*Submitted to the Annals of Applied Statistics*
arXiv: math.PR/0000000

# THE LASSO, CORRELATED DESIGN, AND IMPROVED ORACLE INEQUALITIES*

By Sara van de Geer and Johannes Lederer

*ETH Zürich*

We study high-dimensional linear models and the $\ell_1$-penalized least squares estimator, also known as the Lasso estimator. In literature, oracle inequalities have been derived under restricted eigenvalue or compatibility conditions. In this paper, we complement this with entropy conditions which allow one to improve the dual norm bound, and demonstrate how this leads to new oracle inequalities. The new oracle inequalities show that a smaller choice for the tuning parameter and a trade-off between $\ell_1$-norms and small compatibility constants are possible. This implies, in particular for correlated design, improved bounds for the prediction error of the Lasso estimator as compared to the methods based on restricted eigenvalue or compatibility conditions only.

**1. Introduction.** We derive oracle inequalities for the Lasso estimator for various designs. Results in literature are generally based on restricted eigenvalue or compatibility conditions (see Section 3 for definitions). We refer to [2], [4], [5], [6], [8], [10], [11]. See also [3] and the references therein. In a sense, compatibility or restricted eigenvalue conditions and the so-called dual norm bound we describe below belong together. In contrast, if compatibility constants or restricted eigenvalues are very small, the design may have high correlations, and then the dual norm bound is too rough. In this paper, we discuss an approach that joins both situations. The work is a follow-up of [12]. It combines results of the latter with the parallel developments in the area based on the dual norm bound.

We consider an input space $\mathcal{X}$ and $p$ feature mappings $\psi_j : \mathcal{X} \to \mathbb{R}$, $j = 1, \ldots, p$. We let $(x_1, \ldots, x_n)^T \in \mathcal{X}^n$ be a given input vector, and $\mathbf{Y} :=$

*Research supported by SNF 20PA21-120050
*AMS 2000 subject classifications:* Primary 62J05; secondary 62J99
*Keywords and phrases:* compatibility, correlation, entropy, high-dimensional model, Lasso

$(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^T \in \mathbb{R}^n$ be an output vector, and consider the linear model

$$\mathbf{Y} = \sum_{j=1}^{p} \psi_j \beta_j^0 + \epsilon,$$

with $\epsilon \in \mathbb{R}^n$ a noise vector, and $\beta^0 \in \mathbb{R}^p$ a vector of unknown coefficients. Here, with some abuse of notation, $\psi_j$ denotes the vector $\psi_j = (\psi_j(x_1), \ldots, \psi_j(x_n))^T$. The design matrix is $\mathbf{X} := (\psi_1, \ldots, \psi_p)$ and the Gram matrix is

$$\hat{\Sigma} := \mathbf{X}^T \mathbf{X}/n.$$

Throughout, we assume that $\sum_{i=1}^n \psi_j^2(x_i) \leq n$ for all $j$.

We write a linear function with coefficients $\beta$ as $f_\beta := \sum_{j=1}^p \psi_j \beta_j$, $\beta \in \mathbb{R}^p$. The Lasso estimator is

$$\hat{\beta} := \arg\min_{\beta} \Big\{ \|\mathbf{Y} - f_\beta\|_2^2/n + \lambda \|\beta\|_1 \Big\}.$$

We denote the estimator of the regression function $f^0 := f_{\beta^0}$ by $\hat{f} := f_{\hat{\beta}}$.

Oracle results using compatibility or restricted eigenvalue conditions are based on the dual norm bound

$$\sup_{\|\beta\|_1 = 1} |\epsilon^T f_\beta|/n = \max_{1 \leq j \leq p} |\epsilon^T \psi_j|/n.$$

Let us define

$$\|f_\beta\|_n^2 := \sum_{i=1}^n f_\beta^2(x_i)/n = \beta^T \hat{\Sigma} \beta.$$

The point we make in this paper is that the dual norm bound does not take into account possible small values for $\|f_{\hat{\beta}} - f_{\beta^0}\|_n$. Our results are based on bounds for

$$\sup_{\|\beta\|_1 \leq 1,\ \|f_\beta\|_n \leq R} |\epsilon^T f_\beta|/n$$

as function of $R > 0$. We then apply these to $\hat{\beta} - \beta^0$ (or $\beta^0$ here replaced by a sparse approximation). We use an improvement of the dual norm bound, and show in Theorem 4.1 the consequences. The main observation here is that with highly correlated design, one can generally take the tuning parameter $\lambda$ of much smaller order than the usual $\sqrt{\log p/n}$. Moreover, small compatibility constants may be traded off against the $\ell_1$-norm of the coefficients of an oracle.

**2. Organization of the paper.** In Section 3, we present our notation, and the definitions of compatibility constants and restricted eigenvalues. Section 4 contains the main result, based on a pre-assumed improvement of the dual norm bound. In Section 5, we present a result from empirical process theory, which shows that the improvement of the dual norm bound used in Section 4 holds under entropy conditions on $\mathcal{F} := \{f_\beta : \ \|\beta\|_1 = 1\}$. In Section 6, we first give a geometrical interpretation of the compatibility constant and discuss the relation with eigenvalues. The next question to address is then how to read off the entropy conditions directly from the design. We show that a Gram matrix with strongly decreasing eigenvalues leads to a small entropy of $\mathcal{F}$. Alternatively, we derive an an entropy bound for $\mathcal{F}$ based on the covering number of the design $\{\psi_j\}$, a result much in the spirit of [7]. We moreover link these covering numbers with the correlation structure of the design. Section 7 concludes and Section 8 contains proofs.

## 3. Notation and definitions.

3.1. *The compatibility constant.* Let $S \subset \{1, \ldots, p\}$ be an index set with cardinality $s$. We define for all $\beta \in \mathbb{R}^p$,

$$\beta_{S,j} := \beta_j 1\{j \in S\}, \ j = 1, \ldots, p, \ \beta_{S^c} := \beta - \beta_S.$$

Below, we present for constants $L > 0$ the compatibility constant $\phi(L, S)$ introduced in [10]. For normalized $\psi_j$ (i.e., $\|\psi_j\|_n = 1$ for all $j$), one can view $1 - \phi^2(1, S)/2$ as an $\ell_1$-version of the canonical correlation between the linear space spanned by the variables in $S$ on the one hand, and the linear space of the variables in $S^c$ on the other hand. Instead of all linear combinations with normalized $\ell_2$-norm, we now consider all linear combinations with normalized $\ell_1$-norm of the coefficients. For a geometric interpretation, we refer to Section 6.

**Definition** *The* compatibility constant *is*

$$\phi^2(L, S) := \min\{s\|f_{\beta_S} - f_{\beta_{S^c}}\|_n^2 : \ \|\beta_S\|_1 = 1, \ \|\beta_{S^c}\|_1 \leq L\}.$$

The compatibility constant is closely related to (and never smaller than) the restricted eigenvalue as defined in [2], which is

$$\phi^2_{\mathrm{RE}}(L, S) = \min\left\{\frac{\|f_{\beta_S} - f_{\beta_{S^c}}\|_n^2}{\|\beta_S\|_2^2} : \|\beta_{S^c}\|_1 \leq L\|\beta_S\|_1\right\}.$$

See also [8], and see [13] for a discussion of the relation between restricted eigenvalues and compatibility.

3.2. *Projections.* As the "true" $\beta^0$ is perhaps only approximately sparse, we will consider a sparse approximation. The projection of $f^0 := f_{\beta^0}$ on the space spanned by the variables in $S$ is

$$\mathrm{f}_S := \arg\min_{f=f_{\beta_S}} \|f - f^0\|_n.$$

The coefficients of $\mathrm{f}_S$ are denoted by $b^S$, i.e.,

$$\mathrm{f}_S = f_{b^S}.$$

Note that $\mathrm{f}_S$ only has non-zero coefficients inside $S$, that is, $(b^S)_S = b^S$.

**4. Main result.** We let $\mathcal{T}_\alpha$ be the set

$$\mathcal{T}_\alpha := \Big\{ \sup_\beta \frac{4|\epsilon^T f_\beta|/n}{\|f_\beta\|_n^{1-\alpha}\|\beta\|_1^\alpha} \le \lambda_0 \Big\}.$$

Here, $0 \le \alpha \le 1$ and $\lambda_0 > 0$ are fixed constants.

Note that on $\mathcal{T}_\alpha$,

$$\sup_{\|\beta\|_1=1,\ \|f_\beta\|_n \le R} |\epsilon^T f_\beta|/n \le \lambda_0 R^{1-\alpha}/4,$$

i.e., we have a refinement of the dual norm bound described in Section 1.

Note that for fixed $\lambda_0$ and for $\alpha < \tilde{\alpha}$, it holds that $\mathcal{T}_\alpha \subset \mathcal{T}_{\tilde{\alpha}}$. This is because by the triangle inequality

$$\|f_\beta\|_n = \|\sum_j \psi_j \beta_j\|_n \le \sum_j \|\psi_j\|_n |\beta_j| \le \|\beta\|_1.$$

We want to choose $\alpha$ preferably small, yet keep the probability of the set $\mathcal{T}_\alpha$ large. For $\alpha = 1$, one has

$$\mathcal{T}_1 = \Big\{ \max_{1 \le j \le p} 4|\epsilon^T \psi_j|/n \le \lambda_0 \Big\},$$

by the dual norm bound. Thus, e.g. when $\epsilon \sim \mathcal{N}(0, I)$, the probability $\mathbb{P}(\mathcal{T}_1)$ of $\mathcal{T}_1$ is large when $\lambda_0 \asymp \sqrt{\log p/n}$. We detail in Section 5 how one can lowerbound $\mathbb{P}(\mathcal{T}_\alpha)$ for a proper value of $\alpha$ depending on the design $\{\psi_j\}$.

Generally, the value for $\lambda_0$ will be of order $\sqrt{\log p/n}$, as in the case $\alpha = 1$, or $\lambda \asymp \sqrt{\log n/n}$ or even $\lambda_0 \asymp 1/\sqrt{n}$.

The choice of the tuning parameter $\lambda$ depends on $\lambda_0$. The following technical lemma will be used:

LEMMA 4.1. *Let $0 \leq \alpha \leq 1$ and let $a$, $b$ and $\lambda_0$ be positive numbers. Then*

$$\lambda_0 a^{1-\alpha} b^\alpha \leq \frac{1}{2}a^2 + \lambda b + \frac{1}{2}\left(\frac{\lambda_0}{\lambda^\alpha}\right)^{\frac{2}{1-\alpha}}.$$

*Here, when $\alpha = 1$,*

$$\left(\frac{\lambda_0}{\lambda^\alpha}\right)^{\frac{2}{1-\alpha}} = \left(\frac{\lambda_0}{\lambda}\right)^\infty := \begin{cases} \infty & \lambda < \lambda_0 \\ 1 & \lambda = \lambda_0 \\ 0 & \lambda > \lambda_0 \end{cases}.$$

In the proof of the main result, Theorem 4.1, we invoke Lemma 4.1 to handle the "noise part" $\epsilon^T f_\beta$ with $\beta = \hat{\beta} - \beta^0$ (or actually with $\beta^0$ replaced here by a sparse approximation). On $\mathcal{T}_\alpha$, it holds that

$$4|\epsilon^T f_\beta|/n \leq \frac{1}{2}\|f_\beta\|_n^2 + \lambda\|\beta\|_1 + \frac{1}{2}\left(\frac{\lambda_0}{\lambda^\alpha}\right)^{\frac{2}{1-\alpha}},$$

uniformly in $\beta \in \mathbb{R}^p$. In the right hand side of this inequality, the first term $\|f_\beta\|_n^2/2$ can be incorporated in the risk and the second term $\lambda\|\beta\|_1$ will be overruled by the penalty. Finally, the third term $(\lambda_0/\lambda^\alpha)^{\frac{2}{1-\alpha}}/2$ governs the choice of the tuning parameter $\lambda$.

We now come to the main result. We formulate it for an arbitrary index set $S$ partitioned in sets $S_1$ and $S_2$ in an arbitrary way. We will elaborate on the choice of $S$ in Remarks 4.2 and 4.5. Corollaries 4.1 and 4.2 take for a given $S$ some special choices for the tuning parameter $\lambda$ and for the partition of $S$ into $S_1$ and $S_2$.

Recall that $\mathrm{f}_S$ is the projection of $f^0 = f_{\beta^0}$ and $b^S$ are the coefficients of $\mathrm{f}_S$.

THEOREM 4.1. *Let $S$ be an arbitrary index set, partitioned into two sets $S_1$ and $S_2$, i.e. $S = S_1 \cup S_2$, $S_1 \cap S_2 = \emptyset$. Let $s_1$ be the cardinality of $S_1$. Let $\mathcal{T}_\alpha$ be the set*

$$\mathcal{T}_\alpha := \left\{\sup_\beta \frac{4|\epsilon^T f_\beta|/n}{\|f_\beta\|_n^{1-\alpha}\|\beta\|_1^\alpha} \leq \lambda_0\right\}.$$

*Then on $\mathcal{T}_\alpha$,*

$$\|\hat{f}-f^0\|_n^2+\lambda\|\hat{\beta}-b^S\|_1 \leq \frac{56\lambda^2 s_1}{\phi^2(6,S_1)}+\frac{28}{3}\lambda\|(b^S)_{S_2}\|_1+\frac{7}{6}\left(\frac{\lambda_0}{\lambda^\alpha}\right)^{\frac{2}{1-\alpha}}+7\|\mathrm{f}_S-f^0\|_n^2.$$

REMARK 4.1. *We did not attempt to optimize the constants we provided in Theorem 4.1.*

REMARK 4.2. *Given a value of the tuning parameter $\lambda$, we can now define the estimation error using the variables in $S$ as*

$$\mathcal{E}(S) := \min_{S_1\subset S,\ S_2=S\setminus S_1} \frac{8\lambda^2 s_1}{\phi^2(6,S_1)}+\frac{4}{3}\lambda\|(b^S)_{S_2}\|_1.$$

*The oracle set $S_*$ is then the set which trades off estimation error and approximation error, i.e, the set $S_*$ that minimizes*

$$\mathcal{E}(S)+\|\mathrm{f}_S-f^0\|_n^2.$$

*Note that $S_*$ depends on $\lambda$, say $S_*=S_*(\lambda)$. The best value for the tuning parameter $\lambda^*$ is then obtained by minimizing*

$$\mathcal{E}(S_*(\lambda))+\frac{1}{6}\left(\frac{\lambda_0}{\lambda^\alpha}\right)^{\frac{2}{1-\alpha}}+\|\mathrm{f}_{S_*(\lambda)}-f^0\|_n^2.$$

REMARK 4.3. *In practice, the tuning parameter $\lambda$ can be chosen by cross-validation. As this method tries to mimic minimization of the prediction error, it can be conjectured that one then arrives at rates at least a good as the ones we discuss here choosing values of $\lambda$ depending on the design, the (unknown) error distribution, and the unknown sparsity. This is however not rigorously proven.*

REMARK 4.4. *We have restricted ourselves to improvements of the dual norm bound of the form given by sets $\mathcal{T}_\alpha$. The situation can be generalized by considering sets of the form*

$$\left\{\sup_\beta \frac{4|\epsilon^T f_\beta|/n}{G^{-1}(\|f_\beta\|_n/\|\beta\|_1)\|\beta\|_1}\leq\lambda_0\right\},$$

*where $G$ is a given increasing convex function with $G(0)=0$.*

COROLLARY 4.1.
*a) If we take $S_2=\emptyset$, we have $S_1=S$, and $s_1=|S|=:s$. This is a good*

*choice when the compatibility constants are large for all subsets of $S$. With the choice*

$$\lambda^2 \asymp \lambda_0^2 \Big( \frac{\phi^2(6,S)}{s} \Big)^{1-\alpha},$$

*we get on $\mathcal{T}_\alpha$,*

$$\|\hat{f} - f^0\|_n^2 + \lambda \|\hat{\beta} - b^S\|_1 = \mathcal{O}\left( \lambda_0^2 \Big( \frac{s}{\phi^2(6,S)} \Big)^{\alpha} + \|\mathrm{f}_S - f^0\|_n^2 \right).$$

*Recall that the dual norm bound has $\alpha = 1$. With $\lambda_0 \asymp \sqrt{\log p/n}$ we then arrive at the "usual" oracle inequality as provided by, among others, [2], [4], [5], [6], [8] [10], [11]. When $\alpha < 1$, the compatibility constant may be very small, as the design is highly correlated. The effect is however somewhat tempered by the power $\alpha$ in the bound.*
*b) More generally, let*

$$\lambda^2 \asymp \lambda_0^2 \Big( \frac{\phi^2(6,S_1)}{s_1} \Big)^{1-\alpha},$$

*Then on $\mathcal{T}_\alpha$,*

$$\|\hat{f} - f^0\|_n^2 + \lambda \|\hat{\beta} - b^S\|_1$$
$$= \mathcal{O}\left( \lambda_0^2 \Big( \frac{s_1}{\phi^2(6,S_1)} \Big)^{\alpha} + \lambda_0 \Big( \frac{\phi^2(6,S_1)}{s_1} \Big)^{\frac{1-\alpha}{2}} \|(b^S)_{S_2}\|_1 + \|\mathrm{f}_S - f^0\|_n^2 \right).$$

Corollary 4.2.
*a) With the choice $S_1 = \emptyset$, the result does not involve the compatibility constant. This may be desirable when the design is highly correlated. The result then corresponds to what is sometimes called "slow rates", although we will see that when $\alpha < 1$, the rates can still be much faster than $1/\sqrt{n}$. When $\alpha = 1$, we must take $\lambda > \lambda_0$ (due to the term $(\lambda_0/\lambda^\alpha)^{\frac{2}{1-\alpha}}$). When $\alpha < 1$, we choose*

$$\lambda \asymp \lambda_0^{\frac{2}{1+\alpha}} \|b^S\|_1^{-\frac{1-\alpha}{1+\alpha}}.$$

*We get on $\mathcal{T}_\alpha$,*

$$\|\hat{f} - f^0\|_n^2 + \lambda \|\hat{\beta} - b^S\|_1 = \mathcal{O}\left( \lambda_0^{\frac{2}{1+\alpha}} \|b^S\|_1^{\frac{2\alpha}{1+\alpha}} + \|\mathrm{f}_S - f^0\|_n^2 \right).$$

*b) More generally, let*

$$\lambda \asymp \lambda_0^{\frac{2}{1+\alpha}} \|(b^S)_{S_2}\|_1^{-\frac{1-\alpha}{1+\alpha}}.$$

*Then on* $\mathcal{T}_\alpha$,

$$\|\hat{f} - f^0\|_n^2 + \lambda\|\hat{\beta} - b^S\|_1$$
$$= \mathcal{O}\left(\frac{\lambda_0^{\frac{4}{1+\alpha}}}{\|(b^S)_{S_2}\|_1^{\frac{2(1-\alpha)}{1+\alpha}}}\frac{s_1}{\phi^2(6,S_1)} + \lambda_0^{\frac{2}{1+\alpha}}\|(b^S)_{S_2}\|_1^{\frac{2\alpha}{1+\alpha}} + \|\mathrm{f}_S - f^0\|_n^2\right).$$

REMARK 4.5. *Note that by taking* $S_1$ *smaller, the value of* $s_1/\phi^2(6,S_1)$ *will not increase, but on the other hand, the value of* $\|(b^S)_{S_2}\|_1$ *will become larger. Thus, the best rate will emerge if we trade off these two effects. Indeed, suppose that for some* $S_1$

$$\lambda_0^{\frac{2}{1+\alpha}}\frac{s_1}{\phi^2(S_1)} \asymp \|(b^S)_{S_2}\|_1^{\frac{2}{1+\alpha}}.$$

*Then on* $\mathcal{T}_\alpha$, *for*

$$\lambda \asymp \lambda_0\left(\frac{s_1^2}{\phi^2(6,S_1)}\right)^{(1-\alpha)/2} \asymp \lambda_0^{\frac{2}{1+\alpha}}\|(b^S)_{S_2}\|_1^{-\frac{1-\alpha}{1+\alpha}},$$

*we have*

$$\|\hat{f} - f^0\|_n^2 + \lambda\|\hat{\beta} - b^S\|_1 = \mathcal{O}\left(\lambda_0^2\left(\frac{s_1}{\phi^2(6,S_1)}\right)^{\alpha} + \|\mathrm{f}_S - f^0\|_n^2\right)$$
$$= \mathcal{O}\left(\lambda_0^{\frac{2}{1+\alpha}}\|(b^S)_{S_2}\|_1^{\frac{2\alpha}{1+\alpha}} + \|\mathrm{f}_S - f^0\|_n^2\right).$$

*In particular for the case* $\alpha < 1$, *it is however not clear when such a trade-off is possible. It may well be that for any* $S_1$, $s_1/\phi^2(6,S_1)$ *either heavily dominates or is heavily dominated by the* $\ell_1$*-part* $\|(b^S)_{S_2}\|_1$. *See Section 6 for a further discussion.*

**5. Improving the dual norm bound.** In this section, we provide probability bounds for the set $\mathcal{T}_\alpha$ introduced in Section 4. The results follow from empirical process theory, see e.g. and [14] and [15]. Theorem 5.1 is taken from [3].

**Definition** *Let* $\mathcal{F}$ *be a class of real-valued functions on* $\mathcal{X}$. *Endow* $\mathcal{F}$ *with norm* $\|\cdot\|_n$. *Let* $\delta > 0$ *be some radius. A* $\delta$-packing set *is a set of functions in* $\mathcal{F}$ *that are each at least* $\delta$ *apart. A* $\delta$-covering set *is a set of functions* $\{\phi_1, \ldots, \phi_N\}$, *such that*

$$\sup_{f\in\mathcal{F}}\min_{k=1,\ldots,N}\|f - \phi_k\|_n \leq \delta.$$

*The* $\delta$-covering number $N(\delta, \mathcal{F}, \|\cdot\|_n)$ *of* $\mathcal{F}$ *is the minimum size of a* $\delta$-*covering set. The* entropy *of* $\mathcal{F}$ *is* $H(\cdot, \mathcal{F}, \|\cdot\|_n) = \log N(\cdot, \mathcal{F}, \|\cdot\|_n)$.

It is easy to see that $N(\delta, \mathcal{F}, \|\cdot\|_n)$ can be bounded by the size of a maximal $\delta$-packing set.

We assume the errors are sub-Gaussian, that is, for some positive constants $K$ and $\sigma_0$,

$$K^2(\mathbb{E}\exp[\epsilon_i^2/K^2] - 1) \leq \sigma_0^2, \ i = 1, \ldots, n. \tag{5.1}$$

The following theorem is Corollary 14.6 in [3]. It is in the spirit of a weighted concentration inequality, and uses the notation

$$x_+ := \max\{x, 0\}.$$

THEOREM 5.1. *Assume (5.1). Let* $\mathcal{F}$ *be a class of functions with* $\|f\|_n \leq 1$ *for all* $f \in \mathcal{F}$*, and with, for some* $0 < \alpha < 1$ *and some constant* $A$*,*

$$\log\Big(1 + 2N(\delta, \mathcal{F}, \|\cdot\|_n)\Big) \leq \left(\frac{A}{\delta}\right)^{2\alpha}, \ 0 < \delta \leq 1.$$

*Define*

$$B := \exp\left[\frac{A^{2\alpha}\alpha}{2(2^{1-\alpha} - 1)^2}\right] - 1,$$

*and*

$$K_0 := 3 \times 2^5 \sqrt{K^2 + \sigma_0^2}.$$

*It holds that*

$$\mathbb{E}\exp\left[\sup_{f \in \mathcal{F}}\left[\left(\frac{|\epsilon^T f|/\sqrt{n}}{\|f\|_n K_0} - \frac{A^\alpha \|f\|_n^{-\alpha}}{2^{1-\alpha} - 1}\right)_+\right]^2\right] \leq 1 + 2/B.$$

COROLLARY 5.1. *Assume the conditions of Theorem 5.1. Chebyshev's inequality shows that for all* $t > 0$*,*

$$\mathbb{P}\Big(\exists\ f \in \mathcal{F} : \ |\epsilon^T f|/\sqrt{n} \geq K_0 A^\alpha \|f\|_n^{1-\alpha}(2^{1-\alpha} - 1)^{-1} + K_0 \|f\|_n t\Big)$$

$$\leq \exp[-t^2](1 + 2/B).$$

COROLLARY 5.2. *Consider now linear functions*

$$f_\beta := \sum_{j=1}^p \psi_j \beta_j, \ \beta \in \mathbb{R}^p,$$

*where* $\|\psi_j\|_n \leq 1$. *Then*

$$\|f_\beta\|_n \leq \|\beta\|_1.$$

*Hence,* $\{f_\beta/\|\beta\|_1 : \beta \in \mathbb{R}^p\}$ *is a class of functions with* $\|\cdot\|_n$*-norm bounded by* 1. *Suppose now*

$$\log\Big(1 + 2N(\delta, \{f_\beta : \ \|\beta\|_1 = 1\}, \|\cdot\|_n)\Big) \leq \left(\frac{A}{\delta}\right)^{2\alpha}, \ 0 < \delta \leq 1.$$

*Under the sub-Gaussianity condition (5.1), we then have for all* $t > 0$ *and for*

$$\lambda_0 = \frac{4K_0}{\sqrt{n}}\left(\frac{A^\alpha}{2^{1-\alpha}-1} + t\right),$$

*the lower bound*

$$\mathbb{P}(\mathcal{T}_\alpha) \geq 1 - \exp[-t^2](1 + 2/B).$$

**6. Compatibility, eigenvalues, entropy and correlations.** We study the set

$$\mathcal{F} := \{f_\beta : \|\beta\|_1 = 1\}.$$

It is considered as subset of $L_2(Q_n)$, where $Q_n := \sum_{i=1}^n \delta_{x_i}/n$. The $L_2(Q_n)$-norm is $\|\cdot\|_n$.

6.1. *Geometric interpretation of the compatibility constant.* We first look at the minimal $\ell_1$-eigenvalue

$$\Lambda^2_{\min,1}(S) := \min\left\{s\beta_S^T \hat{\Sigma}\beta_S : \ \|\beta_S\|_1 = 1\right\}$$

as introduced in [3]. Note that $\Lambda_{\min,1}(S)/\sqrt{s}$ is the minimal distance between any point $f_{\beta_S}$ with $\|\beta_S\|_1 = 1$ and the point $\{0\}$. We tacitly assume that the $\{\psi_j\}_{j\in S}$ are linearly independent. The set $\{f_{\beta_S} : \ \|\beta_S\|_1 = 1\}$ is then an $\ell_1$-version of a sphere: it is the boundary of the convex hull of $\{\psi_j\}_{j\in S} \cup \{-\psi_j\}_{j\in S}$ in $s$-dimensional space with $\{0\}$ in its "center". It is a parallelogram when $s = 2$ (see Figure 1) and then a rectangle when the $\psi_j$, $j \in S$, have equal length.

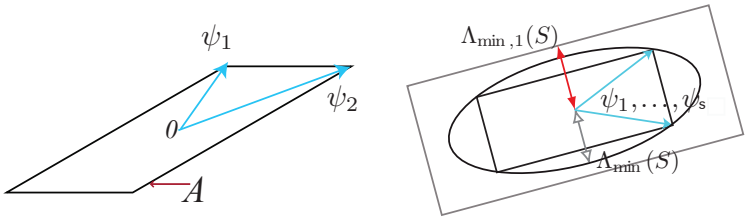


FIG 1. *Left panel: the set* $A = \{f_{\beta_S} : \|\beta_S\|_1 = 1\}$. *Right panel:* $\ell_1$- *and* $\ell_2$-*eigenvalues.*

Let $\hat{\Sigma}_S$ be the Gram matrix of the variables in $S$ and $\Lambda^2_{\min}(S)$ be the minimal ($\ell_2$-)eigenvalue of the matrix $\hat{\Sigma}_S$:

$$\Lambda^2_{\min}(S) := \min\Big\{\beta_S^T \hat{\Sigma} \beta_S : \ \|\beta_S\|_2 = 1\Big\}.$$

Then

$$\Lambda^2_{\min,1}(S) \geq \Lambda^2_{\min}(S) \geq \Lambda^2_{\min,1}(S)/s,$$

One can construct examples where $\Lambda^2_{\min}(S)$ is as small as $3/(s-2)$ ( $s > 2$) and $\Lambda^2_{\min,1}(S)$ is at least $1/2$ (see [13]), that is, they can differ by the maximal amount $s$ in order of magnitude. See also Figure 1 which is to be understood as representing a case $s > 2$. Thus, minimal $\ell_1$-eigenvalues can be much larger than minimal ($\ell_2$-)eigenvalues. The normalized compatibility constant $\phi(L,S)/\sqrt{s}$ is the minimal distance between the sets $A := \{f_{\beta_S} : \ \|\beta_S\|_1 = 1\}$ and $B := \{f_{\beta_{S^c}} : \ \|\beta_{S^c}\|_1 \leq L\}$, that is,

$$\frac{\phi(L,S)}{\sqrt{s}} = \min\Big\{\|a-b\|_n : \ a \in A, \ b \in B\Big\}.$$

See Figure 2 for an impression of the situation. Observe that $A$ is the boundary of the convex hull of $\{+\psi_j\}_{j\in S} \cup \{-\psi_j\}_{j\in S}$, and $B$ is the convex hull of $\{+\psi_j\}_{j\in S^c} \cup \{-\psi_j\}_{j\in S^c}$ including its interior, blown up with a factor $L$ (typically, the $\{\psi_j\}_{j\in S^c}$ form a linearly dependent system in $\mathbb{R}^n$). Furthermore, since $\{0\} \in B$

$$\phi(L,S) \leq \Lambda_{\min,1}(S).$$

This shows that when $\ell_1$-eigenvalues are small, the compatibility constant is necessarily also small. Small $\ell_2$-eigenvalues may have less of this effect.

6.2. *Eigenvalues and entropy.* We now let

$$\hat{\Sigma} = E\Omega^2 E^T$$

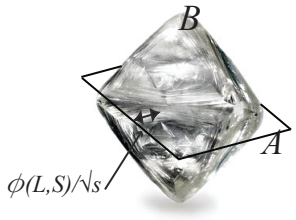


FIG 2. *The compatibility constant*

be the spectral decomposition of the Gram matrix $\hat{\Sigma}$, $E$ being the matrix of eigenvectors, ($E^T E = EE^T = I$) and $\Omega^2 = \text{diag}(\omega_1^2, \cdots, \omega_p^2)$ the matrix of ($\ell_2$-)eigenvalues. We assume they are in decreasing order: $\omega_1^2 \geq \cdots \geq \omega_p^2$.

LEMMA 6.1. *Suppose that for some strictly decreasing function $V$*

$$\omega_{j+1}^2 \leq V^2(j), \ j = 1, \ldots, p.$$

*Then for all $\delta > 0$,*

$$H(2\delta, \{f_\beta : \ \|\beta\|_1 = 1\}, \|\cdot\|_n) \leq V^{-1}(\delta) \log\left(\frac{3}{\delta}\right).$$

EXAMPLE 6.1. *Suppose that for some positive constants $m$ and $C$*

$$\omega_j \leq \frac{C}{j^m}, \ j = 1, \ldots, p.$$

*Then by Lemma 6.1,*

$$H(2\delta, \{f_\beta : \ \|\beta\|_1 = 1\}, \|\cdot\|_n) \leq \left(\frac{C}{\delta}\right)^{\frac{1}{m}} \log\left(\frac{3}{\delta}\right).$$

*For $\delta \geq 1/n$ (say) we therefore have*

$$H(\delta, \{f_\beta : \ \|\beta\|_1 = 1\}, \|\cdot\|_n) \leq \left(\frac{2C}{\delta}\right)^{\frac{1}{m}} \log(6n).$$

*When $m > 1/2$, one can use a minor generalization of Corollary 5.2, where the entropy bound is only required for values of $\delta > 1/n$. One then takes*

$$\alpha = \frac{1}{2m}, \ A = (C_m^2 \log(n))^{\frac{1}{2\alpha}},$$

*where $C_m$ is a constant depending on $m$ and $C$. Then the value of $\lambda_0$ defined there becomes*

$$\lambda_0 = \frac{4K_0}{\sqrt{n}}\left(\frac{C_m\sqrt{\log(n)}}{2^{1-\frac{1}{2m}}-1} + t\right)$$

*which is for fixed $m$ and $K_0$, and a fixed (large) $t$, of order $\sqrt{\log n/n}$.*

6.3. *Entropy based on coverings of $\{\psi_j\}$.* We can consider $\{f_\beta : \|\beta\|_1 = 1\}$ as a subset of

$$\operatorname{conv}(\{\pm\psi_j\}),$$

where $\{\pm\psi_j\} := \{\psi_j\} \cup \{-\psi_j\}$, and $\operatorname{conv}(\{\pm\psi_j\})$ is its convex hull. Infact, if the $\{\psi_j\}$ form a linearly dependent system in $\mathbb{R}^n$, $\mathcal{F}$ is exactly equal to $\operatorname{conv}(\{\pm\psi_j\})$.

The paper [7] gives a bound for the entropy of a convex hull for the case where the $u$-covering number of the extreme points is a polynomial in $1/u$. This result can also be found in [9]. There is a redundant log-term in these entropy bounds, see [1] and [15], but removing this log-term may result in very large constants, depending on the dimension $W$ as given in Example 6.2 (see [3] for some explicit constants). This means that when the dimension $W$ of the extreme points is large (growing with $n$ say), the simple bound with log-term we provide below in Lemma 6.2 may be better than the more involved ones.

We give a bound for the entropy of $\mathcal{F}$ by balancing the $u$-covering number of $\{\psi_j\}$ and the squared radius $u^2$. The result is as in [9], with only new element its extension to general covering numbers (i.e., not only polynomial ones). Lemma 6.2 and its proof can be found in [3].

LEMMA 6.2. *Let*

$$N(u) := N(u, \{\psi_j\}, \|\cdot\|_n), \ u > 0.$$

*We have*

$$H\Big(\delta, \{f_\beta : \ \|\beta\|_1 = 1\}, \|\cdot\|_n\Big)$$

$$\leq \min_{0<u<1} 6\left(N(u) + \frac{6u^2}{\delta^2}\right)\log\left(2\left(\frac{8+\delta}{\delta}\right)N(\delta)\right).$$

EXAMPLE 6.2. *In this example, we assume the $u$-covering numbers of $\{\psi_j\}$ are bounded by a polynomial in $u$. That is, we suppose that for some positive*

*constants $W$ and $C$,*

$$N(u, \{\psi_j\}, \|\cdot\|_n) \leq \left(\frac{C}{u}\right)^W, u > 0.$$

*The constant $W$ can be thought of as the dimension of $\{\psi_j\}$. By Lemma 6.2, we can choose*

$$\alpha = \frac{W}{2+W}, \ A = (C_W^2 \log(n))^{\frac{1}{2\alpha}},$$

*and we get, as in Example 6.1,*

$$\lambda_0 = \frac{4K_0}{\sqrt{n}}\left(\frac{C_W\sqrt{\log(n)}}{2^{\frac{2}{2+W}}-1} + t\right).$$

A refined analysis of the relation between compatibility constants, covering numbers and entropy is still to be carried out. We confine ourselves here to the following, rather trivial, observation (without proof).

LEMMA 6.3. *Consider normalized design: $\|\psi_j\|_n = 1 \ \forall \ j$. Let $\{\psi_{j_1}, \ldots, \psi_{j_N}\}$ be a maximal $u$-packing set of $\{\psi_j\}$. Then for any $S \supset \{j_1, \ldots, j_N\}$, $S \neq \{1, \ldots, p\}$, and any $L \geq 1$,*

$$\phi^2(L, S) \leq su^2.$$

One may argue that as $u$-packing sets are approximations of the original design $\{\psi_j\}$ with fewer covariables, they are good candidates for the sparsity set $S_1$ used in Theorem 4.1. Lemma 6.3 however shows that such sparsity sets will have very small compatibility constants.

6.4. *Decorrelation numbers.* Decorrelation numbers are closely related to packing numbers. First, define the inner product

$$\rho(\phi, \tilde{\phi}) := \phi^T \tilde{\phi}/n.$$

Note that $\Sigma_{j,k} = \rho(\psi_j, \psi_k)$ and that in the case of standardized design (i.e. $\sum_{i=1}^n \psi_j(x_i) = 0$ and $\|\psi_j\|_n = 1 \ \forall \ j$), the inner product $\rho(\psi_j, \psi_k)$ is for $j \neq k$ the (empirical) correlation between $\psi_j$ and $\psi_k$.

**Definition** *For $\rho > 0$, the* $\rho$-decorrelation number $M(\rho)$ *is the largest value of $M$ such that there exists $\{\phi_1, \ldots, \phi_M\} \subset \{\pm\psi_j\}$ with $|\rho(\phi_j, \phi_k)| < \rho$ for all $j \neq k$.*

Hence, if the $\rho$-decorrelation number is small, then there are many large correlations, i.e., then the design is highly correlated.

It is clear that when $\|\psi_j\|_n = \|\psi_k\|_n = 1$, it holds that

$$\|\psi_j - \psi_k\|_n^2 = 2(1 - \rho(\psi_j, \psi_k)).$$

In other words, small correlations correspond to covariables that are near to each other. This can be translated into covering number as shown in Lemma 6.4. Its proof is straightforward and omitted.

LEMMA 6.4. *Consider normalized design:* $\|\psi_j\|_n = 1 \ \forall \ j$. *For all* $0 < u < 1$,

$$N(\sqrt{2}u, \{\pm\psi_j\}, \|\cdot\|_n) \leq M(1 - u^2).$$

**7. Conclusion.** We have combined results for the prediction error of the Lasso with both compatibility conditions and entropy conditions. Small entropies of $\{f_\beta : \|\beta\|_1 = 1\}$ correspond to highly correlated design and possibly to small compatibility constants. Our analysis shows that small entropies allow for a smaller choice of the tuning parameter and possibly for a compensation of small compatibility constants. This means that the Lasso enjoys good prediction error properties, even in the case where the design is highly correlated.

**8. Proofs.**

**Proof of Lemma 4.1.** We use that for positive $u$ and $v$ and for $\mathrm{p} \geq 1$, $\mathrm{q} \geq 1$, $1/\mathrm{p} + 1/\mathrm{q} = 1$, the conjugate inequality

$$uv \leq u^{\mathrm{p}}/\mathrm{p} + v^{\mathrm{q}}/\mathrm{q}$$

holds. Taking $\mathrm{p} = 1/(1-\alpha)$ and replacing $u$ by $u^{1-\alpha}$ gives

$$u^{1-\alpha}v \leq \frac{1-\alpha}{2}u^2 + \frac{1+\alpha}{2}v^{\frac{2}{1+\alpha}}.$$

With $\mathrm{p} = (1+\alpha)/(2\alpha)$, and replacing $u$ by $u^{\frac{2\alpha}{1+\alpha}}$, we get

$$u^{\frac{2\alpha}{1+\alpha}}v \leq \frac{2\alpha}{1+\alpha}u + \frac{1-\alpha}{1+\alpha}v^{\frac{1+\alpha}{1-\alpha}}.$$

Thus,

$$\lambda_0 a^{1-\alpha}b^{\alpha} \leq \frac{1-\alpha}{2}a^2 + \frac{1+\alpha}{2}\left(\lambda_0 b^{\alpha}\right)^{\frac{2}{1+\alpha}}$$

$$\leq \frac{a^2}{2} + \frac{1+\alpha}{2}(\lambda b)^{\frac{2\alpha}{1+\alpha}} \left(\frac{\lambda_0}{\lambda^\alpha}\right)^{\frac{2}{1+\alpha}}$$

$$\leq \frac{a^2}{2} + \frac{1+\alpha}{2}\left(\frac{2\alpha}{1+\alpha}\lambda b + \frac{1-\alpha}{1+\alpha}\left(\frac{\lambda_0}{\lambda^\alpha}\right)^{\frac{2}{1-\alpha}}\right)$$

$$\leq \frac{a^2}{2} + \lambda b + \frac{1}{2}\left(\frac{\lambda_0}{\lambda^\alpha}\right)^{\frac{2}{1-\alpha}}.$$

□

**Proof of Theorem 4.1.** Since

$$\|\mathbf{Y} - \hat{f}\|_2^2/n + \lambda\|\hat{\beta}\|_1 \leq \|\mathbf{Y} - \mathrm{f}_S\|_2^2/n + \lambda\|b^S\|_1,$$

we have the Basic Inequality

$$\|\hat{f} - f^0\|_n^2 + \lambda\|\hat{\beta}\|_1 \leq 2\epsilon^T(\hat{f} - \mathrm{f}_S)/n + \lambda\|b^S\|_1 + \|\mathrm{f}_S - f^0\|_n^2.$$

Hence, on $\mathcal{T}_\alpha$,

$$\|\hat{f} - f^0\|_n^2 + \lambda\|\hat{\beta}\|_1 \leq \lambda_0\|\hat{f} - \mathrm{f}_S\|_n^{1-\alpha}\|\hat{\beta} - b^S\|_1^\alpha/2 + \lambda\|b^S\|_1 + \|\mathrm{f}_S - f^0\|_n^2.$$

Apply Lemma 4.1 to find

$$\|\hat{f} - f^0\|_n^2 + \lambda\|\hat{\beta}\|_1$$

$$\leq \frac{1}{4}\|\hat{f} - \mathrm{f}_S\|_n^2 + \frac{1}{2}\lambda\|\hat{\beta} - b^S\|_1 + \frac{1}{4}\left(\frac{\lambda_0}{\lambda^\alpha}\right)^{\frac{2}{1-\alpha}} + \lambda\|b^S\|_1 + \|\mathrm{f}_S - f^0\|_n^2.$$

$$\leq \frac{1}{2}\|\hat{f} - f^0\|_n^2 + \frac{1}{2}\lambda\|\hat{\beta} - b^S\|_1 + \frac{1}{4}\left(\frac{\lambda_0}{\lambda^\alpha}\right)^{\frac{2}{1-\alpha}} + \lambda\|b^S\|_1 + \frac{3}{2}\|\mathrm{f}_S - f^0\|_n^2.$$

Thus, we get on $\mathcal{T}_\alpha$,

$$\|\hat{f} - f^0\|_n^2 + 2\lambda\|\hat{\beta}\|_1 \leq \lambda\|\hat{\beta} - b^S\|_1 + 2\lambda\|b^S\|_1 + \frac{1}{2}\left(\frac{\lambda_0}{\lambda^\alpha}\right)^{\frac{2}{1-\alpha}} + 3\|\mathrm{f}_S - f^0\|_n^2.$$

Defining $S_3 := S^c$, we rewrite this to

$$\|\hat{f} - f^0\|_n^2 + 2\lambda\|\hat{\beta}_{S_2 \cup S_3}\|_1$$

$$\leq \lambda\|\hat{\beta}_{S_1} - (b^S)_{S_1}\|_1 + \lambda\|\hat{\beta}_{S_2} - (b^S)_{S_2}\|_1 + \lambda\|\hat{\beta}_{S_3}\|_1 + 2\lambda\|(b^S)_{S_1}\|_1 - 2\lambda\|\hat{\beta}_{S_1}\|_1$$

$$+2\lambda\|(b^S)_{S_2}\|_1 + \frac{1}{2}\left(\frac{\lambda_0}{\lambda^\alpha}\right)^{\frac{2}{1-\alpha}} + 3\|\mathrm{f}_S - f^0\|_n^2$$

$$\leq 3\lambda\|\hat{\beta}_{S_1} - (b^S)_{S_1}\|_1 + \lambda\|\hat{\beta}_{S_2\cup S_3}\|_1 + 3\lambda\|(b^S)_{S_2}\|_1 + \frac{1}{2}\Big(\frac{\lambda_0}{\lambda^\alpha}\Big)^{\frac{2}{1-\alpha}} + 3\|\mathrm{f}_S - f^0\|_n^2.$$

Moving the term $\lambda\|\hat{\beta}_{S_2\cup S_3}\|_1$ to the left hand side, and applying a triangle inequality, we obtain

$$\|\hat{f} - f^0\|_n^2 + \lambda\|\hat{\beta}_{S_2\cup S_3} - (b^S)_{S_2}\|_1$$

$$\leq \underbrace{3\lambda\|\hat{\beta}_{S_1} - (b^S)_{S_1}\|_1}_{:=I} + \underbrace{4\lambda\|(b^S)_{S_2}\|_1 + \frac{1}{2}\Big(\frac{\lambda_0}{\lambda^\alpha}\Big)^{\frac{2}{1-\alpha}} + 3\|\mathrm{f}_S - f^0\|_n^2}_{:=II}.$$

**Case i.** If $I \geq II$, we arrive at

$$\|\hat{f} - f^0\|_n^2 + \lambda\|\hat{\beta}_{S_2\cup S_3} - (b^S)_{S_2}\|_1 \leq 6\lambda\|\hat{\beta}_{S_1} - (b^S)_{S_1}\|_1.$$

We first add add a term $\lambda\|\hat{\beta}_{S_1} - (b^S)_{S_1}\|_1$ to the left and right hand side and then apply the compatibility condition to $\hat{\beta} - b^S$, to get

$$\|\hat{f} - f^0\|_n^2 + \lambda\|\hat{\beta} - b^S\|_1 \leq \frac{7\lambda\sqrt{s_1}}{\phi(6, S_1)}\|\hat{f} - \mathrm{f}_S\|_n$$

$$\leq \frac{1}{2}\|\hat{f} - f^0\|_n^2 + \frac{7}{2}\|\mathrm{f}_S - f_0\|_n^2 + \frac{56\lambda^2 s_1}{2\phi^2(6, S_1)}.$$

Here we used the decoupling device

$$2xy \leq bx^2 + y^2/b \;\forall\; x, y \in R,\; b > 0.$$

So then

$$\|\hat{f} - f^0\|_n^2 + 2\lambda\|\hat{\beta} - b^S\|_1 \leq \frac{56\lambda^2 s_1}{\phi^2(6, S_1)} + 7\|\mathrm{f}_S - f^0\|_n^2.$$

**Case ii.** If $I < II$, we get

$$\|\hat{f} - f^0\|_n^2 + \lambda\|\hat{\beta}_{S_2\cup S_3} - (b^S)_{S_2}\|_1 \leq 2II,$$

and hence

$$\|\hat{f} - f^0\|_n^2 + \lambda\|\hat{\beta} - b^S\|_1 \leq \frac{7}{3}II$$

$$= \frac{28}{3}\lambda\|(b^S)_{S_2}\|_1 + \frac{7}{6}\Big(\frac{\lambda_0}{\lambda^\alpha}\Big)^{\frac{2}{1-\alpha}} + 7\|\mathrm{f}_S - f^0\|_n^2.$$

□

**Proof of Lemma 6.1.** Let $\|\beta\|_1 = 1$. Then $\|\beta\|_2 \leq 1$, and hence $\|E^T\beta\|_2 \leq 1$. For $N \leq V^{-1}(\delta)$ it holds that $\omega_{N+1} \leq \delta$ and hence

$$\sum_{j=N+1}^{p} \omega_j^2 (E^T\beta)_j^2 \leq \omega_{N+1}^2 \sum_{j=N+1}^{p} (E^T\beta)_j^2 \leq \delta^2.$$

We now note that $\|\beta\|_1 = 1$ implies $\|f_\beta\|_n \leq 1$ and hence

$$\sum_{j=1}^{N} \omega_j^2 (E^T\beta)_j^2 \leq 1.$$

Lemma 14.27 in [3] states that a ball with radius 1 in $N$-dimensional Euclidean space can be covered by $(3/\delta)^N$ balls with radius $\delta$ (see also Problem 2.1.6 in [15]). □

## References.


[1] K. Ball and A. Pajor. The entropy of convex bodies with few extreme points. *London Mathematical Society Lecture Note Series*, 158:25–32, 1990.

[2] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

[3] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[4] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation and sparsity via $\ell_1$-penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory, COLT 2006. Lecture Notes in Artificial Intelligence 4005*, pages 379–391, Heidelberg, 2006. Springer Verlag.

[5] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35:1674–1697, 2007a.

[6] F. Bunea, A. Tsybakov, and M.H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007b.

[7] R.M. Dudley. Universal Donsker classes and metric entropy. *Annals of Probability*, 15:1306–1326, 1987.

[8] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 45:7–57, 2009.

[9] D. Pollard. *Empirical Processes: Theory and Applications*. IMS Lecture Notes, 1990.

[10] S. van de Geer. The deterministic Lasso. *The JSM Proceedings*, 2007.

[11] S. van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36:614–645, 2008.

[12] S. van de Geer. On non-asymptotic bounds for estimation in generalized linear models with highly correlated design. In *Asymptotics: Particles, Processes and Inverse Problems (E.A. Cator, G. Jongbloed, C. Kraaikamp, H.P. Lopuhaä, J.A. Wellner eds.)*, volume 55, pages 121–134. IMS Lecture Notes Monograph Series, 2007.

[13] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, pages 1360–1392, 2009.

[14] S. A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge, 2000.

[15] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3.

Seminar for Statistics
ETH Zürich
Rämistrasse 101
8092 Zürich
E-mail: geer@stat.math.ethz.ch
lederer@stat.math.ethz.ch