac{1}{m}|\boldsymbol{\delta}_{J_0^c}|_1^2,$$

and, since (4.1) holds on $\mathcal{B}$ (with $c_0 = 1$), we find

$$|\boldsymbol{\delta}_{J_{01}^c}|_2 \le \frac{c_0|\boldsymbol{\delta}_{J_0}|_1}{\sqrt{m}} \le c_0|\boldsymbol{\delta}_{J_0}|_2\sqrt{\frac{s}{m}} \le c_0|\boldsymbol{\delta}_{J_{01}}|_2\sqrt{\frac{s}{m}}.$$

Therefore, on $\mathcal{B}$,

$$|\boldsymbol{\delta}|_2 \le \left(1 + c_0\sqrt{\frac{s}{m}}\right)|\boldsymbol{\delta}_{J_{01}}|_2. \qquad \text{(B.28)}$$

On the other hand, it follows from (B.25) that

$$\frac{1}{n}|X\boldsymbol{\delta}|_2^2 \le 4r\sqrt{s}|\boldsymbol{\delta}_{J_{01}}|_2.$$

Combining this inequality with Assumption RE$(s, m, 1)$, we obtain that, on $\mathcal{B}$,

$$|\boldsymbol{\delta}_{J_{01}}|_2 \le 4r\sqrt{s}/\kappa^2.$$

Recalling that $c_0 = 1$ and applying the last inequality together with (B.28), we get

$$|\boldsymbol{\delta}|_2^2 \le 16\left(1 + c_0\sqrt{\frac{s}{m}}\right)^2 \left(r\sqrt{s}/\kappa^2\right)^2. \tag{B.29}$$

It remains to note that (7.6) is a direct consequence of (7.4) and (B.29). This follows from the fact that inequalities $\sum_{j=1}^M a_j \le b_1$ and $\sum_{j=1}^M a_j^2 \le b_2$ with $a_j \ge 0$ imply

$$\sum_{j=1}^M a_j^p = \sum_{j=1}^M a_j^{2-p} a_j^{2p-2} \le \left(\sum_{j=1}^M a_j\right)^{2-p} \left(\sum_{j=1}^M a_j^2\right)^{p-1}$$
$$\le b_1^{2-p} b_2^{p-1} \qquad \forall 1 < p \le 2. \qquad \square$$

PROOF OF THEOREM 7.2. Set $\boldsymbol{\delta} = \widehat{\beta}_L - \beta^*$ and $J_0 = J(\beta^*)$. Using (B.1), where we put $\beta = \beta^*$, $r_{n,j} \equiv r$ and $\|f_\beta - f\|_n = 0$, we get that, on the event $\mathcal{A}$,

$$\frac{1}{n}|X\boldsymbol{\delta}|_2^2 \le 4r\sqrt{s}|\boldsymbol{\delta}_{J_0}|_2 \tag{B.30}$$

and (4.1) holds with $c_0 = 3$ on the same event. Thus, by Assumption RE$(s, 3)$ and the last inequality, we obtain that, on $\mathcal{A}$,

$$\frac{1}{n}|X\boldsymbol{\delta}|_2^2 \le 16r^2 s/\kappa^2, \qquad |\boldsymbol{\delta}_{J_0}|_2 \le 4r\sqrt{s}/\kappa^2, \tag{B.31}$$

where $\kappa = \kappa(s, 3)$. The first inequality here coincides with (7.8). Next, (7.9) follows immediately from (B.3) and (7.8). To show (7.7), it suffices to note that on the event $\mathcal{A}$ the relations (B.27) hold with $c_0 = 3$, to apply the second inequality in (B.31) and to use (B.4).

Finally, the proof of (7.10) follows exactly the same lines as that of (7.6). The only difference is that one should set $c_0 = 3$ in (B.28) and (B.29), as well as in the display preceding (B.28). $\square$

REFERENCES

[1] BICKEL, P. J. (2007). Discussion of "The Dantzig selector: Statistical estimation when p is much larger than n," by E. Candes and T. Tao. *Ann. Statist.* **35** 2352–2357. MR2382645

[2] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2004). Aggregation for regression learning. Preprint LPMA, Univ. Paris 6–Paris 7, n° 948. Available at arXiv:math.ST/0410214 and at https://hal.ccsd.cnrs.fr/ccsd-00003205.

[3] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2006). Aggregation and sparsity via ℓ_1 penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006)* (G. Lugosi and H. U. Simon, eds.). *Lecture Notes in Artificial Intelligence* **4005** 379–391. Springer, Berlin. MR2280619

[4] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101

[5] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Statist.* **1** 169–194. MR2312149

[6] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Sparse density estimation with ℓ_1 penalties. In *Proceedings of 20th Annual Conference on Learning Theory (COLT 2007)* (N. H. Bshouty and C. Gentile, eds.). *Lecture Notes in Artificial Intelligence* **4539** 530–543. Springer, Berlin. MR2397610

[7] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[8] DONOHO, D. L., ELAD, M. and TEMLYAKOV, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52** 6–18. MR2237332

[9] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–451. MR2060166

[10] FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1** 302–332. MR2415737

[11] FU, W. and KNIGHT, K. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

[12] GREENSHTEIN, E. and RITOV, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli* **10** 971–988. MR2108039

[13] JUDITSKY, A. and NEMIROVSKI, A. (2000). Functional aggregation for nonparametric estimation. *Ann. Statist.* **28** 681–712. MR1792783

[14] KOLTCHINSKII, V. (2006). Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.* To appear.

[15] KOLTCHINSKII, V. (2007). Dantzig selector and sparsity oracle inequalities. Unpublished manuscript.

[16] MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The Group Lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B* **70** 53–71.

[17] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[18] MEINSHAUSEN, N. and YU, B. (2006). Lasso type recovery of sparse representations for high dimensional data. *Ann. Statist.* To appear.

[19] NEMIROVSKI, A. (2000). Topics in nonparametric statistics. In *Ecole d'Eté de Probabilités de Saint-Flour XXVIII—1998. Lecture Notes in Math.* **1738**. Springer, New York. MR1775640

[20] OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A (2000a). On the Lasso and its dual. *J. Comput. Graph. Statist.* **9** 319–337. MR1822089

[21] OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A (2000b). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20** 389–404.

[22] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[23] TSYBAKOV, A. B. (2006). Discussion of "Regularization in Statistics," by P. Bickel and B. Li. *TEST* **15** 303–310. MR2273731

[24] TURLACH, B. A. (2005). On algorithms for solving least squares problems under an L1 penalty or an L1 constraint. In *2004 Proceedings of the American Statistical Association, Statistical Computing Section [CD-ROM]* 2572–2577. Amer. Statist. Assoc., Alexandria, VA.

[25] VAN DE GEER, S. A. (2008). High dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614–645. MR2396809

[26] ZHANG, C.-H. and HUANG, J. (2008). Model-selection consistency of the Lasso in high-dimensional regression. *Ann. Statist.* **36** 1567–1594. MR2435448

[27] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

P. J. BICKEL
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA AT BERKELEY
CALIFORNIA
USA
E-MAIL: bickel@stat.berkeley.edu

Y. RITOV
DEPARTMENT OF STATISTICS
FACULTY OF SOCIAL SCIENCES
THE HEBREW UNIVERSITY
JERUSALEM 91904
ISRAEL
E-MAIL: yaacov.ritov@gmail.com

A. B. TSYBAKOV
LABORATOIRE DE STATISTIQUE
CREST
3, AVENUE PIERRE LAROUSSE,
92240 MALAKOFF
AND
LPMA (UMR CNRS 1599)
UNIVERSITÈ PARIS VI
4, PLACE JUSSIEU,
75252 PARIS, CEDEX 05
FRANCE
E-MAIL: alexandre.tsybakov@upmc.fr