The deterministic Lasso

Sara van de Geer
Seminar für Statistik, ETH Zürich

Abstract

We study high-dimensional generalized linear models and empirical risk minimization using the Lasso. An oracle inequality is presented, under a so called *compatibility* condition. Our aim is three fold: to proof a result announced in van de Geer (2007), to provide a simple proof with simple constants, and to separate the stochastic problem from the deterministic one.

KEY WORDS: coherence, Lasso, oracle, sparsity

1. Introduction

We consider the Lasso for high-dimensional generalized linear models, introducing three new elements. Firstly, we complete the proof of an oracle inequality, which was announced in van de Geer (2007). The result is shown to hold under a so-called *compatibility* condition for certain norms. This condition is a weaker version of Assumption C* in van de Geer (2007), and is related to coherence conditions in e.g. Candes and Tao, and Bunea et al. (2006, 2007). Secondly, we state the result with simple constants, and give a new and rather straightforward proof. Thirdly, in Subsection 1.3, we separate the stochastic part of the problem from the deterministic part. In the subsequent sections, we then only concentrate on the deterministic part (whence the title of this paper). The advantage of this approach is that the behavior of the Lasso under various sampling schemes (e.g., dependent observations), can immediately be deduced once the stochastic problem (a probability inequality for the empirical process) is clear.

Consider data $\{Z_i\}_{i=1}^n$, with (for $i = 1, \ldots, n$) Z_i in some space $\mathcal{Z}$. Let $\mathbf{F} := (\mathbf{F}, \|\cdot\|)$ be a normed real vector space, and, for each $f \in \mathbf{F}$, $\rho_f : \mathcal{Z} \to \mathbf{R}$ be a loss function. We assume that the map $f \mapsto \rho_f(z)$ is convex for all $z \in \mathcal{Z}$. For example, in a regression setup, one has (for $i = 1, \ldots, n$) $Z_i = (X_i, Y_i)$, with covariables $X_i \in \mathcal{X}$ and response variable $Y_i \in \mathcal{Y} \subset \mathbf{R}$, and $f : \mathcal{X} \to \mathbf{R}$ is a regression function. Examples are quadratic loss,

$$\rho_f(\cdot, y) = (y - f)^2,$$

or logistic loss,

$$\rho_f(\cdot, y) = -yf + \log(1 + \exp[f]),$$

etc.

1.1 The target

We denote, for a function $\rho : \mathcal{Z} \to \mathbf{R}$, the theoretical measure by

$$P\rho := \sum_{i=1}^n E\rho(Z_i)/n,$$

and the empirical measure by

$$P_n\rho := \sum_{i=1}^n \rho(Z_i)/n.$$

Define the target

$$f^0 := \arg\min_{f \in \mathbf{F}} P\rho_f.$$

We assume for simplicity that the minimum is attained, and that it is unique.

For $f \in \mathbf{F}$, the excess risk is

$$\mathcal{E}(f) := P(\rho_f - \rho_{f^0}).$$

1.2 The Lasso

Consider a given linear subspace $\mathcal{F} := \{f_\beta : \beta \in \mathbf{R}^p\} \subset \mathbf{F}$, where $\beta \mapsto f_\beta$ is linear. The collection $\mathcal{F}$ will be the model class over which we perform empirical risk minimization. When $\mathcal{F}$ is high-dimensional, a complexity penalty can prevent overfitting. The Lasso is

$$\hat{\beta} = \arg\min_\beta \left\{ P_n\rho_{f_\beta} + \lambda\|\beta\|_1 \right\}, \tag{1}$$

where

$$\|\beta\|_1 := \sum_{j=1}^p |\beta_j|,$$

i.e., $\|\cdot\|_1$ is the ℓ^1-norm. The parameter λ is a smoothing - or tuning - parameter.

We examine the excess risk $\mathcal{E}(f_{\hat{\beta}})$ of the Lasso, and show that it behaves as if it knew which variables are relevant for a linear approximation of the target f^0. Section 2 does this under a so-called *compatibility condition*. Section 3 examines under what circumstances the compatibility condition is met.

In (1), the penalty weights all coefficients equally. In practice, certain terms (e.g., the constant term) will not be penalized, and a weighted version of the ℓ^1 norm is used, with e.g., more weight assigned to variables with large sample variance. In essence (possibly random) weights do not alter the main issues in the theory. We therefore consider the non-weighted ℓ^1 penalty to clarify these issues. More details on the case of empirical weights are in Bunea et al. (2006) and van de Geer (2007).

1.3 The empirical process

Throughout, the study of the stochastic part of the problem (involving the empirical process) is set aside. It can be handled using empirical process theory. In the current paper, we simply restrict ourselves to a set $\mathcal{S}$ where the empirical process behaves "well" (see (6) in Lemma 2.1). This allows us to proceed with purely deterministic, and relatively straightforward, arguments.

More precisely, write for $M > 0$,

$$\mathbf{Z}_M := \sup_{\|\beta-\beta^*\|_1 \leq M} |(P_n - P)(\rho_{f_\beta} - \rho_{f_{\beta^*}})|. \qquad (2)$$

Here, β^* is some fixed parameter value, which will be specified in Section 2. Our results hold on the set $\mathcal{S} = \{\mathbf{Z}_M \leq \varepsilon\}$, with M and ε appropriately chosen. The ratio ε/M will play a major role: it has to be chosen large enough so that $\mathcal{S}$ has large probability. On the other hand, ε/M will form be a lower bound for the smoothing parameter λ (see (5) in Lemma 2.1).

It is shown in van de Geer (2007), that for independent observations Z_i, under mild conditions, and with $\varepsilon/M \asymp \sqrt{\log(2p)/n}$, the set $\mathcal{S}$ has large probability in the case of Lipschitz loss functions, that is, for $f \mapsto \rho_f(z)$ is Lipschitz for all z. For other loss functions (e.g. quadratic loss), similar behavior can be obtained, but only for small enough values of M.

In the result of Lemma 2.1, one may replace P_n by any other (random) probability measure, say $\hat{P}$, in the definition of $\hat{\beta}$, provided one also adjusts the definition (2) of $\mathbf{Z}_M$. The result for the Lasso goes through on $\mathcal{S}$. One should then ensure that $\mathcal{S}$ contains most of the probability mass by taking ε/M appropriately, and this will then put lower bounds on the smoothing parameter.

1.4 The margin behavior

We will make use of the so-called margin condition (see below), which is assumed for a neighborhood, denoted by $\mathbf{F}_\eta$, of the target f^0. Some of the effort goes in proving that the estimator is indeed in this neighborhood. Here, we use arguments that rely on the assumed convexity of $f \mapsto \rho_f$. These arguments also show that the stochastic part of the problem needs only to be studied "locally" (i.e., (2) for "small" values of M).

The margin condition requires that in the neighborhood $\mathbf{F}_\eta \subset \mathbf{F}$ of f^0, the excess risk $\mathcal{E}$ is bounded from below by a strictly convex function. This is true in many particular cases for an L_∞ neighborhood $\mathbf{F}_\eta = \{\|f - f^0\|_\infty \leq \eta\}$ (for $\mathbf{F}$ being a class of functions).

Definition *We say that the* margin condition *holds with strictly convex function G, if for all $f \in \mathbf{F}_\eta$, we have*

$$\mathcal{E}(f) \geq G(\|f - f^0\|).$$

Indeed, in typical cases, the margin condition holds with quadratic function G, that is, $G(u) = cu^2$, $u \geq 0$, where c is a positive constant.

We also need the notion of *convex conjugate* of a function.

Definition *Let G be a strictly convex function on $[0, \infty)$, with $G(0) = 0$. The* convex conjugate *H of G is defined as*

$$H(v) = \sup_u \{uv - G(u)\}, \ v \geq 0.$$

Note that when G is quadratic, its convex conjugate H is quadratic as well. The function H will appear in the estimation error term.

2. An oracle inequality

Our goal is now, to show that the estimator (1) has oracle properties, namely, to show, for a set $\mathcal{B} \subset \mathbf{R}^p$ as large as possible, that

$$\mathcal{E}(f_{\hat{\beta}}) \leq \text{const.} \min_{\beta \in \mathcal{B}} \left\{ \mathcal{E}(f_\beta) + H\left(\lambda \sqrt{J_\beta}\right) \right\}, \qquad (3)$$

with large probability. Here, H is the convex conjugate of the function G appearing in the margin condition. The left hand side of (3) is small if the target f^0 can be well approximated by an f_β with only a few of the coefficients β_j $(j = 1, \ldots, p)$ non-zero, that is, by a *sparse* f_β.

To arrive at such an inequality, we need a certain amount of compatibility between the ℓ^1 norm $\|\cdot\|_1$ and the metric $\|\cdot\|$ on the vector space $\mathbf{F}$.

2.1 The compatibility condition

Let $\mathcal{J} \subset \{1, \ldots, p\}$ be an index set.

Definition *We say that the* compatibility condition *is met for the set $\mathcal{J}$, with constants $\phi(\mathcal{J}) > 0$ and $\nu(\mathcal{J}) > 0$, if for all $\beta \in \mathbf{R}^p$ that satisfy $\sum_{j \notin \mathcal{J}} |\beta_j| \leq 3 \sum_{j \in \mathcal{J}} |\beta_j|$, it holds that*

$$\sum_{j \in \mathcal{J}} |\beta_j| \leq \sqrt{|\mathcal{J}|} \|f_\beta\| / \phi(\mathcal{J}) + \sum_{j \notin \mathcal{J}} |\beta_j| / (1 + \nu(\mathcal{J})). \quad (4)$$

More details on this condition are in Section 3.

2.2 The result

Let $0 < \delta < 1$ be fixed but otherwise arbitrary. For $\beta \in \mathbf{R}^p$, set $\mathcal{J}_\beta := \{j : \ \beta_j \neq 0\}$ and $J_\beta := |\mathcal{J}_\beta|$. The oracle β^* is defined as

$$\beta^* := \arg\min_{\beta \in \mathcal{B}} \left\{ (1 + \delta)\mathcal{E}(f_\beta) + 2\delta H\left(\frac{2\lambda \sqrt{J_\beta}}{\phi(\mathcal{J})\delta} \right) \right\},$$

Here, the minimum is over the set $\mathcal{B}$ of all β, for which the compatibility condition holds for $\mathcal{J}_\beta$.

Let $\mathcal{J}^* := \{j : \ \beta_j^* \neq 0\}$, and let $\phi^* = \phi(\mathcal{J}^*)$ and $\nu^* = \nu(\mathcal{J}^*)$. Set

$$\varepsilon^* := (1 + \delta)\mathcal{E}(f_{\beta^*}) + 2\delta H\left(\frac{2\lambda \sqrt{J_{\beta^*}}}{\phi^* \delta} \right).$$

Lemma 2.1 *(Lasso under the compatibility condition) Consider a constant λ_0 satisfying the inequality*

$$\lambda_0 \leq \frac{\nu^*}{8(2+\nu^*)}\lambda, \tag{5}$$

and let $M^* = \varepsilon^*/\lambda_0$. Assume the margin condition with strictly convex function G, and that the oracle f_{β^*} is in the neighborhood $\mathbf{F}_\eta$ of the target, as well as $f_\beta \in \mathbf{F}_\eta$ for all $\|\beta - \beta^*\|_1 \leq M^*$. Then on the set

$$\mathcal{S} := \{\mathbf{Z}_{M^*} \leq \varepsilon^*\}, \tag{6}$$

we have either

$$(1-\delta)\mathcal{E}(f_{\hat\beta}) + \frac{\nu^*}{2+\nu^*}\lambda\|\hat\beta - \beta^*\|_1 \leq 2\varepsilon^*,$$

or

$$\mathcal{E}(f_{\hat\beta}) + \lambda\|\hat\beta - \beta^*\|_1 \leq \left(\frac{4(2+\nu^*)}{\nu^*}\right)\varepsilon^*.$$

2.3 Asymptotics in n

In the case of independent observations, the set $\mathcal{S}$ has large probability under general conditions, for $\lambda_0 \asymp \sqrt{\log(2p)/n}$ (see van de Geer (2007)). Taking λ also of this order, and assuming a a quadratic margin, and in addition that ϕ^* stays away from zero, the estimation error $H(2\lambda\sqrt{J_{\beta^*}}/(\phi^*\delta))$ behaves like $J_{\beta^*}\log(2p)/n$.

2.4 The case $\nu^* = \infty$

We observe that often λ_0 can be taken independently of oracle values, whereas (5) requires a lower bound on the smoothing parameter λ which depends on the unknown oracle value ν^*. The best possible situation is when ν^* is arbitrarily large. Now, the case $\nu^* > 2$ corresponds after a reparametrization to the case $\nu^* = \infty$. Indeed, if $\nu(\mathcal{J}) > 2$ and $\sum_{j\notin\mathcal{J}}|\beta_j| \leq 3\sum_{j\in\mathcal{J}}|\beta_j|$, then (4) implies

$$\sum_{j\in\mathcal{J}}|\beta_j| \leq \left(\frac{\nu(\mathcal{J})-2}{1+\nu(\mathcal{J})}\right)\sqrt{|\mathcal{J}|}\|f_\beta\|/\phi(\mathcal{J}).$$

With $\nu^* = \infty$, Lemma 2.1 reads

Corollary 2.1 *Let $\lambda_0 \leq \lambda/8$, and $M^* := \varepsilon^*/\lambda_0$. Assume the conditions of Lemma 2.1 with $\nu^* = \infty$. Then on the set $\mathcal{S}$ given in (6), we have either*

$$(1-\delta)\mathcal{E}(f_{\hat\beta}) + \lambda\|\hat\beta - \beta^*\|_1 \leq 2\epsilon^*,$$

or

$$\mathcal{E}(f_{\hat\beta}) + \lambda\|\hat\beta - \beta^*\|_1 \leq 4\epsilon^*.$$

3. On the compatibility condition

Let $\|f_\beta\|^2 := \beta^T\Sigma\beta$, with Σ a $p\times p$ matrix. The entries of Σ are denoted by $\sigma_{j,k}$ $(j,k\in\{1,\ldots,p\})$.

We first present a simple lemma.

Lemma 3.1 *Suppose that Σ has smallest eigenvalue $\phi > 0$. Then the compatibility condition holds for all index sets $\mathcal{J}$, with $\phi(\mathcal{J}) = \phi$ and $\nu(\mathcal{J}) = \infty$.*

3.1 The coherence lemma

For a vector v, and for $q \geq 1$, we shall write

$$\|v\|_q = (\sum_j |v_j|^q)^{1/q},$$

with the usual extension for $q = \infty$.

We now fix an $L \in \{1,\ldots,p\}$. Consider a partition of the form

$$\Sigma := \begin{pmatrix}\Sigma_{1,1} & \Sigma_{1,2}\\ \Sigma_{2,1} & \Sigma_{2,2}\end{pmatrix}$$

where $\Sigma_{1,1}$ is an $L\times L$ matrix, $\Sigma_{2,1}$ is a $(p-L)\times L$ matrix and $\Sigma_{1,2} = \Sigma_{2,1}^T$ is its transpose, and where $\Sigma_{2,2}$ is a $(p-L)\times(p-L)$ matrix. Then for any $b_1 \in \mathbf{R}^L$ and $b_2 \in \mathbf{R}^{p-L}$, and for $b^T := (b_1^T, b_2^T)$, one has

$$b_1^T\Sigma_{1,1}b_1 \leq b^T\Sigma b + 2\|\Sigma_{2,1}\|_q\|b_1\|_2\|b_2\|_r,$$

with $1/q + 1/r = 1$, and

$$\|\Sigma_{2,1}\|_q := \sup_{a\in\mathbf{R}^L,\ \|a\|_2=1}\|\Sigma_{2,1}a\|_q.$$

Consider now a subset $\mathcal{L} \subset \{1,\ldots,p\}$. We introduce the $|\mathcal{L}|\times|\mathcal{L}|$ matrix

$$\Sigma_{1,1}(\mathcal{L}) := (\sigma_{j,k})_{j,k\in\mathcal{L}},$$

and the $(p-|\mathcal{L}|)\times|\mathcal{L}|$ matrix

$$\Sigma_{2,1}(\mathcal{L}) := (\sigma_{j,k})_{j\notin\mathcal{L},k\in\mathcal{L},}.$$

We let $\rho^2(\mathcal{L})$ be the smallest eigenvalue of $\Sigma_{1,1}(\mathcal{L})$ and take

$$\vartheta_q^2(\mathcal{L}) := \|\Sigma_{2,1}(\mathcal{L})\|_q.$$

We moreover define

$$\phi(\mathcal{J}) := \min_{\mathcal{L}\supset\mathcal{J}:\ |\mathcal{L}|=L}\rho(\mathcal{L}),$$

and

$$\theta_q(\mathcal{J}) := \max_{\mathcal{L}\supset\mathcal{J}:\ |\mathcal{L}|=L}\vartheta_q(\mathcal{L}).$$

Lemma 3.2 *(Coherence lemma) Fix $\mathcal{J} \subset \{1,\ldots,p\}$, with cardinality $|\mathcal{J}| := J \leq L$. Then we have for any $\beta \in \mathbf{R}^p$,*

$$\sum_{j\in\mathcal{J}}|\beta_j| \leq \frac{\sqrt{J}}{\phi(\mathcal{J})}\|f_\beta\|_2 + \sum_{j\notin\mathcal{J}}|\beta_j|/(1+\nu(\mathcal{J}))$$

with, for $1/q + 1/r = 1$,

$$\frac{1}{1+\nu(\mathcal{J})} = \begin{cases}\frac{2\sqrt{J}q^{1/r}}{(L+1-J)^{1/q}}\frac{\theta_q^2(\mathcal{J})}{\phi^2(\mathcal{J})} & 1\leq q<\infty\\ 2\sqrt{J}\frac{\theta_\infty^2(\mathcal{J})}{\phi^2(\mathcal{J})} & q=\infty\end{cases}.$$

The coherence lemma uses an auxiliary lemma (Lemma 4.1) which is based on ideas in Candes and Tao (2007).

3.2 Relation with other coherence conditions

In the coherence lemma, the choice of $L \geq J$ and $q \geq 1$ is open. They are allowed to depend on $\mathcal{J}$, and for our purposes should be taken in such a way that $\phi(\mathcal{J})$ and $\nu(\mathcal{J})$ are as large as possible.

To gain insight into what the coherence lemma can bring us, let us again consider the partition

$$\Sigma := \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}.$$

Note first that $\|\Sigma_{2,1}\|_2^2$ is the largest eigenvalue of the matrix $(\Sigma_{1,2}\Sigma_{2,1})$. The coherence lemma with $q = 2$ is along the lines of the coherence condition in Candes and Tao (2007). They choose L not larger than $3J$.

It is easy to see that

$$\|\Sigma_{2,1}\|_q \leq \left(\sum_{j=L+1}^{p} \left(\sqrt{\sum_{k=1}^{L} \sigma_{j,k}^2} \right)^q \right)^{1/q}. \tag{7}$$

Thus,

$$\|\Sigma_{2,1}\|_q \leq \sqrt{L} \left(\sum_{j=L+1}^{p} \max_{1 \leq k \leq L} |\sigma_{j,k}|^q \right)^{1/q}.$$

For $q = \infty$, this reads

$$\|\Sigma_{2,1}\|_\infty \leq \sqrt{L} \max_{L+1 \leq j \leq p} \max_{1 \leq k \leq L} |\sigma_{j,k}|.$$

Note also that with $q = \infty$, the choice $L = J$ in the coherence lemma gives the "best" result. With this choice of L, the coherence lemma with $q = \infty$ gives a positive value for $\nu(\mathcal{J})$ (required in the compatibility conditions), if

$$2J \max_{j \notin \mathcal{J}} \max_{k \in \mathcal{J}} |\sigma_{j,k}|/\phi^2(\mathcal{J}) < 1.$$

This is in the spirit of the maximal local coherence condition in Bunea et al. (2007).

It also follows from (7), that

$$\|\Sigma_{2,1}\|_q \leq \left(\sum_{j=L+1}^{p} \left(\sum_{k=1}^{L} |\sigma_{j,k}| \right)^q \right)^{1/q}.$$

Invoking this bound with $q = 1$ and $L = J$ in the coherence lemma leads to a condition which is similar to the cumulative local coherence condition in Bunea et al. (2007).

4. Proofs

Proof of Lemma 2.1. We write

$$\beta = \beta_{\text{in}} + \beta_{\text{out}},$$

with $\beta_{j,\text{in}} = \beta_j 1\{j \in \mathcal{J}^*\}$ and $\beta_{j,\text{out}} = \beta_j 1\{j \notin \mathcal{J}^*\}$. Note thus that $\beta_{\text{in}}^* = \beta^*$ and $\beta_{\text{out}}^* \equiv 0$.

Throughout the proof, we assume we are on the set $\mathcal{S}$.

Let

$$s := \frac{M^*}{M^* + \|\hat{\beta} - \beta^*\|_1},$$

and $\tilde{\beta} := s\hat{\beta} + (1-s)\beta^*$. Write $\tilde{f} := f_{\tilde{\beta}}$, $\tilde{\mathcal{E}} := \mathcal{E}(\tilde{f})$ and $f^* := f_{\beta^*}$, $\mathcal{E}^* := \mathcal{E}(f^*)$.

The proof is structured as follows. We first note that $\|\tilde{\beta} - \beta^*\|_1 \leq M^*$ and prove the result with $\hat{\beta}$ replaced by $\tilde{\beta}$. As a byproduct we obtain that $\|\hat{\beta} - \beta^*\|_1 \leq M^*$. The latter allows us to repeat the proof with $\tilde{\beta}$ replaced by $\hat{\beta}$.

By the convexity of $f \mapsto \rho_f$, and of $\|\cdot\|_1$,

$$P_n\rho_{\tilde{f}} + \lambda\|\tilde{\beta}\|_1 \leq s[P_n\rho_{f_{\hat{\beta}}} + \lambda\|\hat{\beta}\|_1]$$

$$+(1-s)[P_n\rho_{f^*} + \lambda\|\beta^*\|_1] \leq P_n\rho_{f^*} + \lambda\|\beta^*\|_1.$$

Thus,

$$\begin{aligned} &\tilde{\mathcal{E}} + \lambda\|\tilde{\beta}\|_1 \\ &= -(P_n - P)(\rho_{\tilde{f}} - \rho_{f^*}) + P_n(\rho_{\tilde{f}} - \rho_{f^*}) + \mathcal{E}^* + \lambda\|\tilde{\beta}\|_1 \\ &\leq -(P_n - P)(\rho_{\tilde{f}} - \rho_{f^*}) + \mathcal{E}^* + \lambda\|\beta^*\|_1 \\ &\leq \mathbf{Z}_{M^*} + \mathcal{E}^* + \lambda\|\beta^*\|_1. \end{aligned} \tag{8}$$

So (on $\mathcal{S}$)

$$\tilde{\mathcal{E}} + \lambda\|\tilde{\beta}\|_1 \leq \varepsilon^* + \mathcal{E}^* + \lambda\|\beta^*\|_1. \tag{9}$$

Therefore, we have

$$\begin{aligned} &\tilde{\mathcal{E}} + \lambda\|\tilde{\beta}_{\text{out}}\|_1 \\ &\leq \varepsilon^* + \mathcal{E}^* + \lambda\|\tilde{\beta}_{\text{in}} - \beta^*\|_1 \leq 2\varepsilon^* + \lambda\|\tilde{\beta}_{\text{in}} - \beta^*\|_1. \end{aligned} \tag{10}$$

Case 1. If

$$\lambda\|\tilde{\beta}_{\text{in}} - \beta^*\|_1 \leq \left(\frac{2(2+\nu^*)}{\nu^*} - 1 \right) \varepsilon^*,$$

then (10) implies

$$\tilde{\mathcal{E}} + \lambda\|\tilde{\beta}_{\text{out}}\|_1 \leq \left(\frac{2(2+\nu^*)}{\nu^*} + 1 \right) \varepsilon^*,$$

and hence

$$\tilde{\mathcal{E}} + \lambda\|\tilde{\beta} - \beta^*\|_1 \leq \left(\frac{4(2+\nu^*)}{\nu^*} \right) \varepsilon^*. \tag{11}$$

But then

$$\|\tilde{\beta} - \beta^*\|_1 \leq \frac{4(2+\nu^*)}{\nu^*} \frac{\varepsilon^*}{\lambda} = \frac{4(2+\nu^*)}{\nu^*} \frac{\lambda_0}{\lambda} M^* \leq \frac{M^*}{2},$$

since $\lambda \geq 8(2+\nu^*)\lambda_0/\nu^*$. This implies that $\|\hat{\beta} - \beta^*\|_1 \leq M^*$.

Case 2. If

$$\lambda\|\tilde{\beta}_{\text{in}} - \beta^*\|_1 \geq \left(\frac{2(2+\nu^*)}{\nu^*} - 1 \right) \varepsilon^*,$$

we get from (10) and using

$$\left(\frac{2(2+\nu^*)}{\nu^*}-1\right)=\frac{4+\nu*}{\nu^*}\geq 1,$$

that

$$\lambda\|\tilde{\beta}_{\text{out}}\|_1\leq 2\varepsilon^*+\lambda\|\tilde{\beta}_{\text{in}}-\beta^*\|_1\leq 3\lambda\|\tilde{\beta}_{\text{in}}-\beta^*\|_1.$$

This means that we can apply the weak compatibility condition.

Recall that inequality (9) implies

$$\tilde{\mathcal{E}}+\lambda\|\tilde{\beta}_{\text{out}}\|_1\leq\varepsilon^*+\mathcal{E}^*+\lambda\|\beta^*\|_1-\lambda\|\tilde{\beta}_{\text{in}}\|_1$$
$$\leq\varepsilon^*+\mathcal{E}^*+\lambda\|\tilde{\beta}_{\text{in}}-\beta^*\|_1.$$

We have

$$\|\tilde{\beta}_{\text{out}}\|_1=\frac{2}{2+\nu^*}\|\tilde{\beta}_{\text{out}}\|_1+\frac{\nu^*}{2+\nu^*}\|\tilde{\beta}_{\text{out}}\|_1,$$

and

$$\|\tilde{\beta}_{\text{out}}\|_1=\|\tilde{\beta}-\beta^*\|_1-\|\tilde{\beta}_{\text{in}}-\beta^*\|_1.$$

So

$$\tilde{\mathcal{E}}+\frac{2}{2+\nu*}\lambda\|\tilde{\beta}_{\text{out}}\|_1+\frac{\nu^*}{2+\nu^*}\lambda\|\tilde{\beta}-\beta^*\|_1$$
$$\leq\varepsilon^*+\mathcal{E}^*+\frac{2(1+\nu^*)}{2+\nu^*})\lambda\|\tilde{\beta}_{\text{in}}-\beta^*\|_1.$$

With the compatibility condition on $\mathcal{J}^*$, we find

$$\|\tilde{\beta}_{\text{in}}-\beta^*\|_1\leq\sqrt{J}^*\|f_{\tilde{\beta}}-f_{\beta^*}\|/\phi^*+\|\tilde{\beta}_{\text{out}}\|_1/(1+\nu^*).$$

Thus

$$\tilde{\mathcal{E}}+\frac{2}{2+\nu^*}\lambda\|\tilde{\beta}_{\text{out}}\|_1+\frac{\nu^*}{2+\nu^*}\lambda\|\tilde{\beta}-\beta^*\|_1$$

$$\leq\varepsilon^*+\mathcal{E}^*+\frac{2(1+\nu^*)}{2+\nu^*}\sqrt{J}^*\|f_{\tilde{\beta}}-f_{\beta^*}\|/\phi^*+\frac{2}{2+\nu^*}\lambda\|\tilde{\beta}_{\text{out}}\|_1.$$

The term with $\|\tilde{\beta}_{\text{out}}\|_1$ cancels out. Moreover, we may use the bound $(1+\nu^*)/(2+\nu^*)\leq 1$. This allows us to conclude that

$$\tilde{\mathcal{E}}+\frac{\nu^*}{2+\nu^*}\lambda\|\tilde{\beta}-\beta^*\|_1\leq\varepsilon^*+\mathcal{E}^*+2\lambda\sqrt{J}^*\|f_{\tilde{\beta}}-f_{\beta^*}\|/\phi^*.$$

Now, because we assumed $f_{\beta^*}\in\mathbf{F}_\eta$, and since also $f_{\tilde{\beta}}\in\mathbf{F}_\eta$, we can invoke the margin condition to arrive at

$$2\lambda\sqrt{J}^*\|f_{\tilde{\beta}_n}-f_{\beta_n^*}\|/\phi^*\leq 2\delta H\left(\frac{2\lambda\sqrt{J^*}}{\phi^*\delta}\right)+\delta\tilde{\mathcal{E}}+\delta\mathcal{E}^*.$$

It follows that

$$\tilde{\mathcal{E}}+\frac{\nu^*}{2+\nu^*}\lambda\|\tilde{\beta}-\beta^*\|_1$$
$$\leq\varepsilon^*+(1+\delta)\mathcal{E}^*+2\delta H\left(\frac{2\lambda\sqrt{J^*}}{\phi^*\delta}\right)+\delta\tilde{\mathcal{E}}$$
$$=\varepsilon^*+\varepsilon^*+\delta\tilde{\mathcal{E}}=2\varepsilon^*+\delta\tilde{\mathcal{E}},$$

or

$$(1-\delta)\tilde{\mathcal{E}}+\frac{\nu^*}{2+\nu^*}\lambda\|\tilde{\beta}-\beta^*\|_1\leq 2\varepsilon^*.\qquad(12)$$

This yields

$$\|\tilde{\beta}-\beta^*\|_1\leq\frac{2(2+\nu^*)\lambda_0}{\nu^*\lambda}M^*\leq\frac{M^*}{4},$$

where the last inequality is ensured by the assumption $\lambda\geq 8(2+\nu^*)/\nu^*\lambda_0$.

But $\|\tilde{\beta}-\beta^*\|_1\leq M^*/4$ implies

$$\|\hat{\beta}-\beta^*\|_1\leq\frac{M^*}{3}\leq M^*.$$

So in both Case 1 and Case 2, we arrive at $\|\hat{\beta}-\beta^*\|_1\leq M^*$. Now, repeat the above arguments with $\tilde{\beta}$ replaced by $\hat{\beta}$. Let $\hat{\mathcal{E}}=\mathcal{E}(f_{\hat{\beta}})$. Then, either from (11) (Case 1) with this replacement:

$$\hat{\mathcal{E}}+\lambda\|\hat{\beta}-\beta^*\|_1\leq\left(\frac{4(2+\nu^*)}{\nu^*}\right)\varepsilon^*.$$

Or we are in Case 2 and can apply the compatibility assumption. Then we arrive at (12) with replacement:

$$(1-\delta)\hat{\mathcal{E}}+\frac{\nu^*}{2+\nu^*}\lambda\|\hat{\beta}-\beta^*\|_1\leq 2\varepsilon^*.$$

□

Proof of Lemma 3.1. Let $\beta\in\mathbf{R}^p$ be arbitrary. It is clear that

$$\|\beta\|_2^2\leq\|f_\beta\|^2/\phi^2.$$

Hence,

$$\sum_{j\in\mathcal{J}}|\beta_j|\leq\sqrt{|\mathcal{J}|}\left(\sum_{j\in\mathcal{J}}|\beta_j|^2\right)^{1/2}$$
$$\leq\sqrt{|\mathcal{J}|}\|\beta\|_2\leq\sqrt{|\mathcal{J}|}\|f_\beta\|/\phi.$$

□

Proof of Lemma 3.2. Let $\beta=\beta_{\text{in}}+\beta_{\text{out}}$, with $\beta_{j,\text{in}}:=\beta_j 1_{j\in\mathcal{J}}$ and $\beta_{j,\text{out}}:=\beta_j 1_{j\notin\mathcal{J}}$, $j=1,\ldots,p$. Let $\mathcal{L}\supset\mathcal{J}$, with $\mathcal{L}\backslash\mathcal{J}$ being the set of the $L-J$ largest $|\beta_j|$ with $j\notin\mathcal{J}$. Define $b_{j,1}:=\beta_j 1_{j\in\mathcal{L}}$ and $b_{j,2}:=\beta_j 1_{j\notin\mathcal{L}}$. We have

$$|b_1^T\Sigma b_2|\leq\theta_q^2(\mathcal{J})\|b_1\|_2\|b_2\|_r.$$

By the auxiliary lemma below, for $q<\infty$,

$$\|b_2\|_r\leq\|\beta_{\text{out}}\|_1/(L+1-J)^{1/q}q^{1/r}.\qquad(13)$$

This yields

$$|b_1^T\Sigma b_2|\leq\theta_q^2(\mathcal{J})\|b_1\|_2\|\beta_{\text{out}}\|_1/(L+1-J)^{1/q}q^{1/r}.$$

Hence

$$\|f_{b_1}\|^2\leq\|f_\beta\|^2+2|b_1^T\Sigma b_2|$$
$$\leq\|f_\beta\|^2+2\theta_q^2(\mathcal{J})\|b_1\|_2\|\beta_{\text{out}}\|_1/(L+1-J)^{1/q}q^{1/r}.$$

It follows that

$$\|b_1\|_2^2\leq\frac{\|f_{b_1}\|^2}{\phi^2(\mathcal{J})}$$

$$\leq \frac{\|f_\beta\|^2}{\phi^2(\mathcal{J})} + 2\frac{\theta_q^2(\mathcal{J})}{\phi^2(\mathcal{J})}\|b_1\|_2\|\beta_{\text{out}}\|_1/(L+1-J)^{1/q}q^{1/r}.$$

This implies

$$\|b_1\|_2 \leq \frac{\|f_\beta\|}{\phi(\mathcal{J})} + 2\frac{\theta_q^2(\mathcal{J})}{\phi^2(\mathcal{J})}\|\beta_{\text{out}}\|_1/(L-J)^{1/q}.$$

We thus arrive at

$$\|\beta_{\text{in}}\|_1 \leq \sqrt{J}\|\beta_{\text{in}}\|_2$$

$$\leq \frac{\sqrt{J}\|f_\beta\|}{\phi(\mathcal{J})} + 2\sqrt{J}\frac{\theta_q^2(\mathcal{J})}{\phi^2(\mathcal{J})}\|\beta_{\text{out}}\|_1/(L-J)^{1/q}.$$

The result with $q = \infty$ follows in the same manner, using instead of (13), the trivial bound

$$\|b_1\|_1 \leq \|\beta_{\text{out}}\|_1.$$

□

Lemma 4.1 *(Auxiliary lemma) Let $b_1 \geq b_2 \geq \ldots \geq 0$. For $1 < r \leq \infty$, $1/q + 1/r = 1$, and $L \in \{0, 1, \ldots\}$, we have*

$$(\sum_{j \geq L+1} |b_j|^r)^{1/r} \leq q^{1/r}(L+1)^{-1/q}\|b\|_1.$$

Proof. We have

$$\sum_{j \geq L+1} \frac{1}{j^r} \leq \frac{1}{(L+1)^r} + \int_{L+1}^{\infty} \frac{1}{x^r}\mathrm{d}x \leq \frac{q}{(L+1)^{r-1}}.$$

Moreover, for all j,

$$\|b\|_1 \geq \sum_{l=1}^{j} b_l \geq j b_j,$$

so that $b_j \leq \|b\|_1/j$. Hence

$$\sum_{j \geq L+1} b_j^r \leq \|b\|_1^r \sum_{j \geq L+1} \frac{1}{j^r} \leq \|b\|_1^r \frac{q}{(L+1)^{r-1}}.$$

□

REFERENCES

Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2006) "Sparsity oracle inequalities for the Lasso" submitted.

Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007) "Sparse density estimation and aggregation with ℓ_1 penalties", *COLT 2007* 530–543.

Candes, E. and Tao, T. (2007) "The Dantzig selector: statistical estimation when p is much larger than n" To appear in *Ann. Statist.*

van de Geer, S.A. (2007), "High-dimensional generalized linear models and the Lasso" To appear in *Ann. Statist.*