

---

# A Study of Convex Regularizers for Sparse Recovery and Feature Selection

---

Andreas Argyriou  
Toyota Technological Institute at Chicago

## Abstract

We study the problem of recovering a sparse vector from a set of linear measurements. This problem also relates to feature or variable selection in statistics and machine learning. A widely used method for such problems has been regularization with the  $L_1$  norm. We extend this methodology to allow for a broader class of regularizers which includes the  $L_1$  norm. This class is characterized by a convex analytic condition on the subdifferentials of its members. We derive oracle inequalities for the estimators obtained by solving regularization problems of this type. These bounds show that this class of regularizers may be used for sparse recovery with the convergence rate depending on the first-order properties of the regularizer. In particular, the optimal value in these bounds is attained by the  $L_1$  norm. Finally, we present lower bounds on the recovery error, which hold under no assumptions on the design matrix.

## 1 Introduction

In recent years, a significant amount of research in machine learning and statistics has addressed the problem of sparse recovery and prediction from a set of input/output training data. This problem is relevant in different contexts, such as feature selection or variable selection in machine learning, nonparametric statistics, image reconstruction and compressed sensing in signal processing etc. One of the most widely used approaches has been regularization with the  $L_1$  norm, where the empirical error on the data is penalized against the  $L_1$  norm of the estimator [6, 9]. In the case of a quadratic error term, the method is often referred to as Lasso [12, 15, 16]. Another related formulation with a different error term is the Dantzig selector [5].

The theoretical study of these methods usually assumes a probabilistic model in which training data  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  are generated in a linear way from a sparse vector  $w^*$ , which is to be estimated. Sparsity means that most of the elements of this vector are assumed to equal zero. Moreover, the regime of interest is high-dimensional regression in the sense that the number of variables is much larger than the sample size. An amount of research has demonstrated that recovery of  $w^*$  is possible under certain assumptions on the training data [1, 4, 5, 7, 8]. These assumptions quantify the intuition that the covariance matrix of the data should be close enough to the identity matrix in order to be able to recover the sparsity pattern with high probability. In some of this work, such assumptions have led to bounds on the estimation error or the prediction error in terms of the unknown target  $w^*$ , often referred to as *oracle inequalities*.

In this paper, we investigate further the theoretical properties of such methods in a simple probabilistic setting. Our main goal, however, is to extend the theory to regularization methods other than the ones mentioned above. That is, we consider regularization problems which use penalties other than the  $L_1$  norm. At the same time, we assume that this class of penalties behaves similarly to the  $L_1$  norm, in terms of gradients and subgradients. It turns out that sparse recovery is possible with these regularizers and that the first order properties of the penalties determine the quality of the convergence rates. Specifically, we derive oracle inequalities for the error of such estimators, which

are similar to known inequalities for Lasso [1]. As usual, these results are conditioned on Restricted Eigenvalue assumptions on the data covariance matrix.

There are several reasons for this treatment and some related questions:

- the proof techniques rely on convex analytic properties, which can lead to application in a general context, not limited to norms; as a side benefit, we obtain improved bounds in comparison to [1]
- the results are a first step towards a broader study of oracle inequalities for convex programs; thus, we address a broad class of convex regularizers including the elastic net, order statistics, polyhedral functions etc.; on the other hand, an example which cannot be addressed with our treatment is entropic regularization [10]
- by enlarging the scope, we gain new intuition about why regularizers like the  $L_1$  norm are good for sparse recovery; indeed, we see that the convergence rates become optimal exactly in the case of  $L_1$  (Sec. 3)
- our framework allows for parameterization of “sparsity inducing” effects, that is, by adjusting parameters one may trade off between sparsity and lack thereof (Sec. 2.2); one such example is a variation on the elastic net
- the inequalities also highlight the relation between the rates and the constant in the Restricted Eigenvalue assumption; improving the rate requires strengthening this assumption; they also indicate a trade-off between the estimation error and data error, so that they cannot be simultaneously optimized.

In addition, we derive a new result bounding the estimation error from below (Section 4). This inequality holds under no assumptions and shows that there is a limit to estimation which depends on the regularization parameter.

## 2 Methodology

Let us introduce the problem of interest and the model we use. Our goal is to estimate a vector  $w^* \in \mathbb{R}^d$  of regression coefficients. We are given input and output data drawn from a Gaussian model

$$y = Xw^* + \nu \quad (2.1)$$

where  $X \in \mathbb{R}^{n \times d}$  is a deterministic *design matrix*,  $y \in \mathbb{R}^n$ , and  $\nu$  is an  $n$ -dimensional Gaussian random variable with zero mean and covariance  $\sigma^2 I_n$  ( $\sigma > 0$ ). We work in the cases in which  $d$  is larger than  $n$  and hence it is not possible to uniquely recover  $w^*$  even if  $Xw^*$  were known. However, we shall show that it is possible in this model to learn very sparse vectors  $w^*$ , that is, vectors most of whose components are zero. Learning such a regression vector can be done under very general assumptions, stated in Section 3, which are often satisfied in practice.

### 2.1 Regularization with Bounded Gradient Functions

We now propose a family of new regularization methods for estimating  $w^*$  in (2.1). These methods can be seen as an alternative to convex regularization with the  $L_1$  norm, discussed above. Let us denote the set of  $d$ -dimensional vectors with nonnegative components by  $\mathbb{R}_+^d$  and that with positive components by  $\mathbb{R}_{++}^d$ . We will solve the optimization problem

$$\min \left\{ \frac{1}{n} \|Xw - y\|^2 + 2\lambda f(|w|) : w \in \mathbb{R}^d \right\} \quad (2.2)$$

where  $\lambda > 0$  is a regularization parameter and  $f : \mathbb{R}_+^d \rightarrow \mathbb{R}_+$ . By  $|w|$  we denote the vector of absolute values  $(|w_1|, \dots, |w_d|)^\top$ .

As typical,  $f$  is assumed to be *convex* and *increasing* in each coordinate. Under these assumptions, the minimum (2.2) is well defined and the optimization problem is a convex program (easy to verify).

In addition, we want to require that the gradient of the regularizer  $f(|w|)$  exhibits a discontinuity at zero coordinates. This characteristic is important in inducing sparse solutions, a well known fact

from the study of the  $L_1$  norm. To this end, the gradient of  $f$  should be bounded away from zero on the positive side. Moreover, we assume that the gradient of  $f$  is bounded from above as well. As we shall see in Section 3, this assumption is necessary for obtaining theoretical guarantees for sparse estimation. In fact, the range of the upper and lower bounds quantifies how appropriate the regularizer is for sparse estimation.

Let  $f'(z; \delta)$  denote the one-sided directional derivative of  $f$  at  $z$  in direction  $\delta$ , that is,

$$f'(z; \delta) = \lim_{t \rightarrow 0^+} \frac{f(z + t\delta) - f(z)}{t}.$$

If  $f$  is convex, the directional derivative exists on its domain. Summarizing, we shall make the following assumptions.

**Assumption 2.1.** *The function  $f : \mathbb{R}_+^d \rightarrow \mathbb{R}_+$  in (2.2) satisfies*

- $f$  is convex
- for every  $i \in \{1, \dots, d\}$ , there exist constants  $A_i, B_i > 0$  such that, for all  $z \in \mathbb{R}_+^d \setminus \{0\}$  with  $z_i = 0$ ,  
 $f'(z; e_i) = A_i$  and  
 $B_i = \sup \{f'(z + \tau e_i; e_i) : \tau > 0\}$

Geometrically, this assumption translates as a positive slope equal to  $A_i$  at zero and a maximum slope of  $B_i$  at infinity. Looking at the level sets of  $f(|w|)$ , they are nonsmooth at vectors with zero components and many of the tangent subspaces will tend to be tangent at such vectors. The sparser the vector  $w$  the larger the size of the subdifferential  $\partial f(|w|)$  — see (3.1). This gives a first intuition as to why sparse solutions may be expected with these regularizers.

Finally, we may extend the class of regularization problems of interest to include problems of the form

$$\min \left\{ \frac{1}{n} \|Xw - y\|^2 + 2\lambda h(f(|w|)) : w \in \mathbb{R}^d \right\} \quad (2.3)$$

where  $f$  satisfies Assumption 2.1 and  $h : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is an *increasing* function. It is clear that, if  $\lambda$  is allowed to vary in  $(0, +\infty)$ , the solution paths of (2.2) and (2.3) are the same. The function  $h \circ f$  will not, in general, satisfy Assumption 2.1. However, solving a problem of the form (2.3) may sometimes simplify implementation.

## 2.2 Examples

Let us provide some examples of candidate functions satisfying Assumption 2.1. A simple way to obtain such regularizers is by component-wise addition

$$f(z) = \sum_{i=1}^d \varphi(z_i) \quad \forall z \in \mathbb{R}_+^d,$$

where  $\varphi$  is convex and satisfies  $0 < A \leq \varphi'(t) \leq B$  for all  $t > 0$ .

The  $L_1$  norm is an important boundary case with  $\varphi_{\ell_1}(t) = t$ . It is (up to constants) the only regularizer for which  $A = B$ . As mentioned in Section 1, regularization with the  $L_1$  norm has been extensively studied and used for sparse learning in the past. An extended class of regularizers is given by

$$f(z) = \sum_{i=1}^d \alpha_i z_i$$

with  $\alpha_i > 0$  for all  $i = 1, \dots, d$ . In this case,  $A_i = B_i = \alpha_i$ .

The *logistic regularizer*

$$\varphi_{\log}(t) = \log(\alpha + e^t),$$

parameterized by  $\alpha > 0$ , has gradient bounded by  $A = \frac{1}{\alpha+1}, B = 1$ .

Other examples include hyperbolas

$$\varphi_{hyp}(t) = \sqrt{\alpha + (x + \beta)^2}$$

with  $\alpha, \beta > 0$  and  $A = \frac{\beta}{\sqrt{\alpha+\beta^2}}, B = 1$ ,

trigonometric functions,

$$\varphi_{arctan}(t) = t \arctan(\alpha + \beta t)$$

with  $A = \arctan(\alpha), B = \frac{\pi}{2}$ ,

logarithms of hyperbolic functions,

$$\varphi_{lcosh}(t) = \log(\cosh(\alpha + \beta t))$$

with  $A = \beta \tanh(\alpha), B = \beta$  etc.

Other types of valid regularizers may not decompose across components, an example of which is similar to the *elastic net* [3, 17],

$$f(z) = \|z\|_1 + \alpha \|z\|_2,$$

with  $A_i = 1, B_i = 1 + \alpha$ . One may even define the *p-elastic net*, for  $p > 1$ ,

$$f(z) = \|z\|_1 + \alpha \|z\|_p,$$

again with  $A_i = 1, B_i = 1 + \alpha$ .

Finally, one may view the  $L_1$  norm as a *polyhedral* convex function and observe that many polyhedral functions satisfy Assumption 2.1. One category consists of the order statistics,

$$f(z) = \sum_{i=1}^d \alpha_i z_i^\downarrow$$

where  $\alpha_i > 0$ , for every  $i = 1, \dots, d$  and  $z_1^\downarrow \geq \dots \geq z_d^\downarrow$  denote the descending ordering of vector  $z$ .

Here,  $A_i = \min_{i=1}^d \alpha_i, B_i = \max_{i=1}^d \alpha_i$ . The vector analog of the Ky Fan norms is obtained in the limit as the last  $k$  of the  $\alpha_i$  approach 0, signifying that Ky-Fan-like norms are inappropriate for sparse recovery.

More generally, functions of the form

$$f(z) = \max_{k=1}^r \langle v_k, z \rangle$$

where  $v_1, \dots, v_r \in \mathbb{R}_{++}$ , can be considered. To compute the values of  $A_i$  and  $B_i$  one merely requires the differentiation rule for max functions – see e.g. [13, Lemma 24].

### 3 Sparsity Bounds

In this section, we obtain theoretical bounds justifying the use of bounded gradient regularizers for feature selection.

We work with the model (2.1), where the noise variables are i.i.d. Gaussian with zero mean and variance  $\sigma^2$ . We use  $X_i$  to denote the  $i$ -th column of the design matrix  $X$ . For any vector  $w \in \mathbb{R}^d$ , we denote

$$\begin{aligned} J(w) &:= \{i \in \{1, \dots, d\} : w_i \neq 0\} \\ J^c(w) &:= \{i \in \{1, \dots, d\} : w_i = 0\} \\ M(w) &:= |J(w)|. \end{aligned}$$

We start with a few facts about optimality conditions for problem (2.2).

**Lemma 3.1.** Assume that  $f$  satisfies Assumption 2.1. Let  $u$  be a subgradient of the function  $f(|\cdot|)$  at  $w \in \mathbb{R}^d \setminus \{0\}$ . Then

$$\begin{aligned} A_i \leq -f'(|w|; -e_i) \leq u_i \operatorname{sgn}(w_i) \leq f'(|w|; e_i) \leq B_i & \quad \text{if } w_i \neq 0 \\ -A_i \leq u_i \leq A_i & \quad \text{if } w_i = 0 \end{aligned} \quad (3.1)$$

Moreover, flipping the sign of  $u_i$  at any  $i$  such that  $w_i = 0$  produces again a subgradient of  $f(|\cdot|)$  at  $w$ .

Consequently, any minimizer  $\hat{w}$  of problem (2.2) satisfies the necessary conditions

$$\left| \frac{1}{n} X_i^\top (X \hat{w} - y) \right| \leq \lambda B_i \quad \forall i \in \{1, \dots, d\} \quad (3.2)$$

$$\left| \frac{1}{n} X_i^\top (X \hat{w} - y) \right| \geq \lambda A_i \quad \forall i \text{ s.t. } \hat{w}_i \neq 0. \quad (3.3)$$

**Proof.** We observe that  $(f \circ |\cdot|)'(w; \operatorname{sgn}(w_i)e_i) = f'(|w|; e_i)$ ,  $(f \circ |\cdot|)'(w; -\operatorname{sgn}(w_i)e_i) = f'(|w|; -e_i)$  whenever  $w_i \neq 0$  and  $(f \circ |\cdot|)'(w; \pm e_i) = f'(|w|; e_i)$  whenever  $w_i = 0$ . Using Assumption 2.1 and the property of subgradients that  $\langle u, \delta \rangle \leq (f \circ |\cdot|)'(w; \delta)$  for every  $\delta \in \mathbb{R}^d$  [14, Thm. 23.2], we obtain (3.1). The first lower bound also uses the fact that  $A_i = f'(|w| - |w_i|e_i; e_i) \leq -f'(|w|; -e_i)$  by [14, Thm. 24.1]. The assertion about sign flips is immediate from the definition of subgradients.

It is known that  $\hat{w}$  is a minimizer of problem (2.2) if and only if there exists  $u$  in the subdifferential of  $f \circ |\cdot|$  at  $\hat{w}$  such that

$$\frac{2}{n} X^\top (X \hat{w} - y) + 2\lambda u = 0.$$

This condition, combined with (3.1), yields (3.2) and (3.3). ■

The main lemma we will need in order to obtain oracle inequalities is the following. The proof follows the analysis in [1, 4, 11] (which only applies to the case of Lasso) but uses the optimality conditions of Lemma 3.1 instead of norm properties and leads to somewhat different results.

**Lemma 3.2.** Consider the model (2.1) and assume that  $\|X_i\| = \sqrt{n}$  for all  $i \in \{1, \dots, d\}$ . Let  $\hat{w}$  be a solution of the optimization problem (2.2) under Assumption 2.1 on  $f$  and with

$$\lambda > \frac{\Gamma \sigma}{A} \sqrt{\frac{2 \log d}{n}} \quad (3.4)$$

for some  $\Gamma > 0$ . Then, with probability at least  $1 - d^{-\Gamma^2}$ , the following hold for every  $w \in \mathbb{R}^d$

$$\begin{aligned} (A - \rho) \sum_{i \in J^c(w)} |w_i - \hat{w}_i| + \frac{1}{\lambda n} \|X(\hat{w} - w^*)\|^2 & \leq \\ (B + \rho) \sum_{i \in J(w)} |w_i - \hat{w}_i| + \frac{1}{\lambda n} (\hat{w} - w^*)^\top X^\top X (w - w^*) & \end{aligned} \quad (3.5)$$

$$\left| \frac{1}{n} X_i^\top X (w^* - \hat{w}) \right| \leq \lambda (B_i + \rho) \quad \forall i \in \{1, \dots, d\} \quad (3.6)$$

$$M(\hat{w}) \leq \frac{\phi_{\max}}{\lambda^2 n (A - \rho)^2} \|X(\hat{w} - w^*)\|^2 \quad (3.7)$$

where

$$\rho = \frac{\Gamma\sigma}{\lambda} \sqrt{\frac{2\log d}{n}}, \quad (3.8)$$

$$A = \min_{i=1}^d A_i, \quad B = \max_{i=1}^d B_i$$

and  $\phi_{\max}$  denotes the largest eigenvalue of  $X^\top X/n$ .

**Proof.** Let  $\psi(w) = \frac{1}{n}\|Xw - y\|^2 + 2\lambda f(|w|)$ , which is a convex function. Since  $\hat{w}$  is a solution of (2.2), it satisfies

$$\psi'(\hat{w}; w - \hat{w}) \geq 0$$

for every  $w \in \mathbb{R}^d$ . By the ‘‘max formula’’ [2, Sec. 3.1],[14, Thm. 23.2], for every  $w \in \mathbb{R}^d$ , there exists a subgradient  $g \in \partial\psi(\hat{w})$  such that

$$\langle g, w - \hat{w} \rangle \geq 0. \quad (3.9)$$

By Lemma 3.1, any subgradient  $g$  of  $\psi$  at  $\hat{w}$  equals  $g = \frac{2}{n}X^\top(X\hat{w} - y) + 2\lambda u$  for some  $u$  satisfying (3.1). Substituting in (3.9) yields

$$\frac{1}{\lambda n} \langle X\hat{w} - y, Xw - X\hat{w} \rangle + \langle u, w - \hat{w} \rangle \geq 0$$

hence using (2.1),

$$\begin{aligned} \frac{1}{\lambda n} (-\|X\hat{w} - Xw^*\|^2 + \langle X\hat{w} - Xw^*, Xw - Xw^* \rangle - \langle \nu, Xw - X\hat{w} \rangle) + \langle u, w - \hat{w} \rangle &\geq 0 \implies \\ \frac{1}{\lambda n} (\langle X\hat{w} - Xw^*, Xw - Xw^* \rangle + \langle \nu, X\hat{w} - Xw \rangle) + \sum_{i \in J(w)} u_i(w_i - \hat{w}_i) &\geq \\ \sum_{i \in J^c(w)} u_i \hat{w}_i + \frac{1}{\lambda n} \|X\hat{w} - Xw^*\|^2 &\implies \\ \frac{1}{\lambda n} (\langle X\hat{w} - Xw^*, Xw - Xw^* \rangle + \langle \nu, X\hat{w} - Xw \rangle) + \sum_{i \in J(w)} B_i |w_i - \hat{w}_i| &\geq \\ \sum_{i \in J^c(w)} A_i |\hat{w}_i| + \frac{1}{\lambda n} \|X\hat{w} - Xw^*\|^2 & \end{aligned}$$

by (3.1) and Assumption 2.1, and hence

$$\begin{aligned} \frac{1}{\lambda n} ((\hat{w} - w^*)^\top X^\top X(w - w^*) + \langle \nu, X\hat{w} - Xw \rangle) + \left( \max_{i=1}^d B_i \right) \sum_{i \in J(w)} |w_i - \hat{w}_i| &\geq \\ \left( \min_{i=1}^d A_i \right) \sum_{i \in J^c(w)} |w_i - \hat{w}_i| + \frac{1}{\lambda n} \|X\hat{w} - Xw^*\|^2 & \quad (3.10) \end{aligned}$$

By Hölder’s inequality and the hypothesis, we have

$$\langle \nu, X\hat{w} - Xw \rangle \leq \|X^\top \nu\|_\infty \|\hat{w} - w\|_1 \leq \sqrt{n} \|\nu\| \|\hat{w} - w\|_1.$$

Let us consider the event

$$\mathcal{A} = \{ \|\nu\| \leq \lambda \rho \sqrt{n} \}. \quad (3.11)$$

Using (3.8) and a Chernoff bound, we obtain that  $P(\mathcal{A}) \geq 1 - d^{-\Gamma^2}$ . Thus, on event  $\mathcal{A}$ , inequality (3.10) becomes

$$\begin{aligned} \frac{1}{\lambda n} (\hat{w} - w^*)^\top X^\top X (w - w^*) + \rho \|\hat{w} - w\|_1 + \left( \max_{i=1}^d B_i \right) \sum_{i \in J(w)} |w_i - \hat{w}_i| \geq \\ \left( \min_{i=1}^d A_i \right) \sum_{i \in J^c(w)} |w_i - \hat{w}_i| + \frac{1}{\lambda n} \|X(\hat{w} - w^*)\|^2 \end{aligned}$$

which implies (3.5). The lower bound on  $\lambda$  is obtained by enforcing  $\min_{i=1}^d A_i > \rho$ .

To show (3.6), we recall optimality condition (3.2). In addition, on the event  $\mathcal{A}$  we have

$$\left| \frac{1}{n} X_i^\top \nu \right| \leq \lambda \rho \quad (3.12)$$

and (3.6) follows using (2.1).

To show (3.7), we use optimality condition (3.3), which combined with (3.12) and (2.1) yields, on the event  $\mathcal{A}$ ,

$$\left| \frac{1}{n} X_i^\top X(\hat{w} - w^*) \right| \geq \lambda(A_i - \rho) \quad \forall i \text{ s.t. } \hat{w}_i \neq 0.$$

Therefore, on  $\mathcal{A}$  it holds that

$$\begin{aligned} \frac{\phi_{\max}}{\lambda^2 n} \|X(\hat{w} - w^*)\|^2 &\geq \\ \frac{1}{\lambda^2 n^2} (\hat{w} - w^*)^\top X^\top X X^\top X (\hat{w} - w^*) &= \\ \frac{1}{\lambda^2} \sum_{i=1}^d \left( \frac{1}{n} X_i^\top X(\hat{w} - w^*) \right)^2 &\geq \\ \frac{1}{\lambda^2} \sum_{i \in J(\hat{w})} \left( \frac{1}{n} X_i^\top X(\hat{w} - w^*) \right)^2 &\geq \\ \sum_{i \in J(\hat{w})} (A_i - \rho)^2 &\geq \\ M(\hat{w}) \left( \min_{i=1}^d A_i - \rho \right)^2 & \end{aligned}$$

and the assertion follows. ■

Note that the above lemma recovers the case of Lasso in [1] with  $A_i = 1$ ,  $B_i = 1$ ,  $\rho = \frac{1}{2}$ . However, as we shall see in Theorem 3.1, another value of  $\rho$  yields better results. Lemma 3.2 also indicates that the case of the  $L_1$  norm is optimal, in the sense that inequality (3.5) becomes tightest when  $A_i = B_j$ ,  $\forall i, j$ . It is also worth noting the inverse relationship of the regularization parameter  $\lambda$  with  $A_i$  in condition (3.4), since the smaller the gradient the harder it is to obtain sparse solutions and hence the larger the regularization needed.

We now state the *Restricted Eigenvalue* assumption from [1]. Several well known assumptions from the literature imply this condition, as mentioned in that work. Let a vector  $\delta \in \mathbb{R}^d$  and  $J \subseteq \{1, \dots, d\}$ . We denote by  $J^c$  the complement of  $J$  with respect to  $\{1, \dots, d\}$  and by  $\delta_J$  the vector which has the same components as  $\delta$  on  $J$  and zero components on  $J^c$ .

**Assumption 3.1.**  $RE(s, c)$

$$\kappa(s, c) := \min \left\{ \frac{\|X\delta\|}{\sqrt{n}\|\delta_J\|} : J \subseteq \{1, \dots, d\}, |J| \leq s, \delta \neq 0, \|\delta_{J^c}\|_1 \leq c\|\delta_J\|_1 \right\} > 0$$

Under this assumption for  $s = M(w^*)$  we can bound the quality of recovery and obtain the dependence of  $c$  on the properties of  $f$ .

**Theorem 3.1.** Consider the model (2.1) and assume that  $\|X_i\| = \sqrt{n}$  for all  $i \in \{1, \dots, d\}$ . Let  $A = \min_{i=1}^d A_i$ ,  $B = \max_{i=1}^d B_i$ . Let  $\hat{w}$  be a solution of the optimization problem (2.2) under Assumption 2.1 on  $f$  and with

$$\lambda > \frac{\Gamma\sigma}{A} \sqrt{\frac{2 \log d}{n}}$$

for some  $\Gamma > 0$ . Assume also that  $RE(s, c)$  holds with

$$s = M(w^*)$$

and

$$c = \frac{B + \rho}{A - \rho},$$

where

$$\rho = \frac{\Gamma\sigma}{\lambda} \sqrt{\frac{2 \log d}{n}}.$$

Then, with probability at least  $1 - d^{-\Gamma^2}$ , we have

$$\|\hat{w} - w^*\|_1 \leq \frac{(B + A)(B + \rho)\Gamma}{\rho(A - \rho)\kappa^2(s, c)} \sigma s \sqrt{\frac{2 \log d}{n}} \quad (3.13)$$

$$\|X(\hat{w} - w^*)\|^2 \leq \frac{2(B + \rho)^2 \Gamma^2}{\rho^2 \kappa^2(s, c)} \sigma^2 s \log d \quad (3.14)$$

$$M(\hat{w}) \leq \frac{(B + \rho)^2 \phi_{\max}}{(A - \rho)^2 \kappa^2(s, c)} s. \quad (3.15)$$

**Proof.** Setting  $w = w^*$  in (3.5) yields

$$(A - \rho) \sum_{i \in J^c(w^*)} |w_i^* - \hat{w}_i| + \frac{1}{\lambda n} \|X(\hat{w} - w^*)\|^2 \leq (B + \rho) \sum_{i \in J(w^*)} |w_i^* - \hat{w}_i| \quad (3.16)$$

and hence

$$\sum_{i \in J^c(w^*)} |w_i^* - \hat{w}_i| \leq \frac{(B + \rho)}{(A - \rho)} \sum_{i \in J(w^*)} |w_i^* - \hat{w}_i|. \quad (3.17)$$

Thus, using Assumption  $RE(s, c)$  with  $\delta = w^* - \hat{w}$ ,  $J = J(w^*)$  and  $s, c$  as in the hypothesis, we obtain that

$$\frac{\|X(w^* - \hat{w})\|}{\sqrt{n}\|(w^* - \hat{w})_J\|} \geq \kappa(s, c)$$

and combining with (3.16) that

$$\begin{aligned}
(A - \rho) \sum_{i \in J^c(w^*)} |w_i^* - \hat{w}_i| + \frac{\kappa^2(s, c)}{\lambda} \|(w^* - \hat{w})_J\|^2 &\leq \\
(B + \rho) \sum_{i \in J(w^*)} |w_i^* - \hat{w}_i| &\implies \\
\frac{\kappa^2(s, c)}{\lambda s} \left( \sum_{i \in J(w^*)} |w_i^* - \hat{w}_i| \right)^2 &\leq (B + \rho) \sum_{i \in J(w^*)} |w_i^* - \hat{w}_i| \tag{3.18}
\end{aligned}$$

In combination with (3.17) this yields (3.13).

To show (3.14), we obtain

$$\frac{1}{\lambda n} \|X(\hat{w} - w^*)\|^2 \leq (B + \rho) \sum_{i \in J(w^*)} |w_i^* - \hat{w}_i|$$

from (3.16) and combine with (3.18).

Finally, (3.15) follows from (3.14) and (3.7). ■

The first interesting observations which arise relate to the connection of the bounds and constants in the above theorem. In particular, in the case of Lasso ( $A = B = 1$ ), we can optimize the bound in (3.13) and obtain the values  $\rho = \sqrt{2} - 1$  and  $c = \sqrt{2} + 1$ . However, it is more difficult to decrease the error on the data (3.14), because the bound is optimized as  $\rho$  approaches 1 and  $c$  approaches  $+\infty$ . The results of Theorem 3.1 improve on the constants in [1, Thm. 6.2]. Choosing  $c = \sqrt{2} + 1$ , the bounds (3.13), (3.14), (3.15) become tighter and the Restricted Eigenvalue assumption is weakened. Specifically, we have the following.

**Corollary 3.1** (Lasso). *Consider the model (2.1) and assume that  $\|X_i\| = \sqrt{n}$  for all  $i \in \{1, \dots, d\}$ . Let  $\hat{w}$  be a solution of the optimization problem (2.2) with  $f = \|\cdot\|_1$  and*

$$\lambda = (2 + \sqrt{2})\Gamma\sigma\sqrt{\frac{\log d}{n}}$$

for some  $\Gamma > 0$ . Assume also that  $RE(s, c)$  holds with

$$s = M(w^*)$$

and

$$c = \sqrt{2} + 1.$$

Then, with probability at least  $1 - d^{-\Gamma^2}$ , it holds that

$$\begin{aligned}
\|\hat{w} - w^*\|_1 &\leq \frac{(6\sqrt{2} + 8)\Gamma}{\kappa^2(s, c)} \sigma s \sqrt{\frac{\log d}{n}} \\
\|X(\hat{w} - w^*)\|^2 &\leq \frac{(12 + 8\sqrt{2})\Gamma^2}{\kappa^2(s, c)} \sigma^2 s \log d \\
M(\hat{w}) &\leq \frac{(3 + 2\sqrt{2})\phi_{\max}}{\kappa^2(s, c)} s.
\end{aligned}$$

In general, (3.13) is optimized for  $\rho = \sqrt{B(B + A)} - B$ , which yields a value of  $c = \frac{1}{A}(\sqrt{B(B + A)} + B)$ . We also remark that  $c$  and the sparsity of  $\hat{w}$  can be decreased at the cost of increasing the estimation error and data error bounds. Thus, a region around  $\rho = \sqrt{B(B + A)} - B$  is optimal in the sense that all of the above quantities attain reasonably small values. This trade off between the extremes of underfitting and overfitting is determined by the regularization parameter  $\lambda$ .

In fact, the bounds of Theorem 3.1 are vacuous in the regime of high regularization. Optimizing  $c$  and the sparsity of the solution (3.15) leads to  $\rho = 0$ , that is, to  $\lambda = +\infty$ . This should be expected. For example, in the case of Lasso, it is known that 0 is a minimizer if and only if  $\lambda \geq \max_{i=1}^d \frac{|X_i^\top y|}{n}$ . In general, if  $\lambda \geq \max_{i=1}^d \frac{|X_i^\top y|}{nA_i}$  and also  $\lambda$  is large enough for 0 to be a minimizer, one obtains stronger inequalities which always hold on event  $\mathcal{A}$ , by replacing  $\kappa$  with the larger quantity  $\frac{\|Xw^*\|}{\sqrt{n}\|w^*\|}$ . Indeed,

$$\begin{aligned} \frac{(B+A)(B+\rho)\Gamma n\|w^*\|^2}{\rho(A-\rho)\|Xw^*\|^2} \sigma s \sqrt{\frac{2\log d}{n}} &= \frac{(B+A)(B+\rho)n\|w^*\|^2 s \lambda}{(A-\rho)\|Xw^*\|^2} \geq \\ \frac{(B+A)(B+\rho)n\|w^*\|_1^2 \lambda}{(A-\rho)\|Xw^*\|^2} &\geq \frac{(B+A)(B+\rho)n\|w^*\|_1 \lambda}{(A-\rho)\|X^\top Xw^*\|_\infty} > \frac{(B+\rho)n\|w^*\|_1 \lambda}{\|X^\top Xw^*\|_\infty} \geq \\ &\frac{\left(\max_{i=1}^d A_i + \rho\right)n\|w^*\|_1 \lambda}{\|X^\top Xw^*\|_\infty} \geq \|w^*\|_1, \end{aligned}$$

where the last inequality follows from  $|X_i^\top Xw^*| \leq |X_i^\top y| + |X_i^\top \nu| \leq \lambda n A_i + \lambda \rho n$ . Similar reasoning applies to (3.14), whereas (3.15) becomes trivial.

We conclude the section with the case of the elastic net.

**Corollary 3.2** (Elastic net). *Consider the model (2.1) and assume that  $\|X_i\| = \sqrt{n}$  for all  $i \in \{1, \dots, d\}$ . Let  $\hat{w}$  be a solution of the optimization problem (2.2) with  $f = \|\cdot\|_1 + \alpha\|\cdot\|_p$ ,  $p > 1$ ,  $\alpha > 0$ , and*

$$\lambda = \left( \sqrt{2 + \frac{2}{1+\alpha}} + \sqrt{2} \right) \Gamma \sigma \sqrt{\frac{\log d}{n}}$$

for some  $\Gamma > 0$ . Assume also that  $RE(s, c)$  holds with

$$s = M(w^*)$$

and

$$c = \sqrt{(1+\alpha)(2+\alpha)} + 1 + \alpha.$$

Then, with probability at least  $1 - d^{-\Gamma^2}$ , it holds that

$$\begin{aligned} \|\hat{w} - w^*\|_1 &\leq \sqrt{2}(2+\alpha) \left( 3 + 2\alpha + 2\sqrt{(1+\alpha)(2+\alpha)} \right) \frac{\Gamma}{\kappa^2(s, c)} \sigma s \sqrt{\frac{\log d}{n}} \\ \|X(\hat{w} - w^*)\|^2 &\leq 2(2+\alpha) \left( 3 + 2\alpha + 2\sqrt{(1+\alpha)(2+\alpha)} \right) \frac{\Gamma^2}{\kappa^2(s, c)} \sigma^2 s \log d \\ M(\hat{w}) &\leq (1+\alpha) \left( 3 + 2\alpha + 2\sqrt{(1+\alpha)(2+\alpha)} \right) \frac{\phi_{\max}}{\kappa^2(s, c)} s. \end{aligned}$$

**Proof.** Set  $\rho = \sqrt{B(B+A)} - B = \sqrt{(1+\alpha)(2+\alpha)} - 1 - \alpha$  in Theorem 3.1. ■

Note that all the bounds in the above corollary are monotone with respect to  $\alpha$ . In the limit  $\alpha = 0$  they recover the Lasso bounds of Corollary 3.1. Moreover,  $c$  is a monotone function of  $\alpha$  as well. Thus, the smaller the  $\alpha$  the better the sparse recovery and the weaker the assumptions required. Oracle inequalities for the elastic net have been studied in [3]. The comparison of those results with Corollary 3.2 is not direct, since here the regularizer is similar but not identical to the elastic net. However, allowing all the parameters such as  $\lambda$  and  $\alpha$  to vary corresponds to the same regularization path of solutions. Another difference is that the bounds in [3] can be meaningful even when  $\kappa$  is close to zero.

## 4 Lower Bounds

In Section 3, we were able to bound  $\|w^* - \hat{w}\|_1$  by assuming the Restricted Eigenvalue condition. In this section, we obtain lower bounds on this quantity which hold without assumptions.

**Theorem 4.1.** *Consider the model (2.1) and assume that  $\|X_i\| = \sqrt{n}$  for all  $i \in \{1, \dots, d\}$ . Let  $\hat{w}$  be a solution of the optimization problem (2.2) under Assumption 2.1 on  $f$ . Also let  $\Gamma > 0$ . Then, with probability at least  $1 - d^{-\Gamma^2}$ ,*

$$\|w^* - \hat{w}\|_1 \geq \frac{\lambda A^2}{2B} - \frac{\Gamma^2 \sigma^2 \log d}{\lambda n B} \quad (4.1)$$

provided that  $\frac{\Gamma \sigma}{A} \sqrt{\frac{2 \log d}{n}} < \lambda < \max_{i=1}^d \frac{|X_i^\top y|}{n A_i}$ , and

$$\|w^* - \hat{w}\|_1 \geq \frac{\lambda A^2 M(\hat{w})}{2B \phi_{\max}} - \frac{\Gamma^2 \sigma^2 \log d}{\lambda n B} \quad (4.2)$$

provided that  $\frac{\Gamma \sigma}{A} \sqrt{\frac{2 \log d}{n}} \sqrt{\frac{\phi_{\max}}{M(\hat{w})}} < \lambda$ , where  $A, B$  and  $\phi_{\max}$  are defined as in Lemma 3.2.

**Proof.** Since  $\hat{w}$  is a solution of (2.2), it holds that

$$\frac{1}{n} \|\nu\|^2 + 2\lambda f(|w^*|) \geq \frac{1}{n} \|X\hat{w} - y\|^2 + 2\lambda f(|\hat{w}|) \quad (4.3)$$

In addition, condition (3.3) along with the hypothesis  $\|X_i\| = \sqrt{n}$  implies that

$$\|X\hat{w} - y\| \geq \sqrt{n} \lambda A.$$

Under the upper bound hypothesis, (3.3) holds even when  $\hat{w} \neq 0$ .

Therefore, (4.3) gives

$$\frac{1}{n} \|\nu\|^2 + 2\lambda (f(|w^*|) - f(|\hat{w}|)) \geq \lambda^2 A^2.$$

Thus, on the event  $\mathcal{A}$  (3.11), whose probability is at least  $1 - d^{-\Gamma^2}$ ,

$$\frac{2}{n} \Gamma^2 \sigma^2 \log d + 2\lambda (f(|w^*|) - f(|\hat{w}|)) \geq \lambda^2 A^2.$$

Since  $f \circ |\cdot|$  is convex, this implies that

$$\frac{2}{n} \Gamma^2 \sigma^2 \log d + 2\lambda \langle w^* - \hat{w}, u \rangle \geq \lambda^2 A^2$$

for all subgradients  $u \in \partial(f \circ |\cdot|)(w^*)$ . We may choose  $u_i$  to have the same sign as  $\hat{w}_i$  whenever  $w_i^* = 0$ , by the sign flipping property in Lemma 3.1. Thus we obtain that

$$\frac{2}{n} \Gamma^2 \sigma^2 \log d + 2\lambda B \sum_{i \in J(w^*)} |w_i^* - \hat{w}_i| \geq \lambda^2 A^2$$

which leads to (4.1). This bound is meaningful only when  $\lambda^2 A^2 - \frac{2}{n} \Gamma^2 \sigma^2 \log d > 0$ , which gives the lower bound on  $\lambda$ .

To show (4.2), we note that

$$\begin{aligned} \frac{\phi_{\max}}{n} \|X\hat{w} - y\|^2 &\geq \\ \frac{1}{n^2} (X\hat{w} - y)^\top X X^\top (X\hat{w} - y) &= \\ \sum_{i=1}^d \left( \frac{1}{n} X_i^\top (X\hat{w} - y) \right)^2 &\geq \\ \lambda^2 A^2 M(\hat{w}), & \end{aligned}$$

due to (3.3), and proceed as before. ■

This theorem states that there is a limitation to how well one may estimate  $w^*$ . This limitation vanishes gradually as we approach the noiseless case ( $\sigma \rightarrow 0$ ).

In the case of Lasso, the lower bounds in the theorem become *largest*, contrary to what happens with the upper bounds of Theorem 3.1. Thus, for regularizers other than the  $L_1$  norm the estimation error lies in a wider interval.

Moreover the lower bounds depend on  $\lambda$  in a monotone way. For small enough  $\lambda$  the bounds decrease to zero. However, this takes the upper bound in Theorem 3.1 towards infinity. Note that (4.1),(4.2) and (3.13) are comparable since they both hold on event  $\mathcal{A}$  of (3.11). Moreover, both (4.1) and (3.13) assume the same interval for  $\lambda$  in the hypothesis. In fact, a simple computation shows that bound (4.1) is always a factor of four below bound (3.13) (see Lemma 5.1 in the appendix).

Thus, the upper bound is minimized for  $\lambda = \frac{\Gamma\sigma}{\sqrt{B(B+A)-B}} \sqrt{\frac{2 \log d}{n}}$  but smaller values of  $\lambda$  may be preferable.

## References

- [1] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 2008. Submitted.
- [2] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics. Springer, 2005.
- [3] F. Bunea. Honest variable selection in linear and logistic regression models via  $\ell_1$  and  $\ell_1 + \ell_2$  penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.
- [4] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.
- [5] E. J. Candès and T. Tao. Rejoinder: the Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *Annals of Statistics*, 35:2392–2404, 2009.
- [6] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.
- [7] D. Donoho. For most large underdetermined systems of linear equations, the minimal  $\ell^1$ -norm near-solution approximates the sparsest near-solution. Preprint, Dept. of Statistics, Stanford University, 2004.
- [8] D. L. Donoho and J. Tanner. Sparse nonnegative solution of underdetermined linear equations by linear programming. Preprint, 2005.
- [9] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–451, 2004.
- [10] V. Koltchinskii. Sparse recovery in convex hulls via entropy penalization. *Annals of Statistics*, 37(3):1332–1359, 2009.
- [11] K. Lounici, M. Pontil, A. B. Tsybakov, and S. A. van de Geer. Taking advantage of sparsity in multi-task learning. In *Proceedings of the Twenty-Second Conference on Computational Learning Theory*, 2009.

- [12] N. Meinshausen and B. Yu. Lasso type recovery of sparse representations for high dimensional data. *Annals of Statistics*, 37(1):246–270, 2009.
- [13] C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.
- [14] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
- [15] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 58(1):267–288, 1996.
- [16] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.
- [17] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67(2):301–320, 2005.

## 5 Appendix

**Lemma 5.1.** *The lower bound in (4.1) is at most  $\frac{1}{4}$  the upper bound in (3.13).*

**Proof.** It is to be shown that

$$\begin{aligned} \frac{\lambda A^2}{2B} - \frac{\Gamma^2 \sigma^2 \log d}{\lambda n B} &\leq \frac{1}{4} \frac{(B+A)(B+\rho)\Gamma}{\rho(A-\rho)\kappa^2(s,c)} \sigma s \sqrt{\frac{2 \log d}{n}} \iff \\ \frac{\lambda A^2}{2B} - \frac{\lambda \rho^2}{2B} &\leq \frac{1}{4} \frac{(B+A)(B+\rho)\lambda s}{(A-\rho)\kappa^2(s,c)} \iff \\ (A-\rho)^2(A+\rho) \frac{\kappa^2(s,c)}{s} &\leq \frac{1}{2} B(B+A)(B+\rho). \end{aligned}$$

Since  $A \leq B$ , it suffices to show that  $\kappa^2(s,c) \leq s$ . Indeed, pick any  $\delta \in \mathbb{R}^d \setminus \{0\}$  such that  $|J(\delta)| \leq s$ . It holds that

$$\kappa^2(s,c) \leq \frac{\|X\delta\|^2}{n\|\delta_J\|^2} = \frac{\|X_J\delta_J\|^2}{n\|\delta_J\|^2} \leq \frac{\|\delta_J\|_1 \|X_J^\top X_J \delta_J\|_\infty}{n\|\delta_J\|^2} \leq \frac{\|\delta_J\|_1^2 \max_{i \in J} \|X_i^\top X_J\|_\infty}{n\|\delta_J\|^2} = \frac{\|\delta_J\|_1^2}{\|\delta_J\|^2} \leq s,$$

where  $J := J(\delta)$  and  $X_J$  denotes the submatrix of  $X$  with just the columns indexed by  $J$ . ■