arXiv:1303.5817v3 [math.ST] 18 Apr 2013

ASSUMPTIONLESS CONSISTENCY OF THE LASSO

SOURAV CHATTERJEE

Abstract. The Lasso is a popular statistical tool invented by Robert Tibshirani for linear regression when the number of covariates is greater than or comparable to the number of observations. The validity of the Lasso procedure has been theoretically established under a variety of complicated-looking assumptions by various authors. This article shows that for the loss function considered in Tibshirani's original paper, the Lasso is consistent under almost no assumptions at all.

1. Introduction

The Lasso is a penalized regression procedure introduced by Tibshirani [26] in 1996. Given response variables $y_1, \ldots, y_n$ and p-dimensional covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the Lasso fits the linear regression model

$$\mathbb{E}(y_i \mid \mathbf{x}_i) = \boldsymbol{\beta} \cdot \mathbf{x}_i$$

by minimizing the ℓ^1 penalized squared error

$$\sum_{i=1}^{n} (y_i - \boldsymbol{\beta} \cdot \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is the vector of regression parameters and λ is a penalization parameter. As λ increases, the Lasso estimates are shrunk towards zero. An interesting and useful feature of the Lasso is that it is well-defined even if p is greater than n. Not only that, often only a small fraction of the estimated β_i's turn out to be non-zero, thereby producing an effect of automatic variable selection. And thirdly, there is a fast and simple procedure for computing the Lasso estimates simultaneously for all λ using the Least Angle Regression (LARS) algorithm of Efron et. al. [14]. The success of Lasso stems from all of these factors.

There have been numerous efforts to give conditions under which the Lasso 'works'. Much of this work has its origins in the investigations of ℓ^1 penalization by David Donoho and coauthors (some of it predating Tibshirani's original paper) [9, 10, 11, 12, 13]. Major advances were made by Osborne et. al. [25], Knight and Fu [19], Fan and Li [15], Meinshausen and Bühlmann [23], Yuan and Lin [29], Zhao and Yu [32], Zou [33], Greenshtein and Ritov [17], Bunea et. al. [5, 6], Candès and Tao [7, 8], Zhang and Huang [31], Lounici [21], Bickel et. al. [2], Zhang [30], Koltchinskii [20], Wainwright

Sourav Chatterjee's research was partially supported by the NSF grant DMS-1005312.

[28], Bartlett et. al. [1], and many other authors. Indeed, it is a daunting task to compile a thorough review of the literature. Fortunately, this daunting task has been accomplished in the recent book of Bühlmann and van de Geer [4], to which we refer the reader for an extensive bibliography and a comprehensive treatment of the Lasso and its many variants.

A common feature of most of the above work is that they assume that only a small number of the true β_i's are nonzero, and then look for conditions under which this set is correctly identified with high probability by the Lasso procedure with an appropriate choice of the penalization parameter. This quest invariably leads to complicated non-degeneracy conditions on the covariance matrix of the covariates. The conditions are usually unverifiable or too artificial to hold for real data — and yet, it is known that sometimes such conditions are actually necessary for certain kinds of consistency to hold [33, 32, 24]. The article [27] can serve as a quick reference for the list of all prominent assumptions and their inter-relations.

The main point of this paper is to show that for the loss function considered by Tibshirani in [26] (the 'prediction loss'), the Lasso is consistent under almost no assumptions beyond the bare minimum required for setting up the ordinary least squares regression problem. Results that are similar in spirit (but not the same) have appeared recently in the important works of Bühlmann and van de Geer [4] and Bartlett et. al. [1]. Comparisons will be given later.

2. The setup

Suppose that $X_1, \ldots, X_p$ are (possibly dependent) random variables, and M is a constant such that

$$|X_j| \le M \tag{1}$$

almost surely for each j. Let

$$Y = \sum_{j=1}^{p} \beta_j^* X_j + \varepsilon, \tag{2}$$

where ε is independent of the X_j's and

$$\varepsilon \sim N(0, \sigma^2). \tag{3}$$

Here $\beta_1^*, \ldots, \beta_p^*$ and σ^2 are unknown constants.

Let $\mathbf{Z}$ denote the random vector $(Y, X_1, \ldots, X_p)$. Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be i.i.d. copies of $\mathbf{Z}$. We will write $\mathbf{Z}_i = (Y_i, X_{i,1}, \ldots, X_{i,p})$. The set of vectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ is our data. The conditions (1), (2), (3) and the independence of $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are all that we need to assume in this paper, besides the sparsity condition that $\sum_{j=1}^{n} |\beta_j^*|$ is not too large.

3. Prediction error

Suppose that in the vector $\mathbf{Z}$, the value of Y is unknown and our task is to predict Y using the values of $X_1, \ldots, X_p$. If the parameter vector $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)$ was known, then best predictor of Y based on $X_1, \ldots, X_p$ would be the linear combination

$$\hat{Y} := \sum_{j=1}^{p} \beta_j^* X_j.$$

However $\beta_1^*, \ldots, \beta_p^*$ are unknown, and so we need to estimate them from the data $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$. The 'mean squared prediction error' of any estimator $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)$ is defined as the expected squared error in estimating $\hat{Y}$ using $\tilde{\boldsymbol{\beta}}$, that is,

$$\mathrm{MSPE}(\tilde{\boldsymbol{\beta}}) := \mathbb{E}(\hat{Y} - \tilde{Y})^2, \tag{4}$$

where

$$\tilde{Y} := \sum_{j=1}^{p} \tilde{\beta}_j X_j.$$

Note that here $\tilde{\beta}_1, \ldots, \tilde{\beta}_p$ are computed using the data $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$, and are therefore independent of $X_1, \ldots, X_p$. By this observation it is easy to see that the prediction error may be alternatively expressed as follows. Let Σ be the covariance matrix of $(X_1, \ldots, X_p)$, and let $\|\cdot\|_\Sigma$ be the norm (or seminorm) on $\mathbb{R}^p$ induced by Σ, that is,

$$\|\mathbf{x}\|_\Sigma^2 = \mathbf{x} \cdot \Sigma \mathbf{x}.$$

With this definition, the mean squared prediction error of any estimator $\tilde{\boldsymbol{\beta}}$ may be written as

$$\mathrm{MSPE}(\tilde{\boldsymbol{\beta}}) = \mathbb{E}\|\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}\|_\Sigma^2.$$

While this alternative representation of the mean squared prediction error may make it more convenient to connect it to, say, the ℓ^2 loss, the original definition (4) is more easily interpretable and acceptable from a practical point of view.

As mentioned before, the mean squared prediction error was the measure of error considered by Tibshirani in his original paper [26] and also previously by Breiman [3] in the paper that served as the main inspiration for the invention of the Lasso (see [26]). Although this gives reasonable justification for proving theorems about the prediction error of the Lasso, this measure of error is certainly not the last word in judging the effectiveness of a regression procedure. Indeed, as Tibshirani [26] remarks, "There are two reasons why the data analyst is often not satisfied with the OLS [Ordinary Least Squares] estimates. The first is *prediction accuracy* .... The second is *interpretation.*" Proving that the Lasso has a small prediction error will take care of the first concern, but not the second.

4. Prediction consistency of the Lasso

Take any $K > 0$ and define the estimator $\tilde{\boldsymbol{\beta}}^K = (\tilde{\beta}_1^K, \ldots, \tilde{\beta}_p^K)$ as the minimizer of

$$\sum_{i=1}^n (Y_i - \beta_1 X_{i,1} - \cdots - \beta_p X_{i,p})^2$$

subject to the constraint

$$\sum_{j=1}^p |\beta_j| \le K.$$

If there are multiple minimizers, choose one according to some predefined rule. While this definition of the Lasso is not the same as the one given in Section 1, this is in fact the original formulation introduced by Tibshirani in [26]. The two definitions may be shown to equivalent under a simple correspondence between K and λ, although the correspondence involves some participation of the data.

The following theorem shows that the Lasso estimator defined above is 'prediction consistent' if K is correctly chosen and $n \gg \log p$. This is the main result of this paper.

Theorem 1. *Consider the setup defined in Section* 2*. Let K be any constant such that*

$$\sum_{j=1}^p |\beta_j^*| \le K. \tag{5}$$

Let MSPE *stand for the mean squared prediction error, defined in Section* 3*. If $\tilde{\boldsymbol{\beta}}^K$ is the Lasso estimator defined above, then*

$$\mathrm{MSPE}(\tilde{\boldsymbol{\beta}}^K) \le KM\sigma\sqrt{\frac{2\log(2p)}{n}} + 8K^2M^2\sqrt{\frac{2\log(2p^2)}{n}}.$$

Remarks. (1) Close cousins of Theorem 1 have appeared very recently in the literature. The two closest results are possibly Corollary 6.1 of Bühlmann and van de Geer [4] and Theorem 1.2 of Bartlett et. al. [1]. However, these results do not actually give bounds on the mean squared prediction error defined in Section 3. Indeed, to the author's knowledge, Theorem 1 is the only result till date that gives a bound on the prediction error used by Tibshirani [26] and Breiman [3]. The results of [4] and [1] are more closely related to the notion of 'persistence' defined in Greenshtein and Ritov [17]. See also Foygel and Srebro [16] and Massart and Meynet [22] for some other related results.

(2) The explicit clean bound in terms of K, M, σ, n and p is a new contribution of Theorem 1.

(3) Suppose that a given value of K is used to compute the estimate $\tilde{\boldsymbol{\beta}}^K$. If the true parameter vector $\boldsymbol{\beta}^*$ does not obey the condition (5), we cannot

hope that $\tilde{\boldsymbol{\beta}}^K$ will be a good estimate of $\boldsymbol{\beta}^*$. There does not seem to be a way to avoid the condition (5).

(4) Theorem 1 does not give a prescription for choosing an appropriate K. But that is a separate problem. One may use, for instance, one of the approaches outlined in [26] to choose a value of K. If K is chosen based on the data, the error bound has to be recomputed to incorporate this knowledge. Theorem 1 can serve as a starting point for such a computation.

(5) In most papers on the Lasso, it is assumed that all but a small number of the β_j^*'s are zero. Theorem 1 makes no such assumption.

(6) If K, M and σ remain bounded as n and p tend to infinity, the only condition required for prediction consistency of the Lasso as given by Theorem 1 is that n grows faster than $\log p$. This condition occurs in most modern treatments of the Lasso. The $\log p$ factor arises due to the Gaussian error assumption. Actually, the assumption of Gaussianity is not strictly required; Gaussian tail is enough. A different assumption about the error would lead to a different factor.

(7) The uniform boundedness of the covariates is not strictly necessary, because M may be allowed to grow slowly with n and p. Similarly, if M remains fixed then K can also grow with n and p, as long it grows slower than $(n/\log p)^{1/4}$.

(8) Theorem 1 may be used to get error bounds for other loss functions under additional assumptions. For example, if we assume that the smallest eigenvalue of Σ is bounded below by some number λ, then the inequality

$$\|\tilde{\boldsymbol{\beta}}^K - \boldsymbol{\beta}^*\|^2 \le \lambda^{-1}\|\tilde{\boldsymbol{\beta}}^K - \boldsymbol{\beta}^*\|_\Sigma^2,$$

together with Theorem 1 gives a bound on the ℓ^2 error. Similarly, assuming that $\boldsymbol{\beta}^*$ has only a small number of nonzero entries may allow us to derive stronger conclusions from Theorem 1.

5. Estimated prediction error

Instead of the prediction error defined in Section 3, one may alternatively consider the 'estimated mean squared prediction error' of an estimator $\tilde{\boldsymbol{\beta}}$, defined as

$$\widehat{\mathrm{MSPE}}(\tilde{\boldsymbol{\beta}}) := \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - \tilde{Y}_i)^2,$$

where

$$\hat{Y}_i = \sum_{j=1}^{p}\beta_j^* X_{i,j} \quad \text{and} \quad \tilde{Y}_i = \sum_{j=1}^{p}\tilde{\beta}_j X_{i,j}.$$

Alternatively, this may be expressed as

$$\widehat{\mathrm{MSPE}}(\tilde{\boldsymbol{\beta}}) = \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\hat{\Sigma}}^2$$

where

$$\|\mathbf{x}\|_{\hat{\Sigma}}^2 = \mathbf{x} \cdot \hat{\Sigma}\mathbf{x},$$

and $\hat{\Sigma}$ is the sample covariance matrix of the covariates, that is, the matrix whose $(j,k)$the element is

$$\frac{1}{n}\sum_{i=1}^n X_{i,j}X_{i,k}.$$

The slight advantage of working with the estimated mean squared prediction error over the actual mean squared prediction error is that consistency in the estimated error holds if K grows slower than $(n/\log p)^{1/2}$, rather than $(n/\log p)^{1/4}$ as demanded by the mean squared prediction error. This is made precise in the following theorem.

Theorem 2. *Let all notation be as in Theorem* 1*, and suppose that* (5) *holds. Let* $\widehat{\mathrm{MSPE}}$ *denote the estimated mean squared prediction error, as defined above. Then*

$$\mathbb{E}\bigl(\widehat{\mathrm{MSPE}}(\tilde{\boldsymbol{\beta}}^K)\bigr) \le KM\sigma\sqrt{\frac{2\log(2p)}{n}}.$$

Incidentally, the above theorem is related to the notion of persistence defined in [17] and thoroughly investigated in [1]. Corollary 6.1 of [4] and Theorem 3.1 of [22] are other closely related results.

6. Proofs of Theorems 1 and 2

Let $\mathbf{Y} := (Y_1, \ldots, Y_n)$, and $\tilde{\mathbf{Y}}^K := (\tilde{Y}_1^K, \ldots, \tilde{Y}_n^K)$, where

$$\tilde{Y}_i^K := \sum_{j=1}^p \tilde{\beta}_j^K X_{i,j}.$$

Similarly, let

$$\tilde{Y}^K := \sum_{j=1}^p \tilde{\beta}_j^K X_j.$$

For each $1 \le j \le p$, let $\mathbf{X}_j := (X_{1,j}, \ldots, X_{n,j})$. Finally, let $\hat{\mathbf{Y}} := (\hat{Y}_1, \ldots, \hat{Y}_n)$, where

$$\hat{Y}_i := \sum_{j=1}^p \beta_j^* X_{i,j},$$

Given $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$, define the set

$$C := \{\beta_1\mathbf{X}_1 + \cdots + \beta_p\mathbf{X}_p : |\beta_1| + \cdots + |\beta_p| \le K\}.$$

Note that C is a compact convex subset of $\mathbb{R}^n$. By definition, $\tilde{\mathbf{Y}}^K$ is the projection of $\mathbf{Y}$ on to the set C. Since C is convex, it follows that for any $\mathbf{x} \in C$, the vector $\mathbf{x} - \tilde{\mathbf{Y}}^K$ must be at an obtuse angle to the vector $\mathbf{Y} - \tilde{\mathbf{Y}}^K$. That is,

$$(\mathbf{x} - \tilde{\mathbf{Y}}^K) \cdot (\mathbf{Y} - \tilde{\mathbf{Y}}^K) \le 0.$$

The condition (5) ensures that $\hat{\mathbf{Y}} \in C$. Therefore

$$(\hat{\mathbf{Y}} - \tilde{\mathbf{Y}}^K) \cdot (\mathbf{Y} - \tilde{\mathbf{Y}}^K) \le 0.$$

This may be written as

$$\begin{aligned}\|\hat{\mathbf{Y}} - \tilde{\mathbf{Y}}^K\|^2 &\le (\mathbf{Y} - \hat{\mathbf{Y}}) \cdot (\tilde{\mathbf{Y}}^K - \hat{\mathbf{Y}}) \\ &= \sum_{i=1}^n \varepsilon_i \Big(\sum_{j=1}^p (\tilde{\beta}_j^K - \beta_j^*) X_{i,j} \Big) \\ &= \sum_{j=1}^p (\tilde{\beta}_j^K - \beta_j^*) \Big(\sum_{i=1}^n \varepsilon_i X_{i,j} \Big).\end{aligned}$$

By the condition (5) and the definition of $\tilde{\boldsymbol{\beta}}^K$, the above inequality implies that

$$\|\hat{\mathbf{Y}} - \tilde{\mathbf{Y}}^K\|^2 \le 2K \max_{1 \le j \le p} |U_j|, \tag{6}$$

where

$$U_j := \sum_{i=1}^n \varepsilon_i X_{i,j}.$$

Let $\mathcal{F}$ be the sigma algebra generated by $(X_{i,j})_{1 \le i \le n,\, 1 \le j \le p}$. Let $\mathbb{E}^{\mathcal{F}}$ denote the conditional expectation given $\mathcal{F}$. Conditional on $\mathcal{F}$,

$$U_j \sim N\Big(0,\, \sigma^2 \sum_{i=1}^n X_{i,j}^2\Big).$$

Since $|X_{i,j}| \le M$ almost surely for all i, j, it follows from the standard results about Gaussian random variables (see Lemma 3 in the Appendix) that

$$\mathbb{E}^{\mathcal{F}}(\max_{1 \le j \le p} |U_j|) \le M\sigma\sqrt{2n \log(2p)}.$$

Since the right hand side is non-random, it follows that

$$\mathbb{E}(\max_{1 \le j \le p} |U_j|) \le M\sigma\sqrt{2n \log(2p)}.$$

Using this bound in (6), we get

$$\mathbb{E}\|\hat{\mathbf{Y}} - \tilde{\mathbf{Y}}^K\|^2 \le 2KM\sigma\sqrt{2n \log(2p)}. \tag{7}$$

This completes the proof of Theorem 2. For Theorem 1, we have to work a bit more. Note that by the independence of $\mathbf{Z}$ and $\tilde{\boldsymbol{\beta}}^K$,

$$\mathbb{E}^{\mathcal{F}}(\hat{Y} - \tilde{Y}^K)^2 = \sum_{j,k=1}^p (\beta_j^* - \tilde{\beta}_j^K)(\beta_k^* - \tilde{\beta}_k^K)\mathbb{E}(X_j X_k).$$

Also, we have

$$\begin{aligned}&\frac{1}{n}\|\hat{\mathbf{Y}} - \tilde{\mathbf{Y}}^K\|^2\\ &= \frac{1}{n}\sum_{i=1}^n \sum_{j,k=1}^p (\beta_j^* - \tilde{\beta}_j^K)(\beta_k^* - \tilde{\beta}_k^K) X_{i,j} X_{i,k}.\end{aligned}$$

Therefore, if we define

$$V_{j,k} := \mathbb{E}(X_j X_k) - \frac{1}{n}\sum_{i=1}^n X_{i,j} X_{i,k},$$

then

$$\begin{aligned}\mathbb{E}^{\mathcal{F}}(\hat{Y} - \tilde{Y}^K)^2 - \frac{1}{n}\|\hat{\mathbf{Y}} - \tilde{\mathbf{Y}}^K\|^2 &= \sum_{j,k=1}^p (\beta_j^* - \tilde{\beta}_j^K)(\beta_k^* - \tilde{\beta}_k^K) V_{j,k}\\ &\le 4K^2 \max_{1\le j,k\le p} |V_{j,k}|. \qquad (8)\end{aligned}$$

Since $|\mathbb{E}(X_j X_k) - X_{i,j} X_{i,k}| \le 2M^2$ for all i, j and k, it follows by Hoeffding's inequality (see Lemma 5 in the Appendix) that for any $\beta \in \mathbb{R}$,

$$\mathbb{E}(e^{\beta V_{j,k}}) \le e^{2\beta^2 M^4/n}.$$

Consequently, by Lemma 4 from the Appendix,

$$\mathbb{E}(\max_{1\le j,k\le p} |V_{j,k}|) \le 2M^2\sqrt{\frac{2\log(2p^2)}{n}}.$$

Plugging this into (8) and combining with (7) completes the proof of Theorem 1.

Appendix

The following inequality is a well-known result about the size of the maximum of Gaussian random variables.

Lemma 3. *Suppose that $\xi_i \sim N(0, \sigma_i^2)$, $i = 1, \ldots, m$. The ξ_i's need not be independent. Let $L := \max_{1\le i\le m} \sigma_i$. Then*

$$\mathbb{E}(\max_{1\le i\le m} |\xi_i|) \le L\sqrt{2\log(2m)}.$$

Proof. For any $\beta \in \mathbb{R}$, $\mathbb{E}(e^{\beta\xi_i}) = e^{\beta^2\sigma_i^2/2} \le e^{\beta^2 L^2/2}$. Thus, for any $\beta > 0$,

$$\begin{aligned}\mathbb{E}(\max_{1\le i\le m}|\xi_i|) &= \frac{1}{\beta}\mathbb{E}(\log e^{\max_{1\le i\le m}\beta|\xi_i|}) \\ &\le \frac{1}{\beta}\mathbb{E}\biggl(\log\sum_{i=1}^m (e^{\beta\xi_i}+e^{-\beta\xi_i})\biggr) \\ &\le \frac{1}{\beta}\log\sum_{i=1}^m \mathbb{E}(e^{\beta\xi_i}+e^{-\beta\xi_i}) \le \frac{\log(2m)}{\beta}+\frac{\beta L^2}{2}.\end{aligned}$$

The proof is completed by choosing $\beta = L^{-1}\sqrt{2\log(2m)}$. □

The result extends easily to the maximum of random variables with Gaussian tails.

Lemma 4. *Suppose that for $i = 1, \dots, m$, ξ_i is a random variable such that $\mathbb{E}(e^{\beta\xi_i}) \le e^{\beta^2 L^2/2}$ for each $\beta \in \mathbb{R}$, where L is some given constant. Then*

$$\mathbb{E}(\max_{1\le i\le m}|\xi_i|) \le L\sqrt{2\log(2m)}.$$

Proof. Exactly the same as the proof of Lemma 3. □

The following lemma is commonly known as Hoeffding's inequality [18]. The version we state here is slightly different than the commonly stated version. For this reason, we state the lemma together with its proof.

Lemma 5. *Suppose that $\eta_1, \dots, \eta_m$ are independent, mean zero random variables, and L is a constant such that $|\eta_i| \le L$ almost surely for each i. Then for each $\beta \in \mathbb{R}$,*

$$\mathbb{E}(e^{\beta\sum_{i=1}^m \eta_i}) \le e^{\beta^2 m L^2/2}.$$

Proof. By independence,

$$\mathbb{E}(e^{\beta\sum_{i=1}^m \eta_i}) = \prod_{i=1}^m \mathbb{E}(e^{\beta\eta_i}).$$

Therefore it suffices to prove the result for $m = 1$. Note that

$$\mathbb{E}(e^{\beta\eta_1}) = \int_{-L}^{L} e^{\beta x} d\mu_1(x),$$

where μ_1 is the law of η_1. By the convexity of the map $x \mapsto e^{\beta x}$, it follows that for each $x \in [-L, L]$,

$$e^{\beta x} = e^{\beta(tL+(1-t)(-L))} \le te^{\beta L} + (1-t)e^{-\beta L}, \tag{9}$$

where

$$t = t(x) = \frac{x}{2L} + \frac{1}{2}.$$

Since $\mathbb{E}(\eta_1) = 0$, therefore $\int t(x) d\mu_1(x) = 1/2$. Thus by (9), $\mathbb{E}(e^{\beta\eta_1}) \le \cosh(\beta L)$. The inequality $\cosh x \le e^{x^2/2}$ completes the proof. □

Acknowledgments. The author thanks Peter Bickel, Peter Bühlmann, Peter Bartlett, Mathias Drton, Gilles Blanchard and Persi Diaconis for helpful comments.

References

[1] Bartlett, P. L., Mendelson, S. and Neeman, J. (2012). ℓ^1-regularized linear regression: persistence and oracle inequalities. *Probab. Theory Related Fields,* **154** no. 1-2, 193–224.

[2] Bickel, P. J., Ritov, Y. A. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.,* **37** no. 4, 1705–1732.

[3] Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics,* **37** no. 4, 373–384.

[4] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data. Methods, theory and applications.* Springer Series in Statistics. Springer, Heidelberg.

[5] Bunea, F., Tsybakov, A. and Wegkamp, M. (2006). Sparsity oracle inequalities for the Lasso. *Electron. J. Statist.,* **1**, 169–194.

[6] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** no. 4, 1674–1697.

[7] Candes, E. and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory,* **51** no. 12, 4203-4215.

[8] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.,* **35** no. 6, 2313–2351.

[9] Donoho, D. (2004). For Most Large Underdetermined Systems of Equations, the Minimal ℓ^1-Norm Solution is the Sparsest Solution. *Comm. on Pure and Appl. Math.,* **59** no. 7, 907–934.

[10] Donoho, D. and Elad, M. (2002). Optimally Sparse Representation in General (Nonorthogonal) Dictionaries via ℓ^1-Norm Minimizations. *Proc. of National Acad. of Science USA,* **1005**, 2197–2202.

[11] Donoho, D. and Huo, X. (2002). Uncertainty Principles and Ideal Atomic Decompositions. *IEEE Transactions on Information Theory,* **47**, 2845–2863.

[12] Donoho, D. and Johnstone, I. (1994). Ideal Spatial Adaptation via Wavelet Shrinkages. *Biometrika,* **81**, 425–455.

[13] Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D. (1995). Wavelet Shrinkage: Asymptopia? *J. of the Royal Statist. Soc., Ser. B,* **57**, 301–337.

[14] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.,* **32** no. 2, 407–499.

[15] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.,* **96** no. 456, 1348–1360.

[16] Foygel, R. and Srebro, N. (2011). Fast-rate and optimistic-rate error bounds for L1-regularized regression. *Preprint.* Available at http://arxiv.org/abs/1108.0373

[17] Greenshtein, E. and Ritov, Y. A. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli,* **10** no. 6, 971–988.

[18] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.,* **58**, 13–30.

[19] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.,* **28** no. 5, 1356–1378.

[20] Koltchinskii, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli,* **15**, 799–828.

[21] Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Elec. J. Statist.,* **2**, 90–102.

[22] Massart, P. and Meynet, C. (2011). The Lasso as an ℓ_1-ball model selection procedure. *Elec. J. Statist.,* **5**, 669–687.
[23] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.,* **34** no. 3, 1436–1462.
[24] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.,* **37** no. 1, 246–270.
[25] Osborne, M. R., Presnell, B. and Turlach, B. A. (2000). On the LASSO and its dual. *J. Comput. Graph. Statist.,* **9** no. 2, 319–337.
[26] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc., Ser. B,* **58** no. 1, 267–288.
[27] van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Elec. J. Statist.,* **3**, 1360–1392.
[28] Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory,* **55**, 2183–2202.
[29] Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. *J. Royal Statist. Soc., Ser. B,* **69** no. 2, 143–161.
[30] Zhang, T. (2009). Some sharp performance bounds for least squares regression with L1 regularization. *Ann. Statist.,* **37**, 2109–2144.
[31] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.,* **36** no. 4, 1567–1594.
[32] Zhao, P. and Yu, B. (2007). On model selection consistency of Lasso. *J. Machine Learning Research,* **7** no. 2, 2541–2563.
[33] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.,* **101** no. 476, 1418–1429.

Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012
E-mail address: sourav@cims.nyu.edu