On ℓ_2 Error Bounds of the Elastic Net when $p \gg n$

Hanzhong Liu and Jinzhu Jia

School of Mathematical Sciences and Center for Statistical Science,
Peking University, P. R. China

Abstract: We study the estimation property of the Elastic Net in high-dimensional settings where the number of parameters p is comparable to or larger than the sample size n. In such a situation one often assumes sparsity of the true regression parameter $\beta^* \in R^p$, i.e., assuming that β^* belongs to an l_q-ball $\mathbb{B}_q(R_q)$ for some $q \in [0, 1]$. In this paper we provide an ℓ_2-estimation error bound for the Elastic Net and the naive Elastic Net using a unified framework for high-dimensional analysis of M-estimators proposed by Negahban et al. (2009). We show that for an exact sparse linear model, the Elastic Net can get the same convergence rate as the Lasso by suitably choosing the tuning parameters but under a weaker restricted strong convexity condition; and for a weak sparse linear models, under the same condition on design matrix, with suitably chosen tuning parameters, the Elastic Net have slightly better error bounds.

Key words and phrases: Lasso; naive Elastic Net; Elastic Net; model selection consistency; estimation consistency

1. Introduction

The literature on high-dimensional statistical inference that deals with models with the number of parameters (p) comparable to or larger than the sample size (n) has enjoyed substantial growth over the last few years. Regularization methods have been shown to have a better accuracy of prediction on future data (Hoerl and Kennard (1970)). The Lasso (Tibshirani (1996)) which regularizes with an ℓ_1 penalty, can generate sparse models and has been studied in much of the recent literature; see, e.g., Osborne, Presnell and Turlach (2000), Efron. Hastie, and Tibshirani (2004), Zhao and Yu (2006), and Wainwright (2007). The Lasso has the advantage of simultaneously performing model selection and estimation, and has been shown to be effective even in high-dimensional settings. But it has some limitations, such as it selects at most n variables before it saturates

in the $p > n$ case and does not perform well when the predictors are highly correlated. Zou and Hastie (2005) proposed the Elastic Net, which regularizes with a combination of the ℓ_1- and ℓ_2-penalties and also has the property of sparsity, to solve the above problems. They stated that the Elastic Net "simultaneously does automatic variable selection and continuous shrinkage, and is able to select groups of correlated variables" and that "simulation studies and real data examples show that the Elastic Net often outperforms the lasso in terms of prediction accuracy". Jia and Yu (2010) study the model selection of the Elastic Net in the general case when p (the number of predictors), s (the number of predictors with non-zero coefficients in the true linear model), and n (the sample size) all go to infinity. They give a sufficient condition EIC (Elastic Irrepresentable Condition) to guarantee the Elastic Net's model selection consistency when Gaussian noise is assumed.

In this paper, we intend to understand the estimation performance of the Elastic Net under the unified framework proposed by Negahban et al. (2009), which can be used for high-dimensional analysis of M-estimators with decomposable regularizers. We provide a ℓ_2-estimation error bound and show that the Elastic Net can be estimation consistent in the high-dimensional settings which allow $p \gg n$.

Assume our data is generated by a linear regression model

$$Y = X\beta^* + \epsilon, \tag{1.1}$$

where $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$ is a vector of i.i.d. random variables with mean 0 and variance σ^2. $Y \in \mathbb{R}^n$ is the response vector, and $X \in \mathbb{R}^{n \times p}$ is the design matrix which is treated as a deterministic one. $\beta^* \in \mathbb{R}^p$ is the vector of model coefficients. The model is assumed to be "sparse", i.e. $\beta^* \in \mathbb{B}_q(R_q)$ (defined below) for some $q \in [0, 1]$:

Definition 1. *The ℓ_q-balls for $q \in [0, 1]$ is defined as*

$$\mathbb{B}_q(R_q) := \left\{ \beta \in \mathbb{R}^p : \ ||\beta||_q^q = \sum_{j=1}^{p} |\beta_j|^q \leq R_q \right\}, \tag{1.2}$$

where in the special case $q = 0$, we have the ℓ_0-ball

$$\mathbb{B}_0(s) := \{\beta \in \mathbb{R}^p : \sum_{j=1}^{p} \mathbb{I}[\beta_j \neq 0] \leq s\} \tag{1.3}$$

corresponding to the set of vectors β with at most s non-zero elements.

The assumption that $\beta^* \in \mathbb{B}_q(R_q)$ with $q \in (0, 1]$ is called a week sparsity assumption and $\beta^* \in \mathbb{B}_0(s)$ exact sparsity assumption.

The Lasso estimate $\hat{\beta}(Lasso)$ is defined by

$$\hat{\beta}(Lasso) = \underset{\beta}{argmin} \left\{ \frac{1}{n}||y - X\beta||_2^2 + \lambda||\beta||_1 \right\}, \tag{1.4}$$

where $\lambda \geq 0$ is the tuning parameter which controls the amount of regularization applied to the estimate. Setting $\lambda = 0$ reverses the the Lasso problem to Ordinary Least Squares which minimizes the unregularized empirical loss. For any vector $a = (a_1, ..., a_m)'$ we have adopted the notation $||a||_2^2 = \sum_{i=1}^m a_i^2$, $||a||_1 = \sum_{i=1}^m |a_i|$, and $\|a\|_\infty = \max_{i=1,...,m} |a_i|$.

The naive Elastic Net estimate $\hat{\beta}(naive)$ is

$$\hat{\beta}(naive) = \underset{\beta}{argmin} \left\{ \frac{1}{n}||y - X\beta||_2^2 + \lambda_1||\beta||_1 + \lambda_{2n}||\beta||_2^2 \right\}, \tag{1.5}$$

where parameters $\lambda_1 \geq 0$ and $\lambda_{2n} \geq 0$ control the amount of regularization applied to the estimate. $\lambda_{2n} = 0$ leads the naive Elastic Net back to the Lasso estimate. We refer to naive Elastic Net as naiveEN. Under the exact sparsity assumption with $\beta^* \in \mathbb{B}_0(s)$, the estimation performance of this naiveEN estimator has already been considered in Bunea (2008) and Hebiri and Van de Geer (2011). Bunea (2008) provides an upper bound on the ℓ_1-estimation error $||\hat{\beta}(naive) - \beta^*||_1$ under the condition called *Condition Stabil*. The results in Hebiri and Van de Geer (2011) are quite close to those in Bunea (2008) except that the former authors also consider to bound some other forms of error, such as $\|X\hat{\beta}(naive) - X\beta^*\|_2$ and $||\hat{\beta}(naive) - \beta^*||_2$.

The Elastic Net estimate $\hat{\beta}(EN)$ are $(1+\lambda_{2n})\hat{\beta}(naive)$ which select the same model as the naive Elastic Net and can improve the prediction performance. Moreover, The correction factor $(1 + \lambda_{2n})$ leads to the fact that the Elastic Net equals the Lasso for the case of orthonormal design where $n^{-1}X^TX = I$ and

Lasso is known to be minimax optimal in this orthonormal design case (Donoho et al. 1995). It has been shown that the Elastic Net is model selection consistent (Jia and Bin Yu (2010)) even when $p \gg n$ under the condition EIC (Elastic Irrepresentable Condition), and it can select the true model even when the lasso cannot. Here we study the estimation property of the Elastic Net and show that in high-dimensional settings and under a weaker condition comparing with Lasso, both the Elastic Net and naive Elastic Net achieve the same convergence rate as the Lasso.

The rest of the paper is organized as follows. In section 2, we review a unified framework proposed in Negahban et al. (2009), which can be used for high-dimensional analysis of M-estimate with decomposable regularizers, and we give general ℓ_2-estimation error bounds for three estimators: the Lasso, the naive Elastic Net and the Elastic Net estimators, using this framework. In section 3, we study error bounds under both exact and weak sparsity assumptions. For general scaling p and n, conditions on the data matrix X and tuning parameters λ_1, λ_{2n} are given such that the Elastic Net estimator has good ℓ_2-estimation error bound and has estimation consistence property. We conclude in section 4. The proofs can be found in the Appendix.

2. General Error Bounds

In this section we provide a general bound on ℓ_2-estimation error $||\hat{\beta} - \beta^*||_2$, where $\hat{\beta}$ can be the Lasso, naiveEN or Elastic Net estimator. In order to get this error bound, we use a unified framework proposed by Negahban et al. (2009), which established general convergence rates for high-dimensional M-estimator. With the general bound established in this section, we give further analysis and results on the naiveEN and the Elastic Net estimator for sparsity settings of β^* in Section 3.

We first review the main results in Negahban et al. (2009) for regularized M-estimators with a decomposable regularizer. Then we transform the Elastice Net estimator to regularized M-estimators with a decomposable regularizer and use the framework from Negahban et al. (2009) to give a general bound on the ℓ_2-estimation error $\|\hat{\beta} - \beta^*\|_2$.

2.1. Review of main results in Negahban et al. (2009)

Given n i.i.d observations $Z_1^n := \{Z_1, ..., Z_n\}$ from some distribution P with the parameter $\beta^* \in \mathbb{R}^p$, the regularized M-estimators are defined by

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{argmin}\{\mathcal{L}(\beta; Z_1^n) + \lambda_n r(\beta)\} \tag{2.1}$$

where $\mathcal{L}$ is a loss function that assigns a cost to any parameter $\beta \in \mathbb{R}^p$, for a given set of observations Z_1^n, r denotes a regularization function (or regularizor), and $\lambda_n \geq 0$ is a tuning parameter. For the ease of notation, we adopt the shorthand $\mathcal{L}(\beta)$ for $\mathcal{L}(\beta; Z_1^n)$. There are two key conditions to obtain the general bound on the ℓ_2-estimation error: the regularization function $r(\beta)$ is a decomposable norm (defined later) and the loss function $\mathcal{L}(\beta)$ satisfies restricted strong convexity (RSC) condition (defined later).

Decomposability Throughout the paper, let $\langle \cdot, \cdot \rangle$ denote the regular Euclidian inner product.

Definition 2. *A norm-based regularizer r is decomposable with respect to the subspace pair $(A, B^\perp)$ of $\mathbb{R}^p$ if $r(\cdot)$ is a norm and*

$$r(u + v) = r(u) + r(v) \text{ for all } u \in A \text{ and } v \in B^\perp, \tag{2.2}$$

where $B^\perp$ is the orthogonal complement of B defined as

$$B^\perp = \{v \in \mathbb{R}^p : \langle u, v \rangle = 0, \text{ for all } u \in B\}.$$

The decomposability property of a regularizer leads to an important consequence. It is shown in Negahban et al. (2009) that under a few mild conditions, the estimation error vector of the regularized M-estimator (2.1) $\hat{\Delta} := \hat{\beta} - \beta^*$ belongs to the set

$$\mathcal{C}(A, B, \beta^*) := \{\Delta \in R^p : \ r(\Delta_{B^\perp}) \leq 3r(\Delta_B) + 4r(\beta^*_{A^\perp})\}, \tag{2.3}$$

where Δ_A denote the Euclidean projection of Δ on subspace A. This consequence plays an essential role in the definition of restricted strong convexity and subsequent analysis. See Negahban et al. (2009) for more details. Below is one example of a decomposable norm.

Example 1. For a subset $S \subseteq \{1, 2, ..., p\}$ with cardinality s, we define the subspace

$$A(S) := \{\beta \in \mathbb{R}^p : \ \beta_j = 0 \ for \ all \ j \notin S\}$$

Let $B(S) = A(S)$, so that the orthogonal complement is given by

$$B^{\perp}(S) = A^{\perp}(S) = \{\beta \in \mathbb{R}^p : \ \beta_j = 0 \ for \ all \ j \in S\}$$

Then the ℓ_1-norm regularizer $r(u) = ||u||_1$ is decomposable with respect to the pair $(A(S), B^{\perp}(S))$. Indeed, by construction of the subspaces, for any $u \in A(S)$ and $v \in B^{\perp}(S)$, we have

$$||u + v||_1 = ||(u_S, 0) + (0, v_{S^c})||_1 = ||u||_1 + ||v||_1$$

Restricted strong convexity

We now move to define the RSC property of the loss function. We know that convexity, especially strong convexity defined below is very important in convex optimization problems. We say that loss function $\mathcal{L}(\beta)$ is strongly convex if and only if there exists a positive number γ,

$$\mathcal{L}(\beta + \Delta) - \mathcal{L}(\beta) - \langle \nabla \mathcal{L}(\beta), \Delta \rangle \geq \gamma ||\Delta||_2^2, \text{ for all } \beta \text{ and } \Delta.$$

The strong convexity makes the solution of $\min_\beta \mathcal{L}(\beta)$ unique and when $\hat{\beta}$ is close to β^*, $\mathcal{L}(\hat{\beta})$ is close to $\mathcal{L}(\beta^*)$ and vice verser. Unfortunately, when $p > n$, it is impossible to guarantee strong convexity for all directions. In fact, based on the consequence of decomposability of the regularizer $r(\cdot)$, the error belongs to a restricted set $\mathcal{C}$ defined in Equation (2.3), hence it is natural to require that the strong convexity holds for the restricted set of directions. That is, we may instead suppose that the loss function satisfies a form of restricted strong convexity (RSC). Let

$$\mathcal{K}(\delta; A, B, \beta^*) := \mathcal{C}(A, B, \beta^*) \cap \{\Delta \in \mathbb{R}^p : \ ||\Delta||_2 = \delta\} \tag{2.4}$$

where $\delta > 0$ is a tolerance parameter.

Definition 3. *The loss function satisfies the RSC$(\delta, \gamma; A, B, \beta^*)$ condition if*

$$\mathcal{L}(\beta^* + \Delta) - \mathcal{L}(\beta^*) - \langle \nabla \mathcal{L}(\beta^*), \Delta \rangle \geq \gamma ||\Delta||_2^2 \quad \textit{for all } \Delta \in \mathcal{K}(\delta; A, B, \beta^*) \tag{2.5}$$

With the decomposable regularizer $r(\cdot)$ and a loss function $\mathcal{L}(\beta)$ with restricted strong convexity, we can now introduce the main result in Negahban et al. (2009). Define

$$\Psi(B) := \inf\{c > 0 : r(u) \leq c\|u\|_2, \text{ for all } u \in B\},$$
$$r^*(v) := \sup_{u \in \mathbb{R}^p \setminus \{0\}} \frac{\langle u, v \rangle}{r(u)}.$$

Theorem 1 (Negahban et al. (2009).)**.** *Suppose that the regularizer $r(\beta)$ is decomposable with respect to the subspace pair $(A, B^\perp)$ with $A \subseteq B$. Consider the regularized M-estimator defined in (2.1) with a convex and differentiable loss function and a strictly positive $\lambda_n \geq 2r^*(\nabla \mathcal{L}(\beta^*))$. Define the critical tolerance*

$$\delta_n := \inf_{\delta > 0} \left\{ \delta : \ \delta \geq \frac{2\lambda_n}{\gamma} \Psi(B) + \sqrt{\frac{2\lambda_n r(\beta^*_{A^\perp})}{\gamma}} \quad \textit{and (2.5) holds} \right\} \tag{2.6}$$

then any optimal solution $\hat{\beta}$ to the convex program (2.1) satisfies the bound $||\hat{\beta} - \beta^*||_2 \leq \delta_n$.

With the tool established above, we are now ready to study the property of the naiveEN and the Elastic Net estimators.

2.2. Elastic Net

The regularizer corresponding to naiveEN defined as (1.5) is $r(\beta) = ||\beta||_1 + \frac{\lambda_{2n}}{\lambda_{1n}}||\beta||_2^2$, which is not a decomposable norm. Fortunately, we can transform the naive Elastic Net estimator and Elastic Net estimator to ℓ_1 regularized M-estimators by the following proposition.

Lemma 1. *Given data (y, X) and $(\lambda_{1n}, \lambda_{2n})$, the naiveEN and Elastic Net estimators are given by*

$$\hat{\beta}(naive) = \underset{\beta}{argmin} \left\{ \beta^T \left(\frac{1}{n} X^T X + \lambda_{2n} I\right)\beta - \frac{2}{n} y^T X \beta + \lambda_{1n} ||\beta||_1 \right\}, \tag{2.7}$$

$$\hat{\beta}(EN) = \underset{\beta}{argmin} \left\{ \beta^T \left(\frac{\frac{1}{n} X^T X + \lambda_{2n} I}{1 + \lambda_{2n}}\right)\beta - \frac{2}{n} y^T X \beta + \lambda_{1n} ||\beta||_1 \right\}. \tag{2.8}$$

The proof of Lemma 1 can be found in Zou and Hastie (2005), but for completeness we provide it in the Appendix.

It is easy to see that

$$\hat{\beta}(Lasso) = \underset{\beta}{argmin} \left\{ \beta^T (\frac{1}{n} X^T X)\beta - \frac{2}{n} y^T X\beta + \lambda_{1n}||\beta||_1 \right\}. \tag{2.9}$$

Hence, all these three estimators belong to a special type of M-estimate as follows

$$\hat{\beta} = \underset{\beta}{argmin} \left\{ \beta^T W\beta - \frac{2}{n} y^T X\beta + \lambda_{1n}||\beta||_1 \right\}. \tag{2.10}$$

Moreover, Proposition 1 interprets the naiveEN and Elastic Net as modified versions of the lasso. The loss function corresponding to naiveEN and Elastic Net, $\beta^T(\frac{1}{n}X^TX + \lambda_{2n}I)\beta - \frac{2}{n}y^TX\beta$ and $\beta^T(\frac{\frac{1}{n}X^TX+\lambda_{2n}I}{1+\lambda_{2n}})\beta - \frac{2}{n}y^TX\beta$ respectively, are strictly convex, while the loss function of Lasso, $\beta^T(\frac{1}{n}X^TX)\beta - \frac{2}{n}y^TX\beta$, is not strictly convex in high-dimensional settings. But the curvature of the loss function depends on the magnitude of the tuning parameter λ_{2n}; if λ_{2n} is small, the advantage of Elastic Net over lasso maybe minor for parameter estimation.

General convergence rate for Elastic Net

Since we already transformed the navieEN and Elastic Net problem to the framework of M-estimator with a decomposable regularizer, we can apply the tool described in Section 2.1 to obtain a general result on the bound of estimation error $\|\hat{\beta} - \beta^*\|_2$.

Consider problem (2.10) which can be the Lasso, the naiveEN and the Elastic Net depending on the choice of W. The loss function is $\mathcal{L}(\beta) = \beta^T W\beta - \frac{2}{n}y^TX\beta$, where W is a nonnegative definite matrix. Observe that this type of loss function is twice-continuously differentiable and the Hessian is $\nabla^2\mathcal{L}(\beta) = 2W$. As a result, the RSC$(\delta, \gamma; A, B, \beta^*)$ condition defined via (2.5) has the following form

$$\Delta^T W\Delta \geq \gamma||\Delta||_2^2 \quad \text{for all} \quad \Delta \in \mathcal{K}(\delta; A, B, \beta^*). \tag{2.11}$$

Applying the result of Theorem 1, we have the general convergence rate for the Elastic Net estimator. Let

$$\tilde{\Psi}(B) := \underset{u\in B\setminus\{0\}}{sup} \frac{||u||_1}{||u||_2} \tag{2.12}$$

Proposition 1. *Suppose that the regularizer $r(\beta) = ||\beta||_1$ is decomposable with respect to the subspace pair $(A, B^\perp)$ with $A \subseteq B$. Consider the regularized M-estimator defined in (2.10) with a nonnegative definition matrix W and a strictly positive $\lambda_{1n} \geq 4||W\beta^* - \frac{1}{n}X^T y||_\infty$. Define the critical tolerance*

$$\delta_n := \inf_{\delta>0} \left\{ \delta : \ \delta \geq \frac{2\lambda_{1n}}{\gamma}\tilde{\Psi}(B) + \sqrt{\frac{2\lambda_{1n}||\beta^*_{A^\perp}||_1}{\gamma}} \text{ and (2.11) holds} \right\} \tag{2.13}$$

then any optimal solution $\hat{\beta}$ to problem (2.10) satisfies the bound $||\hat{\beta} - \beta^*||_2 \leq \delta_n$.

With the general results, we are now ready to establish conditions and discuss how to choose suitable λ_{1n} and λ_{2n} such that the Elastic Net estimator is ℓ_2 consistent: $||\hat{\beta} - \beta^*||_2 \to 0$ when $n \to \infty$.

3. Convergence Rates under Sparsity Assumption

In the high-dimensional setting, i.e., when $p \gg n$, the linear regression model (1.1) is unidentifiable, since the rank of the design matrix X is at most $n \ll p$. In order to obtain an identifiable and interpretable model, we have to impose additional assumption on the regression coefficients β^*. In this section, we study the convergence rates of the Elastic Net under two types of sparsity assumption: (1) exact sparsity, which assumes that β^* has at most s non-zero coefficients; and (2) weak sparsity, which assumes that $\beta^* \in \mathbb{B}_q(R_q)$ for some $q \in (0, 1]$.

Convergence rates under exact sparsity assumption

Let $S = \left\{ j \in \{1, 2, ..., p\} : \ \beta^*_j \neq 0 \right\}$ be the support of β^*, with cardinality $|S| = s$. We note that β^* belongs to subspace $A(S) := \{\beta \in \mathbb{R}^p : \ \beta_j = 0 \text{ for all } j \notin S\}$. By choosing $B(S) = A(S)$, the restricted set $\mathcal{C}(A, B, \beta^*)$ is given by

$$\mathcal{C}(A, B, \beta^*) := \{\Delta \in \mathbb{R}^P : ||\Delta_{S^c}||_1 \leq 3||\Delta_S||_1\}. \tag{3.1}$$

As a consequence, the RSC condition is equivalent to a type of restricted eigenvalue condition on the matrix W,

$$\Delta^T W \Delta \geq \gamma||\Delta||_2^2 \quad \textit{for all } \Delta \in R^p \textit{ such that } ||\Delta_{S^c}||_1 \leq 3||\Delta_S||_1. \tag{3.2}$$

We still require two additional assumptions to get explicit convergence rates.

Letting $X_j \in \mathbb{R}^n$ be the j^{th} column of X, we require

$$\frac{||X_j||_2}{\sqrt{n}} \leq 1 \quad \forall j = 1, 2, ..., p \tag{3.3}$$

In addition, we assume that the noise vector $\epsilon \in \mathbb{R}^n$ is zero-mean and has sub-Gaussian tails, i.e., there is a constant $\sigma > 0$, such that for any fixed $||u||_2 = 1$,

$$P(|\langle u, \epsilon \rangle| \geq t) \leq 2\exp(-\frac{t^2}{2\sigma^2}) \quad \forall t > 0 \tag{3.4}$$

Under these conditions, we can obtain the following corollaries of Proposition 1.

Corollary 1. *Consider the linear regression model (1.1) with the true parameter $\beta^* \in \mathbb{R}^p$ exactly s-sparse and assume the conditions (3.3), (3.4), and the following conditions (a) and (b) hold.*
(a) restricted eigenvalue condition: there exists a constant $\gamma > 0$ such that

$$||X\beta||_2^2/n \geq (\gamma - \lambda_{2n})||\beta||_2^2 \quad \forall \beta \in R^p, ||\beta_{S^c}||_1 \leq 3||\beta_S||_1; \tag{3.5}$$

(b) the tuning parameters satisfy

$$\lambda_{1n} = 8\sigma\sqrt{\frac{\log p}{n}} \quad \textit{and} \quad 16\lambda_{2n}(max_j|\beta_j^*|) \leq \lambda_{1n}, \tag{3.6}$$

then with probability at least $1 - c_1\exp(-c_2 n\lambda_{1,n}^2)$ for some constants $c_1 > 0, c_2 > 0$, the naiveEN $\hat{\beta}$ satisfies

$$||\hat{\beta} - \beta^*||_2 \leq \frac{16\sigma}{\gamma}\sqrt{\frac{s\log p}{n}}. \tag{3.7}$$

We point out that Bunea (2008) and Hebiri and Van de Geer (2011) also have results to bound estimation error of the naiveEN estimator. Bunea (2008) considers to bound ℓ_1 error $\|\hat{\beta} - \beta^*\|_1$; Hebiri and Van de Geer (2011) considers to bound ℓ_1 error $\|\hat{\beta} - \beta^*\|_1$ and some other forms of error, such as $\|X\hat{\beta} - X\beta^*\|_2^2$. In both papers, restricted eigenvalue condition is a key assumption.

Similarly, for the Elastic Net estimator we have the following result:

Corollary 2. *Consider the linear regression model (1.1) with the true parameter $\beta^* \in \mathbb{R}^p$ exactly s-sparse and assume the conditions (3.3), (3.4), and the following conditions (a), (b) hold.*
(a) restricted eigenvalue condition: there exists a constant $\gamma > 0$ such that

$$||X\beta||_2^2/n \geq (\gamma(1 + \lambda_{2n}) - \lambda_{2n})||\beta||_2^2 \quad \forall \beta \in R^p, ||\beta_{S^c}||_1 \leq 3||\beta_S||_1 \tag{3.8}$$

(b) the tuning parameters satisfy

$$\lambda_{1n} = 8\sigma\sqrt{\frac{\log p}{n}} \quad \textit{and} \quad 16\left\|\frac{\lambda_{2n}}{1+\lambda_{2n}}(I - \frac{1}{n}X^TX)\beta^*\right\|_\infty \leq \lambda_{1n}, \tag{3.9}$$

then with probability at least $1 - c_1\exp(-c_2 n\lambda_{1n}^2)$ *for some constants* $c_1 > 0, c_2 > 0$*, the Elastic Net estimator* $\hat{\beta}$ *satisfies*

$$||\hat{\beta} - \beta^*||_2 \leq \frac{16\sigma}{\gamma}\sqrt{\frac{s\log p}{n}}. \tag{3.10}$$

The proof of these corollaries can be found in Appendix. We now compare the results for the Elastic Net with the Lasso. In order to do this comparison, we present the following known result (see Bickel et al. 2009, Meinshausen and Yu 2009, and Van de Geer 2007) which is also a corollary of proposition 2.

Corollary 3. *Consider the linear regression model* (1.1) *with the true parameter* $\beta^* \in \mathbb{R}^p$ *exactly* s*-sparse and assume the conditions* (3.3), (3.4)*, and the following restricted eigenvalue condition hold.*
restricted eigenvalue condition: there exists a constant $\gamma > 0$ *such that*

$$||X\beta||_2^2/n \geq \gamma||\beta||_2^2, \quad \forall\beta \in \mathbb{R}^p, ||\beta_{S^c}||_1 \leq 3||\beta_S||_1. \tag{3.11}$$

If we solve the Lasso with $\lambda_{1n} = 8\sigma\sqrt{\frac{\log p}{n}}$*, then with probability at least* $1 - c_1\exp(-c_2 n\lambda_{1n}^2)$ *for some constants* $c_i > 0$ $(i = 1, 2)$*, the Lasso estimate* $\hat{\beta}$ *satisfies*

$$||\hat{\beta} - \beta^*||_2 \leq \frac{16\sigma}{\gamma}\sqrt{\frac{s\log p}{n}}. \tag{3.12}$$

Remarks: All three estimators (naiveEN, Elastic Net and the Lasso) have the same convergence rate. But the conditions for naiveEN and Elastic Net are a little bit weaker in the sense that when conditions for the Lasso in Corollary 3 hold, conditions for naiveEN or Elastic Net in Corollary 1 and Corollary 2 hold. This makes sense, because the Lasso can be thought as a model nested in naive Elastic net or Elastic net model. But if we look more into the conditions, we find that for the resulting convergence rate, both naiveEN and Elastic Net needs λ_2 small enough, see Conditions (3.6) and (3.9). Both conditions suggest that if $\max_j |\beta_j| < \infty$, λ_{2n} should be chosen in the same order as (or higher order than) λ_{1n}. Because of this small λ_{2n}, the advantage of the naiveEN or Elastic Net over

Lasso is minor for parameter estimation. This phenomenon is also found in Jia and Yu (2010) and Bunea (2008).

From the error bounds obtain in Corollaries 1, 2 and 3, we can have estimation consistency for these estimators in high-dimensional settings: in addition to conditions in Corollaries 1, 2 and 3 respectively, if we further have $\frac{s\log p}{n} \to 0$ then the three estimators–Lasso, naiveEN, and Elastic Net, are estimation consistence, i.e. $||\hat{\beta} - \beta^*||_2 \to_p 0, as\ n \to \infty$.

We've discussed convergence rates under exact sparsity assumption. It is well known that in practice, there is no true model. Even when β^* is not exact sparse, sometimes a sparse vector can be used to well approximate it. In this situation, regularized method could give a good estimator, that is $\|\hat{\beta} - \beta^*\|_2 \to 0$. We next discuss the case when β^* belongs to an ℓ_q ball, for some $q \in (0, 1]$.

Convergence rates under weakly sparsity assumption

We now consider a weak sparsity assumption based on imposing a certain decay rate on the ordered entries of β^*, i.e. $\beta^* \in \mathbb{B}_q(R_q)$ for some $q \in (0, 1]$. One suitable condition for this case is that there exist constants $\kappa_1, \kappa_2 > 0$ such that

$$\frac{||X\beta||_2}{\sqrt{n}} \geq \sqrt{\kappa_1}||\beta||_2 - \sqrt{\frac{\kappa_2 \log p}{n}}||\beta||_1, \quad \forall \beta \in \mathbb{R}^p. \tag{3.13}$$

It has been shown by Raskutti et al. (2011) that, when X has each row drawn i.i.d. from a $N(0, \Sigma)$ distribution, then there exist constants $\kappa_1 > 0$ and $\kappa_2 > 0$ depending only on Σ such that property (3.13) holds with probability at least $1 - c\exp(-c'n)$ for some constants $c > 0$ and $c' > 0$. We point out that Condition (3.13) in fact implies the RSC condition (see Negahban et al. (2009) and our proof of Corollary 6 in the Appendix). Now we have the following corollaries derived from Proposition 1.

Corollary 4. *Consider the linear regression model (1.1) with the true parameter $\beta^* \in \mathbb{B}_q(R_q)$ for some $q \in (0, 1]$. Assume $\sqrt{R_q}(\frac{\log p}{n})^{\frac{1}{2}-\frac{q}{4}} = o(1)$ and conditions (3.3), (3.4), and (3.13) hold. If we solve the Lasso with $\lambda_{1n} = 8\sigma\sqrt{\frac{\log p}{n}}$, then there are universal constants $c_1, c_2 > 0$ such that the Lasso estimate satisfies*

$$||\hat{\beta} - \beta^*||_2 \leq 38\sqrt{R_q}(\frac{\sigma^2}{\kappa_1^2}\frac{\log p}{n})^{\frac{1}{2}-\frac{q}{4}} \tag{3.14}$$

with probability at least $1 - c_1\exp(-c_2 n\lambda_{1n}^2)$.

This result has been obtained in Negahban et al. (2009). We state it here to compare it with the naiveEN and Elastic Net. For the naiveEN and the Elastic Net, we have the following results.

Corollary 5. *Consider the linear regression model (1.1) with the true parameter $\beta^* \in \mathbb{B}_q(R_q)$ for some $q \in (0,1]$. Assume $\sqrt{R_q}(\frac{\log p}{n})^{\frac{1}{2}-\frac{q}{4}} = o(1)$ and conditions (3.3), (3.4), and (3.13) hold. If we solve the naiveEN with λ_{1n} and λ_{2n} satisfying*

$$\lambda_{1n} = 8\sigma\sqrt{\frac{\log p}{n}} \quad \text{and} \quad 16\lambda_{2n}(max_j|\beta_j^*|) \leq \lambda_{1n} \tag{3.15}$$

then there are universal constants $c_1, c_2 > 0$ such that the naiveEN estimate satisfies

$$||\hat{\beta} - \beta^*||_2 \leq 38\sqrt{R_q}(\frac{\sigma^2}{4\gamma^2}\frac{\log p}{n})^{\frac{1}{2}-\frac{q}{4}}, \tag{3.16}$$

with probability at least $1 - c_1\exp(-c_2 n\lambda_{1,n}^2)$, where $\gamma = \frac{\kappa_1}{2} + \lambda_{2n}$.

Corollary 6. *Consider the linear regression model (1.1) with the true parameter $\beta^* \in \mathbb{B}_q(R_q)$ for some $q \in (0,1]$, and Assume $\sqrt{R_q}(\frac{\log p}{n})^{\frac{1}{2}-\frac{q}{4}} = o(1)$ and conditions(3.3), (3.4), and (3.13) hold. If we solve the Elastic Net with λ_{1n} and λ_{2n} satisfying*

$$\lambda_{1n} = 8\sigma\sqrt{\frac{\log p}{n}} \quad \text{and} \quad 16||\frac{\lambda_{2n}}{1+\lambda_{2n}}(I - \frac{1}{n}X^TX)\beta^*||_\infty \leq \lambda_{1n} \tag{3.17}$$

then there are universal constants $c_1, c_2 > 0$ such that the Elastic Net estimate satisfies

$$||\hat{\beta} - \beta^*||_2 \leq 38\sqrt{R_q}(\frac{\sigma^2}{4\gamma^2}\frac{\log p}{n})^{\frac{1}{2}-\frac{q}{4}} \tag{3.18}$$

with probability at least $1 - c_1\exp(-c_2 n\lambda_{1n}^2)$, where $\gamma = \frac{\kappa_1/2+\lambda_{2n}}{1+\lambda_{2n}}$

From these results for weak sparsity assumption, we see that for suitable choice of λ_{2n}, the naive Elastic Net and the Elastic Net have the same convergence rate as the Lasso. Compare the error bounds (3.16) and (3.18) with (3.14), we see that for the naiveEN and Elastic Net, the error bounds are better than that of the Lasso. Once again, because the scale of λ_{2n} is the same order as (or higher order) than λ_{1n}, the advantages over the Lasso is minor.

4. Conclusion

We have discussed the estimation performance of the Elastic Net in high-dimensional settings under the unified framework proposed by Negahban et al. (2009). More precisely, we provide the ℓ_2-estimation error bounds for the naiveEN and Elastic Net for exact and weak sparse linear models. For exactly sparse linear models, we show that under a key assumption of restricted eigenvalue condition and some other minor conditions, with suitable choices of λ_{1n} and λ_{2n}, all three estimators–the Lasso, naiveEN and the Elastic Net are all estimation consistency. Results also show that the conditions for the naiveEN and the Elastic Net are slightly weaker than that needed for the Lasso. For weakly sparse linear models, the key condition is

$$\frac{||X\beta||_2}{\sqrt{n}} \geq \sqrt{\kappa_1}||\beta||_2 - \sqrt{\frac{\kappa_2 \log p}{n}}||\beta||_1, \quad \forall \beta \in \mathbb{R}^p.$$

This condition is not strong. Since for a design matrix X with each row drawn i.i.d. from a $N(0, \Sigma)$ distribution, there exist constants $\kappa_1 > 0$ and $\kappa_2 > 0$ depending only on Σ such that the above property holds with probability at least $1 - c\exp(-c'n)$ for some constants $c > 0$ and $c' > 0$ (Raskutti et al. (2011)). We also find that under the same conditions, with suitable choices of λ_{2n}, the Elastic Net estimator can have better upper error bounds than the Lasso estimator.

Acknowledgment The authors are very grateful to Professor Bin Yu for many insightful comments and helpful advices, which lead to a substantially improved manuscript. This research is partially supported by a grant from MSRA.

Appendix: Proofs

Proof of Lemma 1

By definition (1.5) and simple algebraic calculation we have

$$\begin{aligned}
& \hat{\beta}(naive) \\
= \; & \underset{\beta}{argmin}\{\frac{1}{n}||y - X\beta||_2^2 + \lambda_{1n}||\beta||_1 + \lambda_{2n}||\beta||_2^2\} \\
= \; & \underset{\beta}{argmin}\{\frac{1}{n}y^T y - \frac{2}{n}y^T X\beta + \beta^T(\frac{1}{n}X^T X)\beta + \lambda_{2n}\beta^T I\beta + \lambda_{1n}||\beta||_1\} \\
= \; & \underset{\beta}{argmin}\{\beta^T(\frac{1}{n}X^T X + \lambda_{2n}I)\beta - \frac{2}{n}y^T X\beta + \lambda_{1n}||\beta||_1\}
\end{aligned}$$

and

$$\begin{aligned}
& \hat{\beta}(EN) = (1+\lambda_{2n})\hat{\beta}(naive) \\
= \; & (1+\lambda_{2n}) \underset{\beta}{argmin}\{\beta^T(\frac{1}{n}X^TX + \lambda_{2n}I)\beta - \frac{2}{n}y^TX\beta + \lambda_{1n}||\beta||_1\} \\
= \; & \underset{\beta}{argmin}\{(\frac{\beta}{1+\lambda_{2n}})^T(\frac{1}{n}X^TX + \lambda_{2n}I)(\frac{\beta}{1+\lambda_{2n}}) - \frac{2}{n}y^TX\frac{\beta}{1+\lambda_{2n}} + \lambda_{1n}||\frac{\beta}{1+\lambda_{2n}}||_1\} \\
= \; & \underset{\beta}{argmin}\{\beta^T(\frac{\frac{1}{n}X^TX + \lambda_{2n}I}{1+\lambda_{2n}})\beta - \frac{2}{n}y^TX\beta + \lambda_{1n}||\beta||_1\}
\end{aligned}$$

Proof of Proposition 1

By definition (2.10), we have

$$\begin{aligned}
\hat{\beta} &= \underset{\beta}{argmin}\{\beta^TW\beta - \frac{2}{n}y^TX\beta + \lambda_{1n}||\beta||_1\} \\
&= \underset{\beta}{argmin}\{\mathcal{L}(\beta) + \lambda_{1n}||\beta||_1\}
\end{aligned}$$

where $\mathcal{L}(\beta) = \beta^TW\beta - \frac{2}{n}y^TX\beta$ is obvious a convex and differentiable function and $\nabla\mathcal{L}(\beta) = 2(W\beta - \frac{1}{n}X^Ty)$, then

$$\begin{aligned}
& \mathcal{L}(\beta^* + \Delta) - \mathcal{L}(\beta^*) - \langle\nabla\mathcal{L}(\beta^*), \Delta\rangle \\
= \; & (\beta^* + \Delta)^TW(\beta^* + \Delta) - \frac{2}{n}y^TX(\beta^* + \Delta) - \{(\beta^*)^TW\beta^* - \frac{2}{n}y^TX\beta^*\} - 2\langle W\beta^* - \frac{1}{n}X^Ty, \Delta\rangle \\
= \; & \Delta^TW\Delta
\end{aligned}$$

by condition (2.11), we have

$$\mathcal{L}(\beta^* + \Delta) - \mathcal{L}(\beta^*) - \langle\nabla\mathcal{L}(\beta^*), \Delta\rangle \geq \gamma||\Delta||_2^2 \quad for\ all\ \Delta \in \mathcal{K}(\delta; A, B, \beta^*) \quad (5.1)$$

hence, condition (2.11) indicates that the RSC$(\delta, \gamma; A, B, \beta^*)$ holds. We also assume that the regularizer $r(\beta) = ||\beta||_1$ is decomposable with respect to the subspace pair $(A, B^\perp)$, and

$$\lambda_{1,n} \geq 4||W\beta^* - \frac{1}{n}X^Ty||_\infty = 2||\nabla\mathcal{L}(\beta^*)||_\infty \quad (5.2)$$

then with application of Theorem 1 (Negahban et al. (2009).), we have any optimal solution $\hat{\beta}$ satisfies the bound $||\hat{\beta} - \beta^*||_2 \leq \delta_n$, where the critical tolerance

δ_n is defined by (2.13).

Proof of Corollary 1-3

The proof of these corollaries are very similar, hence we only prove corollary 2 here, the other two can be verified with very few modifications. Let $A(S)$ and $A^{\perp}(S)$ be the model subspace and its orthogonal complement

$$A(S) = \{\beta \in R^p \mid \beta_j = 0 \quad for\ all\ j \notin S\} \tag{5.3}$$

$$A^{\perp}(S) = \{\beta \in R^p \mid \beta_j = 0 \quad for\ all\ j \in S\} \tag{5.4}$$

By example 1 we know that the norm $||\cdot||_1$ is decomposable with respect to $(A(S), A^{\perp}(S))$. By lemma 1, the Elastic Net estimate has the form

$$\begin{aligned} \hat{\beta}(EN) &= \underset{\beta}{argmin}\{\beta^T(\frac{\frac{1}{n}X^TX + \lambda_{2n}I}{1+\lambda_{2n}})\beta - \frac{2}{n}y^TX\beta + \lambda_{1n}||\beta||_1\} \\ &= \underset{\beta}{argmin}\{\beta^TW\beta - \frac{2}{n}y^TX\beta + \lambda_{1n}||\beta||_1\} \end{aligned} \tag{5.5}$$

where $W = (\frac{\frac{1}{n}X^TX+\lambda_{2n}I}{1+\lambda_{2n}})$ is a nonnegative definite matrix. For any $\Delta \in R^p, ||\Delta_{S^c}||_1 \leq 3||\Delta_S||_1$

$$\begin{aligned} \Delta^TW\Delta &= \frac{1}{1+\lambda_{2n}}(\frac{1}{n}||X\Delta||_2^2 + \lambda_{2n}||\Delta||_2^2) \\ &\geq \frac{1}{1+\lambda_{2n}}\{(\gamma(1+\lambda_{2n}) - \lambda_{2n})||\Delta||_2^2 + \lambda_{2n}||\Delta||_2^2\} = \gamma||\Delta||_2^2 \end{aligned}$$

where the inequality holds because the restricted eigenvalue condition (3.8) holds. By proposition 1, if $\lambda_{1n} \geq 4||W\beta^* - \frac{1}{n}X^Ty||_{\infty}$, we have

$$||\hat{\beta} - \beta^*||_2 \leq \frac{2\lambda_{1n}}{\gamma}\tilde{\Psi}(A(S)) + \sqrt{\frac{2\lambda_{1n}||\beta^*_{S^c}||_1}{\gamma}} = \frac{2\lambda_{1n}}{\gamma}\tilde{\Psi}(A(S)) \tag{5.6}$$

where the equality holds because of the exactly sparsity assumption. It is easy to verify

$$\tilde{\Psi}(A(S)) = \underset{u \in A(S)\setminus\{0\}}{sup} ||u||_1/||u||_2 = \sqrt{s} \tag{5.7}$$

Consequently, we need to verify that

$$\lambda_{1n} \geq 4||W\beta^* - \frac{1}{n}X^Ty||_{\infty} = 4||\frac{\lambda_{2n}}{1+\lambda_{2n}}(I - \frac{1}{n}X^TX)\beta^* - \frac{1}{n}X^T\epsilon||_{\infty} \tag{5.8}$$

with high probability. Using condition (3.3) and sub-Gaussian condition (3.4), we have the tail bound $P(|\langle X_j, \epsilon\rangle|/n \geq t) \leq 2\exp(-\frac{nt^2}{2\sigma^2})$. By union bound, we conclude that

$$P(||\frac{1}{n}X^T\epsilon||_\infty \geq t) \leq 2\exp(-\frac{nt^2}{2\sigma^2} + \log p) \tag{5.9}$$

Setting $t^2 = \frac{9\sigma^2 \log p}{4n}$, and recall that our choice of $\lambda_{1n} = 8\sigma\sqrt{\frac{\log p}{n}}$, then we have

$$P(||\frac{1}{n}X^T\epsilon||_\infty \geq \frac{3}{16}\lambda_{1n}) \leq 2\exp(-\frac{1}{8}\log p) = 2\exp(-c_2 n\lambda_{1n}^2) \tag{5.10}$$

Moreover, λ_{2n} satisfies $16||\frac{\lambda_{2n}}{1+\lambda_{2n}}(I - \frac{1}{n}X^TX)\beta^*||_\infty \leq \lambda_{1n}$, then

$$\begin{aligned} & P(\lambda_{1n} \geq 4||W\beta^* - \frac{1}{n}X^Ty||_\infty) \\ \geq\ & P(\lambda_{1n} \geq 4||\frac{\lambda_{2n}}{1+\lambda_{2n}}(I - \frac{1}{n}X^TX)\beta^*||_\infty + 4||\frac{1}{n}X^T\epsilon||_\infty) \\ \geq\ & P(\frac{3}{4}\lambda_{1n} \geq 4||\frac{1}{n}X^T\epsilon||_\infty) \end{aligned} \tag{5.11}$$

therefore, with probability at least $1 - c_1\exp(-c_2 n\lambda_{1n}^2)$, we have

$$||\hat{\beta} - \beta^*||_2 \leq \frac{2\lambda_{1n}}{\gamma}\tilde{\Psi}(A(S)) = \frac{16\sigma}{\gamma}\sqrt{\frac{s\log p}{n}}$$

that yields corollary 2.

Proof of Corollary 4-6

Once again, we only give a proof of corollary 6 here, the other two can be verified similarly. Let us first show that the RSC condition (2.11) holds when the ℓ_2-norm $||\hat{\beta} - \beta^*||_2$ is sufficiently large. For a threshold $\tau > 0$ to be chosen, define the threshold set

$$S_\tau := \{j \in \{1, 2, ..., p\} \mid |\beta_j^*| > \tau\} \tag{5.12}$$

Since $\beta^* \in \mathbb{B}_q(R_q)$, we have

$$R_q \geq \sum_{j=1}^{p}|\beta_j^*|^q \geq \sum_{j\in S_\tau}|\beta_j^*|^q \geq \tau^q|S_\tau| \tag{5.13}$$

hence the cardinality of S_τ can be upper bounded in terms of the threshold τ and l_q-ball radius R_q, to be specific, we have $|S_\tau| \leq \tau^{-q}R_q$ for any $\tau > 0$. Moreover, let S_τ^c denote the complementary set $S_\tau \setminus \{1, 2, ..., p\}$, we have

$$||\beta_{S_\tau^c}^*||_1 = \sum_{j\in S_\tau^c}|\beta_j^*| = \sum_{j\in S_\tau^c}|\beta_j^*|^q|\beta_j^*|^{1-q} \leq R_q\tau^{1-q} \tag{5.14}$$

Define the subspaces $A(S_\tau)$, $A^\perp(S_\tau)$ and $B(S_\tau)$ as previous in Example 1 with $S = S_\tau$ and let the tolerance parameter δ^* to be

$$\delta^* := 32\sqrt{\frac{\kappa_2}{\kappa_1}\frac{\log p}{n}}R_q\tau^{1-q} \tag{5.15}$$

Next, we will show that the RSC$(\delta^*, \gamma; A(S_\tau), B(S_\tau), \beta^*)$ condition holds with $\gamma = \frac{1}{1+\lambda_{2n}}(\frac{\kappa_1}{2} + \lambda_{2n})$ and $\tau = \frac{\lambda_{1n}}{16\gamma}$. Consider the constrained set $\mathcal{C}$ takes the form

$$\mathcal{C}(A, B, \beta^*) := \{\Delta \in R^p \mid ||\Delta_{S_\tau^c}||_1 \leq 3||\Delta_{S_\tau}||_1 + 4||\beta^*_{S_\tau^c}||_1\} \tag{5.16}$$

For any Δ in the set $\mathcal{C}$, we have

$$\begin{aligned} ||\Delta||_1 &\leq 4||\Delta_{S_\tau}||_1 + 4||\beta^*_{S_\tau^c}||_1 \\ &\leq 4\sqrt{|S_\tau^c|}||\Delta||_2 + 4R_q\tau^{1-q} \\ &\leq 4\sqrt{R_q}\tau^{-q/2}||\Delta||_2 + 4R_q\tau^{1-q} \end{aligned}$$

Therefore, for any $\Delta \in \mathcal{C}$, condition (3.13) implies that

$$\begin{aligned} \frac{||X\Delta||_2}{\sqrt{n}} &\geq \sqrt{\kappa_1}||\Delta||_2 - 4\sqrt{\frac{\kappa_2 \log p}{n}}\{\sqrt{R_q}\tau^{-q/2}||\Delta||_2 + R_q\tau^{1-q}\} \\ &\geq ||\Delta||_2\{\sqrt{\kappa_1} - 4\sqrt{\frac{\kappa_2 R_q \log p}{n}}\tau^{-q/2}\} - \frac{\sqrt{\kappa_1}}{8}\delta^* \end{aligned}$$

Then, for all $\Delta \in \mathcal{K}(\delta^*; A(S_\tau), B(S_\tau), \beta^*)$, we have

$$\frac{||X\Delta||_2}{\sqrt{n}} \geq ||\Delta||_2\{\frac{7\sqrt{\kappa_1}}{8} - 4\sqrt{\frac{\kappa_2 R_q \log p}{n}}\tau^{-q/2}\} \geq \frac{\sqrt{\kappa_1}}{\sqrt{2}}||\Delta||_2 \tag{5.17}$$

the second inequality holds since $\tau = \frac{\lambda_{1n}}{16\gamma}$ and

$$4\sqrt{\frac{\kappa_2 R_q \log p}{n}}\tau^{-q/2} = 4\sqrt{\kappa_2}\sqrt{R_q}(\frac{\log p}{n})^{\frac{1}{2}-\frac{q}{4}}(\frac{2\gamma}{\sigma})^{q/2} = o(1) \tag{5.18}$$

Recall that, for the Elastic Net, $W = (\frac{\frac{1}{n}X^TX + \lambda_{2n}I}{1+\lambda_{2n}})$. Therefore, for all $\Delta \in \mathcal{K}(\delta^*; A(S_\tau), B(S_\tau), \beta^*)$,

$$\begin{aligned} \Delta^T W\Delta &= \frac{1}{1+\lambda_{2n}}(\frac{1}{n}||X\Delta||_2^2 + \lambda_{2n}||\Delta||_2^2) \\ &\geq \frac{1}{1+\lambda_{2n}}(\frac{\kappa_1}{2} + \lambda_{2n})||\Delta||_2^2 = \gamma||\Delta||_2^2 \end{aligned} \tag{5.19}$$

From above, we have verified that the RSC$(\delta^*, \gamma; A(S_\tau), B(S_\tau), \beta^*)$ condition holds with $\gamma = \frac{1}{1+\lambda_{2n}}(\frac{\kappa_1}{2} + \lambda_{2n})$ and $\tau = \frac{\lambda_{1n}}{16\gamma}$. On the other hand, for our choice of λ_{1n} and λ_{2n}, we have proved in corollary 2 that with probability at least $1 - c_1 \exp(-c_2 n \lambda_{1n}^2)$

$$\lambda_{1n} \geq 4||W\beta^* - \frac{1}{n}X^T y||_\infty \tag{5.20}$$

Finally, we may apply proposition 1 to obtain

$$\begin{aligned} ||\hat{\beta} - \beta^*||_2 &\leq \max\{\delta^*, \frac{2\lambda_{1n}}{\gamma}\Psi(A(S_\tau)) + \sqrt{\frac{2\lambda_{1n}||\beta^*_{S_\tau^c}||_1}{\gamma}}\} \\ &\leq \max\{\delta^*, 32\sqrt{R_q}(\frac{\lambda_{1n}}{16\gamma})^{1-q/2} + \sqrt{32R_q(\frac{\lambda_{1n}}{16\gamma})^{2-q}}\} \\ &\leq \max\{\delta^*, 38\sqrt{R_q}(\frac{\lambda_{1n}}{16\gamma})^{1-q/2}\} \end{aligned}$$

Under the assumption $\sqrt{R_q}(\frac{\log p}{n})^{\frac{1}{2}-\frac{q}{4}} = o(1)$, the critical tolerance δ^* in (5.15) is of lower order than the second term, so that the claim follows.

References

Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37(4)**, 1705-1732.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via l_1 and $l_1 + \ell_2$ penalization. *Electron. J. Stat.* **2**, 1153-1194.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? *J. R. Statist. Soc.* **B, 57**, 301-369.

Efron, B., Hastie, T. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407-499.

Hebiri, M. and Van de Geer, S. (2011). The Smooth-Lasso and other $l_1 + \ell_2$ penalized methods. *Electronic Journal of Statistics* **5**, 1184C1226.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

Jia, J. and Yu, B. (2010). On model selection consistency of elastic net when $p \gg n$. *Statistica Sinica* **20**, 595-611.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* **37(1)**, 246-270.

Negahban, S., Ravikumar, P., Wainwright, M. J. and Yu, B. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *In Advances in Neural Information Processing Systems* **22**, 1348-1356.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics* **9**, 319-37.

Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over l_q-balls. *IEEE Transactions on Information Theory* **57**, 6976-6994.

Van de Geer, S. (2007). The deterministic lasso. In *Proc. of Joint Statistical Meeting*.

Wainwright, M. (2007). Sharp thresholds for high-dimensional and noisy recovery of sparsity using l_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55**, 2183-2202.

Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *The Journal of Machine Learning Research* **7**, 2541-2563.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67**, 301-320.

School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China
E-mail: (lhz2009@pku.edu.cn)

School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, P.R. China
E-mail: (jzjia@math.pku.edu.cn)