Electronic Journal of Statistics
Vol. 10 (2016) 1874–1895
ISSN: 1935-7524
DOI: 10.1214/15-EJS1100

On the finite-sample analysis of Θ-estimators

Yiyuan She

Department of Statistics
Florida State University
Tallahassee, FL 32306-4330
e-mail: yshe@stat.fsu.edu

Abstract: In large-scale modern data analysis, first-order optimization methods are usually favored to obtain sparse estimators in high dimensions. This paper performs theoretical analysis of a class of iterative thresholding based estimators defined in this way. Oracle inequalities are built to show the nearly minimax rate optimality of such estimators under a new type of regularity conditions. Moreover, the sequence of iterates is found to be able to approach the statistical truth within the best statistical accuracy geometrically fast. Our results also reveal different benefits brought by convex and nonconvex types of shrinkage.

MSC 2010 subject classifications: Primary 62J07, 90C26; secondary 68Q87.
Keywords and phrases: Sparsity, thresholding, nonconvex optimization, oracle inequalities, statistical algorithmic analysis.

Received April 2015.

1. Introduction

Big data naturally arising in machine learning, biology, signal processing, and many other areas, call for the need of scalable optimization in computation. Although for low-dimensional problems, Newton or quasi-Newton methods converge fast and have efficient implementations, they typically do not scale well to high dimensional data. In contrast, *first-order* optimization methods have recently attracted a great deal of attention from researchers in statistics, computer science and engineering. They iterate based on the gradient (or a subgradient) of the objective function, and have each iteration step being cost-effective. In high dimensional statistics, a first-order algorithm typically proceeds in the following manner

$$\boldsymbol{\beta}^{(t+1)} = \mathcal{P} \circ (\boldsymbol{\beta}^{(t)} - \alpha \nabla l(\boldsymbol{\beta}^{(t)})), \tag{1}$$

where $\mathcal{P}$ is an operator that is easy to compute, ∇l denotes the gradient of the loss function l, and α gives the stepsize. Such a simple iterative procedure is suitable for large-scale optimization, and converges in arbitrarily high dimensions provided α is properly small.

$\mathcal{P}$ can be motivated from the perspective of statistical shrinkage or regularization and is necessary to achieve good accuracy when the dimensionality is moderate or high. For example, a proximity operator [11] is associated with a convex penalty function. But the problems of interest may not always be convex. Quite often, $\mathcal{P}$ is taken as a certain **thresholding rule** Θ in statistical learning, such as SCAD [5]. The resulting computation-driven estimators, which we call Θ-*estimators*, are fixed points of $\boldsymbol{\beta} = \Theta(\boldsymbol{\beta} - \nabla l(\boldsymbol{\beta}); \lambda)$. To study the non-asymptotic behavior of Θ-estimators (regardless of the sample size and dimensionality), we will establish some oracle inequalities.

During the last decade, people have performed rigorous finite-sample analysis of many high-dimensional estimators defined as *globally* optimal solutions to some convex or nonconvex problems—see [3], [19], [2], [9], [20], [14], among many others. Θ-estimators pose some new questions. First, although nicely, an associated optimization criterion can be constructed for any given Θ-estimator, the objective may not be convex, and the estimator may not correspond to any functional local (or global) minimum. Second, there are various types of Θ-estimators due to the abundant choices of Θ, but a comparative study regarding their statistical performance in high dimensions is lacking in the literature. Third, Θ-estimators are usually computed in an inexact way on big datasets. Indeed, most practitioners (have to) terminate (1) before full computational convergence. These disconnects between theory and practice when using iterative thresholdings motivate our work.

The rest of the paper is organized as follows. Section 2 introduces the Θ-estimators, the associated iterative algorithm–TISP, and some necessary notation. Section 3 presents the main results, including some oracle inequalities, and sequential analysis of the iterates generated by TISP. Section 4 provides proof details.

2. Background and notation

2.1. Thresholding functions

Definition 1 (Thresholding function)**.** *A thresholding function is a real valued function* $\Theta(t;\lambda)$ *defined for* $-\infty < t < \infty$ *and* $0 \le \lambda < \infty$ *such that (i)* $\Theta(-t;\lambda) = -\Theta(t;\lambda)$*; (ii)* $\Theta(t;\lambda) \le \Theta(t';\lambda)$ *for* $t \le t'$*; (iii)* $\lim_{t\to\infty} \Theta(t;\lambda) = \infty$*; (iv)* $0 \le \Theta(t;\lambda) \le t$ *for* $0 \le t < \infty$*.*

A vector version of Θ (still denoted by Θ) is defined componentwise if either t or λ is replaced by a vector. From the definition,

$$\Theta^{-1}(u;\lambda) := \sup\{t : \Theta(t;\lambda) \le u\}, \forall u > 0 \tag{2}$$

must be monotonically non-decreasing and so its derivative is defined almost everywhere on $(0,\infty)$. Given Θ, a critical number $\mathcal{L}_\Theta \le 1$ can be introduced such that $\mathrm{d}\Theta^{-1}(u;\lambda)/\mathrm{d}u \ge 1 - \mathcal{L}_\Theta$ for almost every $u \ge 0$, or

$$\mathcal{L}_\Theta := 1 - \operatorname{ess\,inf}\{\mathrm{d}\Theta^{-1}(u;\lambda)/\mathrm{d}u : u \ge 0\}, \tag{3}$$

where ess inf is the essential infimum. For the perhaps most popular soft-thresholding and hard-thresholding functions

$$\Theta_S(t;\lambda) = \mathrm{sgn}(t)(|t|-\lambda)_+, \quad \Theta_H(t;\lambda) = t1_{|t|\geq\lambda},$$

$\mathcal{L}_\Theta$ equals 0 and 1, respectively.

For any arbitrarily given Θ, we construct a penalty function $P_\Theta(t;\lambda)$ as follows

$$P_\Theta(t;\lambda) = \int_0^{|t|} (\Theta^{-1}(u;\lambda) - u)\,\mathrm{d}u = \int_0^{|t|} (\sup\{s : \Theta(s;\lambda) \leq u\} - u)\,\mathrm{d}u \tag{4}$$

for any $t \in \mathbb{R}$. This penalty will be used to make a proper objective function for Θ-estimators.

The threshold $\tau(\lambda) := \Theta^{-1}(0;\lambda)$ may not equal λ in general. For ease in notation, in writing $\Theta(\cdot;\lambda)$, we always assume that λ is the *threshold parameter*, i.e., $\lambda = \tau(\lambda)$, unless otherwise specified. Then an important fact is that given λ, any thresholding rule Θ satisfies $\Theta(t;\lambda) \leq \Theta_H(t;\lambda)$, $\forall t \geq 0$, due to property (iv), from which it follows that

$$P_\Theta(t;\lambda) \geq P_H(t;\lambda), \tag{5}$$

where

$$P_H(t;\lambda) = \int_0^{|t|} (\Theta_H^{-1}(u;\lambda) - u)\,\mathrm{d}u = (-t^2/2 + \lambda|t|)1_{|t|<\lambda} + (\lambda^2/2)1_{|t|\geq\lambda}. \tag{6}$$

In particular, $P_H(t;\lambda) \leq P_0(t;\lambda) := \frac{\lambda^2}{2}1_{t\neq 0}$ and $P_H(t;\lambda) \leq P_1(t;\lambda) := \lambda|t|$.

When Θ has discontinuities, such as $t = \pm\lambda$ in $\Theta_H(t;\lambda)$, ambiguity may arise in definition. To avoid the issue, we assume the quantity to be thresholded never corresponds to any discontinuity of Θ. This assumption is mild because practically used thresholding rules have few discontinuity points and such discontinuities rarely occur in real applications.

2.2. Θ-estimators

We assume a model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{7}$$

where $\boldsymbol{X}$ is an $n \times p$ design matrix, $\boldsymbol{y}$ is a response vector in $\mathbb{R}^n$, $\boldsymbol{\beta}^*$ is the unknown coefficient vector, and $\boldsymbol{\epsilon}$ is a *sub-Gaussian* random vector with mean zero and scale bounded by σ, cf. Definition 2 in Section 4 for more detail. Then a Θ-estimator $\hat{\boldsymbol{\beta}}$, driven by the computational procedure (1), is defined as a solution to the Θ-equation

$$\rho\boldsymbol{\beta} = \Theta(\rho\boldsymbol{\beta} + \boldsymbol{X}^T\boldsymbol{y}/\rho - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}/\rho;\lambda), \tag{8}$$

where ρ, the scaling parameter, does not depend on $\boldsymbol{\beta}$. Having ρ appropriately large is crucial to guarantee the convergence of the computational procedure.

All popularly used penalty functions are associated with thresholdings, such as the ℓ_r $(0 < r \leq 1)$, ℓ_2, SCAD [5], MCP [18], capped ℓ_1 [21], ℓ_0, elastic net [22], Berhu [10, 6], $\ell_0 + \ell_2$ [12], to name a few. Table 1 lists some examples. From a shrinkage perspective, thresholding rules usually suffice in statistical learning.

Equation (8) can be re-written in terms of the scaled deign $\tilde{\boldsymbol{X}} = \boldsymbol{X}/\rho$ and the corresponding coefficient vector $\tilde{\boldsymbol{\beta}} = \rho\boldsymbol{\beta}$

$$\tilde{\boldsymbol{\beta}} = \Theta(\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{X}}^T\boldsymbol{y} - \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\tilde{\beta}; \lambda). \tag{9}$$

We will show that the λ in the scaled form does not have to adjust for the sample size, which is advantageous in regularization parameter tuning.

A simple iterative procedure can be defined based on (8) or (9):

$$\tilde{\boldsymbol{\beta}}^{(t+1)} = \Theta(\tilde{\boldsymbol{\beta}}^{(t)} + \tilde{\boldsymbol{X}}^T\boldsymbol{y} - \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}^{(t)}; \lambda), \boldsymbol{\beta}^{(t+1)} = \tilde{\boldsymbol{\beta}}^{(t+1)}/\rho, \tag{10}$$

which is called the Thresholding-based Iterative Selection Procedure (**TISP**) [12]. From Theorem 2.1 of [13], given an arbitrary Θ, TISP ensures the following function-value descent property when $\rho \geq \frac{\|\boldsymbol{X}\|_2}{2-\mathcal{L}_\Theta}$:

$$f(\boldsymbol{\beta}^{(t+1)}; \lambda) \leq f(\boldsymbol{\beta}^{(t)}; \lambda). \tag{11}$$

Here, the energy function (objective function) is constructed as

$$f(\boldsymbol{\beta}; \lambda) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \sum_{j=1}^{p} P(\rho|\beta_j|; \lambda), \tag{12}$$

where the penalty P can be P_Θ as defined in (4), or more generally,

$$P(t; \lambda) = P_\Theta(t; \lambda) + q(t; \lambda), \tag{13}$$

with q an arbitrary function satisfying $q(t, \lambda) \geq 0$, $\forall t \in \mathbb{R}$ and $q(t; \lambda) = 0$ if $t = \Theta(s; \lambda)$ for some $s \in \mathbb{R}$. Furthermore, we can show that when $\rho > \|\boldsymbol{X}\|_2/(2 - \mathcal{L}_\Theta)$, any limit point of $\boldsymbol{\beta}^{(t)}$ is necessarily a fixed point of (8), and thus a Θ-estimator. See [13] for more detail. Therefore, f is not necessarily unique when Θ has discontinuities—for example, penalties like the capped ℓ_1, $P_0(t; \lambda) = \frac{\lambda^2}{2}1_{t\neq 0}$ and P_H are all associated with the same Θ_H. Because of the *many-to-one* mapping from penalty functions to thresholding functions, iterating (1) with a well-designed thresholding rule is perhaps more convenient than solving a nonconvex penalized optimization problem. Indeed, some penalties (like SCAD) are designed from the thresholding viewpoint.

The following theorem shows that the set of Θ-estimators include all locally optimal solutions of $\frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \sum_{j=1}^{p} P_\Theta(|\beta_j|; \lambda) =: f_\Theta(\boldsymbol{\beta})$.

Theorem 1. *Let $\hat{\boldsymbol{\beta}}$ be a local minimum point (or a coordinate-wise minimum point) of $f_\Theta(\cdot)$. If Θ is continuous at $\hat{\boldsymbol{\beta}} + \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}$ must satisfy $\boldsymbol{\beta} = \Theta(\boldsymbol{\beta} + \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}; \lambda)$.*

Table 1
Some examples of thresholding functions and their associated quantities

	soft	**ridge**	**hard**
Θ	$(t-\lambda\mathrm{sgn}(t))1_{\lvert t\rvert>\lambda}$	$\frac{t}{1+\eta}$	$t1_{\lvert t\rvert>\lambda}$
$\mathcal{L}_\Theta$	0	$-\eta$	1
P_Θ	$\lambda\lvert t\rvert$	$\frac{\eta}{2}t^2$	$\begin{cases} -\frac{1}{2}t^2+\lambda\lvert t\rvert, & \text{if } \lvert t\rvert<\lambda \\ \frac{1}{2}\lambda^2, & \text{if } \lvert t\rvert\geq\lambda \end{cases}$
P			$\min(\lambda\lvert t\rvert, \frac{\lambda^2}{2})$ (‘capped ℓ_1’), $\frac{\lambda^2}{2}1_{t\neq 0}$
	elastic net $(\eta\geq 0)$	**berhu** $(\eta\geq 0)$	**hard-ridge** $(\eta\geq 0)$
Θ	$\frac{t-\lambda\mathrm{sgn}(t)}{1+\eta}1_{\lvert t\rvert\geq\lambda}$	$\begin{cases} 0 & \text{if } \lvert t\rvert<\lambda \\ t-\lambda\mathrm{sgn}(t) & \text{if } \lambda\leq\lvert t\rvert\leq\lambda+\lambda/\eta \\ \frac{t}{1+\eta} & \text{if } \lvert t\rvert>\lambda+\lambda/\eta \end{cases}$	$\frac{t}{1+\eta}1_{\lvert t\rvert>\lambda}$
$\mathcal{L}_\Theta$	$-\eta$	0	1
P_Θ	$\lambda\lvert t\rvert+\frac{1}{2}\eta t^2$	$\begin{cases} \lambda\lvert t\rvert & \text{if } \lvert t\rvert\leq\lambda/\eta \\ \frac{\eta t^2}{2}+\frac{\lambda^2}{2\eta} & \text{if } \lvert t\rvert>\lambda/\eta. \end{cases}$	$\begin{cases} -\frac{1}{2}t^2+\lambda\lvert t\rvert, & \text{if } \lvert t\rvert<\frac{\lambda}{1+\eta} \\ \frac{1}{2}\eta t^2+\frac{1}{2}\frac{\lambda^2}{1+\eta}, & \text{if } \lvert t\rvert\geq\frac{\lambda}{1+\eta}. \end{cases}$
P			$\frac{1}{2}\frac{\lambda^2}{1+\eta}1_{t\neq 0}+\frac{\eta}{2}t^2$ (‘$\ell_0+\ell_2$’)
	scad $(a>2)$		**mcp** $(\gamma\geq 1)$
Θ	$\begin{cases} 0, & \text{if } \lvert t\rvert\leq\lambda \\ t-\lambda\,\mathrm{sgn}(t), & \text{if } \lambda<\lvert t\rvert\leq 2\lambda \\ \frac{(a-1)t-a\lambda\,\mathrm{sgn}(t)}{a-2}, & \text{if } 2\lambda<\lvert t\rvert\leq a\lambda \\ t, & \text{if } \lvert t\rvert>a\lambda \end{cases}$		$\begin{cases} 0, & \text{if } \lvert t\rvert<\lambda \\ \frac{t-\lambda\mathrm{sgn}(t)}{1-1/\gamma}, & \text{if } \lambda\leq\lvert t\rvert<\gamma\lambda \\ t, & \text{if } \lvert t\rvert\geq\gamma\lambda \end{cases}$
$\mathcal{L}_\Theta$	$1/(a-1)$		$1/\gamma$
P_Θ	$\frac{\mathrm{d}P}{\mathrm{d}t}=\begin{cases} \lambda\,\mathrm{sgn}(t), & \text{if } \lvert t\rvert\leq\lambda \\ \frac{a\lambda\,\mathrm{sgn}(t)-t}{a-1}, & \text{if } \lambda<\lvert t\rvert\leq a\lambda \\ 0, & \text{if } \lvert t\rvert>a\lambda \end{cases}$		$\begin{cases} -\frac{t^2}{2\gamma}+\lambda\lvert t\rvert, & \text{if } \lvert t\rvert<\gamma\lambda \\ \frac{\gamma\lambda^2}{2}, & \text{if } \lvert t\rvert\geq\gamma\lambda \end{cases}=\frac{1}{\gamma}P_H(t;\gamma\lambda)$
	$\mathbf{l_r}$ $(0<r<1,\ \zeta\geq 0)$		
Θ	$\begin{cases} 0,\ \text{if } \lvert t\rvert\leq\zeta^{1/(2-r)}(2-r)(2-2r)^{(r-1)/(2-r)} \\ \mathrm{sgn}(t)\max\{\zeta^{1/(2-r)}[r(1-r)]^{1/(2-r)}\leq\theta\leq\lvert t\rvert:\theta+\zeta r\theta^{r-1}=\lvert t\rvert\},\ \text{otherwise.(The set is a singleton.)} \end{cases}$		
$\mathcal{L}_\Theta$	1		
P	$\zeta\lvert t\rvert^r$		

The converse is not necessarily true. Namely, Θ-estimators may not guarantee functional local optimality, let alone global optimality. This raises difficulties in statistical analysis. We will give a novel and unified treatment which can yield nearly optimal error rate for various thresholdings.

3. Main results

To address the problems in arbitrary dimensions (with possibly large p and/or n), we aim to establish non-asymptotic oracle inequalities [4]. For any $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$, define

$$\mathcal{J}(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\}, \qquad J(\boldsymbol{\beta}) = |\mathcal{J}(\boldsymbol{\beta})| = \|\boldsymbol{\beta}\|_0. \tag{14}$$

Recall $P_1(t;\lambda) = \lambda|t|$, $P_0(t;\lambda) = \frac{\lambda^2}{2}1_{t\neq 0}$, $P_H(t;\lambda) = (-t^2/2 + \lambda|t|)1_{|t|<\lambda} + (\lambda^2/2)1_{|t|\geq\lambda}$. For convenience, we use $P_1(\boldsymbol{\beta};\lambda)$ to denote $\lambda\|\boldsymbol{\beta}\|_1$ when there is no ambiguity. $P_0(\boldsymbol{\beta};\lambda)$ and $P_H(\boldsymbol{\beta};\lambda)$ are used similarly. We denote by $\lesssim$ an inequality that holds up to a multiplicative constant.

Unless otherwise specified, we study *scaled* Θ-estimators satisfying equation (9), where $\tilde{\boldsymbol{\beta}} = \rho\boldsymbol{\beta}$, $\tilde{\boldsymbol{X}} = \boldsymbol{X}/\rho$, and $\rho \geq \|\boldsymbol{X}\|_2$ (and so $\|\tilde{\boldsymbol{X}}\|_2 \leq 1$). By abuse of notation, we still write $\boldsymbol{\beta}$ for $\tilde{\boldsymbol{\beta}}$, and $\boldsymbol{X}$ for $\tilde{\boldsymbol{X}}$. As mentioned previously, we always assume that Θ is continuous at $\hat{\boldsymbol{\beta}} + \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}$ in Sections 3.1 & 3.2; similarly, Section 3.3 assumes that Θ is continuous at $\boldsymbol{\beta}^{(t)} + \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}^{(t)}$.

The past works on the lasso show that a certain incoherence requirement must be assumed to obtain sharp error rates. In most theorems, we also need to make similar assumptions to prevent the design matrix from being too collinear. We will state a new type of regularity conditions, which are called *comparison regularity conditions*, under which oracle inequalities and sequential statistical error bounds can be obtained for any Θ.

3.1. P_Θ-type oracle inequalities under $\mathcal{R}_0$

In this subsection, we use P_Θ to make a bound of the prediction error of Θ-estimators. Our regularity condition is stated as follows.

Assumption $\mathcal{R}_0(\delta, \vartheta, K, \boldsymbol{\beta}, \lambda)$ Given $\boldsymbol{X}$, Θ, $\boldsymbol{\beta}$, λ, there exist $\delta > 0$, $\vartheta > 0$, $K \geq 0$ such that the following inequality holds for any $\boldsymbol{\beta}' \in \mathbb{R}^p$

$$\begin{aligned} &\vartheta P_H(\boldsymbol{\beta}' - \boldsymbol{\beta};\lambda) + \frac{\mathcal{L}_\Theta}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2 \\ \leq\ &\frac{2-\delta}{2}\|\boldsymbol{X}(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2 + P_\Theta(\boldsymbol{\beta}';\lambda) + KP_\Theta(\boldsymbol{\beta};\lambda). \end{aligned} \tag{15}$$

Roughly, (15) means that $2\|\boldsymbol{X}(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2$ can dominate $\mathcal{L}_\Theta\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2$ with the help from $P_\Theta(\boldsymbol{\beta}';\lambda)$ and $KP_\Theta(\boldsymbol{\beta};\lambda)$ for some $K > 0$.

Theorem 2. *Let $\hat{\boldsymbol{\beta}}$ be any Θ-estimator satisfying $\boldsymbol{\beta} = \Theta(\boldsymbol{\beta} + \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}; \lambda)$ with $\lambda = A\sigma\sqrt{\log(ep)}$ and A a constant. Then for any sufficiently large A, the following oracle inequality holds for $\boldsymbol{\beta} \in \mathbb{R}^p$*

$$\mathbb{E}[\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2] \lesssim \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 + P_\Theta(\boldsymbol{\beta}; \lambda) + \sigma^2, \tag{16}$$

provided $\mathcal{R}_0(\delta, \vartheta, K, \boldsymbol{\beta}, \lambda)$ is satisfied for some constants $\delta > 0, \vartheta > 0$, $K \geq 0$.

Theorem 2 is applicable to any Θ. Let's examine two specific cases. First, consider $\mathcal{L}_\Theta \leq 0$, which indicates that P_Θ is convex. Because $P_H \leq P_\Theta$ and P_H is sub-additive: $P_H(t + s) \leq P_H(t) + P_H(s)$ due to its concavity [20], $\mathcal{R}_0(\delta, \vartheta, K, \boldsymbol{\beta}, \lambda)$ is *always* satisfied (for any $\delta \leq 2$, $0 < \vartheta \leq 1$, $K \geq \vartheta$).

Corollary 1. *Suppose Θ satisfies $\mathcal{L}_\Theta \leq 0$. Then, (16) holds for all corresponding Θ-estimators, without requiring any regularity condition.*

In the case of hard-thresholding or SCAD thresholding, $P_\Theta(\boldsymbol{\beta}; \lambda)$ does not depend on the magnitude of $\boldsymbol{\beta}$, and we can get a finite complexity rate in the oracle inequality. Also, $\mathcal{R}_0$ can be slightly relaxed, by replacing $KP_\Theta(\boldsymbol{\beta}; \lambda)$ with $KP_0(\boldsymbol{\beta}; \lambda)$ in (15). We denote the modified version by $\mathcal{R}_0'(\delta, \vartheta, K, \boldsymbol{\beta}, \lambda)$.

Corollary 2. *Suppose that Θ corresponds to a bounded nonconvex penalty satisfying $P_\Theta(t; \lambda) \leq C\lambda^2$, $\forall t \in \mathbb{R}$, for some constant $C > 0$. Then in the setting of Theorem 2,*

$$\mathbb{E}[\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2] \lesssim \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 + \sigma^2 J(\boldsymbol{\beta}) \log(ep) + \sigma^2, \tag{17}$$

provided $\mathcal{R}_0'(\delta, \vartheta, K, \boldsymbol{\beta}, \lambda)$ is satisfied for some constants $\delta > 0, \vartheta > 0$, $K \geq 0$.

Remark 1. The right-hand side of the oracle inequalities involves a *bias* term $\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2$ and a *complexity* term $P_\Theta(\boldsymbol{\beta}; \lambda)$. Letting $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ in, say, (16), the bias vanishes, and we obtain a prediction error bound of the order $\sigma^2 J^* \log(ep)$ (omitting constant factors), where J^* denotes the number of nonzero components in $\boldsymbol{\beta}^*$. On the other hand, the existence of the bias term ensures the applicability of our results to approximately sparse signals. For example, when $\boldsymbol{\beta}^*$ has many small but nonzero components, we can use a reference $\boldsymbol{\beta}$ with a much smaller support than $\mathcal{J}(\boldsymbol{\beta}^*)$ to get a lower error bound, as a benefit from the bias-variance tradeoff.

Remark 2. When $\mathcal{R}_0$ holds with $\delta > 1$, the proof of Theorem 2 shows that the multiplicative constant for $\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2$ can be as small as 1. The corresponding oracle inequalities are called 'sharp' in some works [7]. This also applies to Theorem 3. Our proof scheme can also deliver high-probability form results, without requiring an upper bound of $\|\boldsymbol{X}\|_2$.

Remark 3. Corollary 2 applies to all "hard-thresholding like" Θ, because when $\Theta(t; \lambda) = t$ for $|t| > c\lambda$, $P_\Theta(t; \lambda) \leq c^2\lambda^2$. It is worth mentioning that the error rate of $\sigma^2 J^* \log(ep)$ cannot be significantly improved in a minimax sense. In fact, under the Gaussian noise contamination and some regularity conditions, there exist constants C, $c > 0$ such that $\inf_{\check{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}^* : J(\boldsymbol{\beta}^*) \leq J} \mathbb{E}[\|\boldsymbol{X}(\check{\boldsymbol{\beta}} -$

$\boldsymbol{\beta}^*)\|_2^2)/(CP_o(J))] \geq c > 0$, where $\check{\boldsymbol{\beta}}$ denotes an arbitrary estimator of $\boldsymbol{\beta}^*$ and $P_o(J) = \sigma^2\{J + J\log(ep/J)\}$. See, e.g., [9] for a proof. The bound in (17) achieves the minimax optimal rate up to a mild logarithm factor for any n and p.

3.2. $\boldsymbol{P_0}$*-type oracle inequalities under* $\boldsymbol{\mathcal{R}_1}$

This part uses P_0 instead of P_Θ to make an oracle bound. We will show that under another type of comparison regularity conditions, *all* thresholdings can attain the essentially optimal error rate given in Corollary 2. We will also show that in the case of soft-thresholding, our condition is more relaxed than many other assumptions in the literature.

Assumption $\mathcal{R}_1(\delta, \vartheta, K, \boldsymbol{\beta}, \lambda)$ Given $\boldsymbol{X}$, Θ, $\boldsymbol{\beta}$, λ, there exist $\delta > 0$, $\vartheta > 0$, $K \geq 0$ such that the following inequality holds for any $\boldsymbol{\beta}' \in \mathbb{R}^p$

$$\begin{aligned} &\vartheta P_H(\boldsymbol{\beta}' - \boldsymbol{\beta}; \lambda) + \frac{\mathcal{L}_\Theta}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2 + P_\Theta(\boldsymbol{\beta}; \lambda) \\ \leq\; &\frac{2-\delta}{2}\|\boldsymbol{X}(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2 + P_\Theta(\boldsymbol{\beta}'; \lambda) + K\lambda^2 J(\boldsymbol{\beta}). \end{aligned} \tag{18}$$

Theorem 3. *Let $\hat{\boldsymbol{\beta}}$ be a Θ-estimator and $\lambda = A\sigma\sqrt{\log(ep)}$ with A a sufficiently large constant. Then $\mathbb{E}[\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2] \lesssim \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 + \lambda^2 J(\boldsymbol{\beta}) + \sigma^2$ holds for any $\boldsymbol{\beta} \in \mathbb{R}^p$ if $\mathcal{R}_1(\delta, \vartheta, K, \boldsymbol{\beta}, \lambda)$ is satisfied for some constants $\delta > 0, \vartheta > 0$, $K \geq 0$.*

Remark 4. Some fusion thresholdings, like those associated with elastic net, Berhu and Hard-Ridge (cf. Table 1), involve an additional ℓ_2 shrinkage. In the situation, the complexity term in the oracle inequality should involve both $J(\boldsymbol{\beta})$ and $\|\boldsymbol{\beta}\|_2^2$. We can modify our regularity conditions to obtain such $\ell_0+\ell_2$ bounds using the same proof scheme. The details are however not reported in this paper. In addition, our results can be extended to Θ-estimators with a stepsize parameter. Given $\lambda > 0$ and $0 < \alpha \leq 1$, suppose λ_α is introduced such that $\alpha P_\Theta(t; \lambda) = P_\Theta(t; \lambda_\alpha)$ for any t. Then, for any $\hat{\boldsymbol{\beta}}$ as a fixed point of $\boldsymbol{\beta} = \Theta(\boldsymbol{\beta} - \alpha\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \alpha\boldsymbol{X}^T\boldsymbol{y}; \lambda_\alpha)$, an analogous result can be obtained (the only change is that $\mathcal{L}_\Theta$ is replaced by $\mathcal{L}_\Theta/\alpha$).

To give some more intuitive regularity conditions, we suppose P_Θ is concave on $[0, \infty)$. Examples include ℓ_r $(0 \leq r \leq 1)$, MCP, SCAD, and so on. The concavity implies $P_\Theta(t + s) \leq P_\Theta(t) + P_\Theta(s)$, and so $P_\Theta(\boldsymbol{\beta}'_{\mathcal{J}}; \lambda) - P_\Theta(\boldsymbol{\beta}_{\mathcal{J}}; \lambda) \leq P_\Theta((\boldsymbol{\beta}' - \boldsymbol{\beta})_{\mathcal{J}}; \lambda)$ and $P_\Theta(\boldsymbol{\beta}'_{\mathcal{J}^c}; \lambda) = P_\Theta((\boldsymbol{\beta}' - \boldsymbol{\beta})_{\mathcal{J}^c}; \lambda)$, where $\mathcal{J}^c$ is the complement of $\mathcal{J}$ and $\boldsymbol{\beta}_{\mathcal{J}}$ is the subvector of $\boldsymbol{\beta}$ indexed by $\mathcal{J}$. Then $\mathcal{R}_1$ is implied by $\mathcal{R}_1'$ below for given $\mathcal{J} = \mathcal{J}(\boldsymbol{\beta})$.

Assumption $\mathcal{R}_1'(\delta, \vartheta, K, \mathcal{J}, \lambda)$ Given $\boldsymbol{X}$, Θ, $\mathcal{J}$, λ, there exist $\delta > 0$, $\vartheta > 0$, $K \geq 0$ such that for any $\boldsymbol{\Delta} \in \mathbb{R}^p$,

$$\begin{aligned}&P_\Theta(\boldsymbol{\Delta}_{\mathcal{J}};\lambda)+\vartheta P_H(\boldsymbol{\Delta}_{\mathcal{J}};\lambda)+\frac{\mathcal{L}_\Theta}{2}\|\boldsymbol{\Delta}\|_2^2\\ &\le\frac{2-\delta}{2}\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2+K\lambda^2J+P_\Theta(\boldsymbol{\Delta}_{\mathcal{J}^c};\lambda)-\vartheta P_H(\boldsymbol{\Delta}_{\mathcal{J}^c};\lambda),\end{aligned}\tag{19}$$

or

$$(1+\vartheta)P_\Theta(\boldsymbol{\Delta}_{\mathcal{J}};\lambda)+\frac{\mathcal{L}_\Theta}{2}\|\boldsymbol{\Delta}\|_2^2\le\frac{2-\delta}{2}\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2+K\lambda^2J+(1-\vartheta)P_\Theta(\boldsymbol{\Delta}_{\mathcal{J}^c};\lambda).\tag{20}$$

When Θ is the soft-thresholding, it is easy to verify that a sufficient condition for (20) is

$$(1+\vartheta)\|\boldsymbol{\Delta}_{\mathcal{J}}\|_1\le K\sqrt{J}\|\boldsymbol{X}\boldsymbol{\Delta}\|_2+\|\boldsymbol{\Delta}_{\mathcal{J}^c}\|_1,\tag{21}$$

for some $\vartheta>0$ and $K\ge 0$. (21) has a simper form than $\mathcal{R}_1$. In the following, we give the definitions of the RE and the compatibility condition [2, 16] to make a comparison to (21).

Assumption $\mathcal{RE}(\kappa_{RE},\vartheta_{RE},\mathcal{J})$. Given $\mathcal{J}\subset[p]$, we say that $\boldsymbol{X}\in\mathbb{R}^{n\times p}$ satisfies $\mathcal{RE}(\kappa_{RE},\vartheta_{RE},\mathcal{J})$, if for positive numbers κ_{RE}, $\vartheta_{RE}>0$,

$$J\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2\ge\kappa_{RE}\|\boldsymbol{\Delta}_{\mathcal{J}}\|_1^2,\tag{22}$$

or more restrictively,

$$\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2\ge\kappa_{RE}\|\boldsymbol{\Delta}_{\mathcal{J}}\|_2^2,\tag{23}$$

for all $\boldsymbol{\Delta}\in\mathbb{R}^p$ satisfying

$$(1+\vartheta_{RE})\|\boldsymbol{\Delta}_{\mathcal{J}}\|_1\ge\|\boldsymbol{\Delta}_{\mathcal{J}^c}\|_1.\tag{24}$$

Assume $\mathcal{RE}(\kappa_{RE},\vartheta_{RE},\mathcal{J})$ holds. When $(1+\vartheta_{RE})\|\boldsymbol{\Delta}_{\mathcal{J}}\|_1\le\|\boldsymbol{\Delta}_{\mathcal{J}^c}\|_1$, (21) holds trivially with $\vartheta=\vartheta_{RE}$; otherwise, (22) indicates $(1+\vartheta)\|\boldsymbol{\Delta}_{\mathcal{J}}\|_1\le K\sqrt{J}\|\boldsymbol{X}\boldsymbol{\Delta}\|_2$ with $K=(1+\vartheta_{RE})/\sqrt{\kappa_{RE}}$. So intuitively, we have the following relationship:

$$(23)+(24)\Rightarrow(22)+(24)\Rightarrow(21)\Rightarrow(20)\Rightarrow(19)\Rightarrow(18).$$

In particular, $\mathcal{R}_1$ is less demanding than RE.

Next, let's compare the regularity conditions required by Θ_S and Θ_H to achieve the nearly optimal error rate. Recall $\mathcal{R}_1(\delta,\vartheta,K,\boldsymbol{\beta},\lambda)$ and $\mathcal{R}_0'(\delta,\vartheta,K,\boldsymbol{\beta},\lambda)$ in Theorem 3 and Corollary 2, respectively

$$\vartheta P_H(\boldsymbol{\beta}'-\boldsymbol{\beta};\lambda)+\lambda\|\boldsymbol{\beta}\|_1\le\frac{2-\delta}{2}\|\boldsymbol{X}(\boldsymbol{\beta}'-\boldsymbol{\beta})\|_2^2+\lambda\|\boldsymbol{\beta}'\|_1+K\lambda^2J,$$

$$\vartheta P_H(\boldsymbol{\beta}'-\boldsymbol{\beta};\lambda)+\frac{1}{2}\|\boldsymbol{\beta}'-\boldsymbol{\beta}\|_2^2\le\frac{2-\delta}{2}\|\boldsymbol{X}(\boldsymbol{\beta}'-\boldsymbol{\beta})\|_2^2+P_H(\boldsymbol{\beta}';\lambda)+K\lambda^2J.$$

$\mathcal{R}_0'(\delta, \vartheta, K, \boldsymbol{\beta}, \lambda)$ implies $\mathcal{R}_1(\delta, \vartheta, K+1, \boldsymbol{\beta}, \lambda)$. Indeed, for $\boldsymbol{\Delta} = \boldsymbol{\beta}' - \boldsymbol{\beta}$,

$$\begin{aligned}
\lambda\|\boldsymbol{\beta}\|_1 - \lambda\|\boldsymbol{\beta}'\|_1 &\le \lambda\|\boldsymbol{\Delta}_{\mathcal{J}}\|_1 - \lambda\|\boldsymbol{\beta}'_{\mathcal{J}^c}\|_1 \\
&\le \frac{1}{2}\lambda^2 J + \frac{1}{2}\|\boldsymbol{\Delta}_{\mathcal{J}}\|_2^2 - P_H(\boldsymbol{\beta}'_{\mathcal{J}^c}; \lambda) \\
&\le \frac{1}{2}\lambda^2 J + \frac{1}{2}\|\boldsymbol{\Delta}_{\mathcal{J}}\|_2^2 - P_H(\boldsymbol{\beta}'; \lambda) + P_H(\boldsymbol{\beta}'_{\mathcal{J}}; \lambda) \\
&\le \frac{1}{2}\lambda^2 J + \frac{1}{2}\|\boldsymbol{\Delta}_{\mathcal{J}}\|_2^2 - P_H(\boldsymbol{\beta}'; \lambda) + P_0(\boldsymbol{\beta}'_{\mathcal{J}}; \lambda) \\
&\le \lambda^2 J + \frac{1}{2}\|\boldsymbol{\Delta}_{\mathcal{J}}\|_2^2 - P_H(\boldsymbol{\beta}'; \lambda).
\end{aligned}$$

On the other hand, Corollary 2 studies when *all* Θ_H-estimators have the optimal performance guarantee, while practically, one may initialize (10) with a carefully chosen starting point.

Theorem 4. *Given any Θ, there exists a Θ-estimator (which minimizes (12)) such that (16) holds without requiring any regularity condition. In particular, if Θ corresponds to a bounded nonconvex penalty as described in Corollary 2, then there exists a Θ-estimator such that (17) holds free of regularity conditions.*

Theorem 4 does *not* place any requirement on $\boldsymbol{X}$. So it seems that applying Θ_H may have some further advantages in practice. (How to efficiently pick a Θ_H-estimator to completely remove all regularity conditions is however beyond the the scope of the current paper. For a possible idea of relaxing the conditions, see Remark 6.)

Finally, we make a discussion of the scaling parameter ρ. Our results so far are obtained after performing $\boldsymbol{X} \leftarrow \boldsymbol{X}/\rho$ with $\rho \ge \|\boldsymbol{X}\|_2$. The prediction error is invariant to the transformation. But it affects the regularity conditions.

Seen from (8), $1/\rho^2$ is related to the stepsize α appearing in (1), also known as the *learning rate* in the machine learning literature. From the computational results in Section 2.2, ρ must be large enough to guarantee TISP is convergent. The larger the value of ρ is, the smaller the stepsize is (and so the slower the convergence is). Based on the machine learning literature, slow learning rates are always recommended when training a nonconvex learner (e.g., artificial neural networks). Perhaps interestingly, in addition to computational efficiency reasons, all our statistical analyses caution against using an extremely large scaling when $\mathcal{L}_\Theta > 0$. For example, $\mathcal{R}_0'(\delta, \vartheta, K, \boldsymbol{\beta}, \lambda)$ for an unscaled $\boldsymbol{X}$ reads $\vartheta P_H(\rho(\boldsymbol{\beta}' - \boldsymbol{\beta}); \lambda) + \rho^2\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2/2 \le (2-\delta)\|\boldsymbol{X}(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2/2 + P_H(\rho\boldsymbol{\beta}'; \lambda) + K\lambda^2 J$, which becomes difficult to hold when ρ is very large. This makes the statistical error bound break down easily. Therefore, a good idea is to have ρ just appropriately large (mildly greater than $\|\boldsymbol{X}\|_2$). The sequential analysis of the iterates in the next part also supports the point.

3.3. Sequential algorithmic analysis

We perform statistical error analysis of the sequence of iterates defined by TISP: $\boldsymbol{\beta}^{(t+1)} = \Theta(\boldsymbol{\beta}^{(t)} + \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}^{(t)}; \lambda)$, where $\|\boldsymbol{X}\|_2 \le 1$ and $\boldsymbol{\beta}^{(0)}$ is the starting point. The study is motivated from the fact that in large-scale applications, Θ-estimators are seldom computed exactly. Indeed, why bother to run TISP till computational convergence? How does the statistical accuracy improve (or deteriorate) at t increases? Lately, there are some key advances on the topic. For example, [1] showed that for convex problems (not necessarily strongly convex), proximal gradient algorithms can be geometrically fast to approach a globally optimal solution $\hat{\boldsymbol{\beta}}$ within the desired statistical precision, under a set of conditions. We however care about the statistical error between $\boldsymbol{\beta}^{(t)}$ and the genuine $\boldsymbol{\beta}^*$ in this work.

We will introduce two comparison regularity conditions (analogous to $\mathcal{R}_0$ and $\mathcal{R}_1$) to present both P_Θ-type and P_0-type error bounds. Hereinafter, denote $(\boldsymbol{\beta}^T\boldsymbol{A}\boldsymbol{\beta})^{1/2}$ by $\|\boldsymbol{\beta}\|_{\boldsymbol{A}}$, where $\boldsymbol{A}$ is a positive semi-definite matrix.

Assumption $\mathcal{S}_0(\delta, \vartheta, K, \boldsymbol{\beta}, \boldsymbol{\beta}', \lambda)$ Given $\boldsymbol{X}$, Θ, $\boldsymbol{\beta}$, $\boldsymbol{\beta}'$, λ, there exist $\delta > 0$, $\vartheta > 0$, $K \geq 0$ such that the following inequality holds

$$\begin{aligned} &\vartheta P_H(\boldsymbol{\beta}' - \boldsymbol{\beta}; \lambda) + \frac{\mathcal{L}_\Theta + \delta}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2 \\ &\leq \|\boldsymbol{X}(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2 + P_\Theta(\boldsymbol{\beta}'; \lambda) + KP_\Theta(\boldsymbol{\beta}; \lambda). \end{aligned} \tag{25}$$

Assumption $\mathcal{S}_1(\delta, \vartheta, K, \boldsymbol{\beta}, \boldsymbol{\beta}', \lambda)$ Given $\boldsymbol{X}$, Θ, $\boldsymbol{\beta}$, $\boldsymbol{\beta}'$, λ, there exist $\delta > 0$, $\vartheta > 0$, $K \geq 0$ such that the following inequality holds

$$\begin{aligned} &\vartheta P_H(\boldsymbol{\beta}' - \boldsymbol{\beta}; \lambda) + \frac{\mathcal{L}_\Theta + \delta}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2 + P_\Theta(\boldsymbol{\beta}; \lambda) \\ &\leq \|\boldsymbol{X}(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2 + P_\Theta(\boldsymbol{\beta}'; \lambda) + K\lambda^2 J(\boldsymbol{\beta}). \end{aligned} \tag{26}$$

(25) and (26) require a bit more than (15) and (18), respectively, due to $\|\boldsymbol{X}\|_2 \leq 1$. The theorem and the corollary below perform sequential analysis of the iterates and reveal the explicit roles of δ, ϑ, K (which can often be treated as constants).

Theorem 5. *Suppose $\mathcal{S}_0(\delta, \vartheta, K, \boldsymbol{\beta}^*, \boldsymbol{\beta}^{(t)}, \lambda)$ is satisfied for some $\delta > 0, \vartheta > 0$, $K \geq 0$, then for $\lambda = A\sigma\sqrt{\log(ep)}/\sqrt{(\delta \wedge \vartheta)\vartheta}$ with A sufficiently large, the following error bound holds with probability at least $1 - Cp^{-cA^2}$:*

$$\frac{1+\delta}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_{(\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X})}^2 \leq \frac{1}{2}\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_{(\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X})}^2 + (K+1)P_\Theta(\boldsymbol{\beta}^*; \lambda), \tag{27}$$

where C, c are universal positive constants.

Similarly, under the same choice of regularity parameter, if $\mathcal{S}_1(\delta, \vartheta, K, \boldsymbol{\beta}^*, \boldsymbol{\beta}^{(t)}, \lambda)$ is satisfied for some $\delta > 0, \vartheta > 0$, $K \geq 0$, (28) is true with probability at least $1 - Cp^{-cA^2}$:

$$\frac{1+\delta}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_{(\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X})}^2 \leq \frac{1}{2}\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_{(\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X})}^2 + K\lambda^2 J^*. \tag{28}$$

Corollary 3. *In the setting of Theorem 5, for any initial point $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^p$, we have*

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|^2_{(\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X})} \le \kappa^t \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|^2_{(\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X})} + \frac{\kappa}{1-\kappa} K' P_\Theta(\boldsymbol{\beta}^*; \lambda), \tag{29}$$

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|^2_{(\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X})} \le \kappa^t \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|^2_{(\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X})} + \frac{\kappa}{1-\kappa} K'' \lambda^2 J^*, \tag{30}$$

under $\mathcal{S}_0(\delta, \vartheta, K, \boldsymbol{\beta}^*, \boldsymbol{\beta}^{(s)}, \lambda)$ and $\mathcal{S}_1(\delta, \vartheta, K, \boldsymbol{\beta}^*, \boldsymbol{\beta}^{(s)}, \lambda), 0 \le s \le t-1$, respectively, with probability at least $1 - Cp^{-cA^2}$. Here, $\kappa = 1/(1+\delta)$, $K' = 2(K+1)$, $K'' = 2K$.

Remark 5. We can get some sufficient conditions for $\mathcal{S}_0$ and $\mathcal{S}_1$, similar to the discussions made in Section 3.2. When $\|\boldsymbol{X}\|_2$ is strictly less than 1, (25) can be relaxed to $\vartheta P_H(\boldsymbol{\beta}' - \boldsymbol{\beta}; \lambda) + (\mathcal{L}_\Theta + \delta)\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2/2 \le (2+\delta)\|\boldsymbol{X}(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2/2 + P_\Theta(\boldsymbol{\beta}'; \lambda) + K P_\Theta(\boldsymbol{\beta}; \lambda)$ for some $\delta > 0$. The proof in Section 4.4 also gives expectation-form results, with an additional additive term $C\sigma^2/(\delta \wedge \vartheta)$ in the upper bounds. Similar to Remark 4, we can also study Θ-iterates with stepsize α, in which case the weighting matrix in (27)-(30) changes from $\boldsymbol{I} - \boldsymbol{X}^T\boldsymbol{X}$ to $\boldsymbol{I}/\alpha - \boldsymbol{X}^T\boldsymbol{X}$, and the factor $(\mathcal{L}_\Theta + \delta)/2$ in (25) and (26) is replaced by $(\mathcal{L}_\Theta + \delta)/(2\alpha)$.

Remark 6. Theorem 5 still applies when δ, ϑ, K and λ are dependent on t. For example, if we use a varying threshold sequence, i.e., $\boldsymbol{\beta}^{(t+1)} = \Theta(\boldsymbol{\beta}^{(t)} + \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}^{(t)}; \lambda^{(t)})$, then (30) becomes

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|^2_{(\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X})} \le \kappa^t \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|^2_{(\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X})} + K'' J^* \sum_{s=0}^{t-1} \kappa^{t-s} \lambda_s^2.$$

This allows for much larger values of λ_s to be used in earlier iterations to attain the same accuracy. It relaxes the regularity condition required by applying a fixed threshold level.

At the end, we re-state some results under $\rho > \|\boldsymbol{X}\|_2$, to get more intuition and implications. For a general $\boldsymbol{X}$ (unscaled), (30) reads

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|^2_{(\rho^2\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X})} \le \kappa^t \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|^2_{(\rho^2\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X})} + \frac{\kappa}{1-\kappa} K'' \sigma^2 \lambda^2 J^*.$$

Set ρ to be a number slightly larger than $\|\boldsymbol{X}\|_2$, i.e., $\rho = (1+\epsilon)\|\boldsymbol{X}\|_2$, $\epsilon > 0$. Then, we know that the prediction error $\|\boldsymbol{X}\boldsymbol{\beta}^{(t)} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2$ decays geometrically fast to $O(\sigma^2 J^* \log(ep))$ with high probability, when ϵ, δ, ϑ, K are viewed as constants; a similar conclusion is true for the estimation error. This is simply due to

$$\frac{\rho^2 - \|\boldsymbol{X}\|_2^2}{\|\boldsymbol{X}\|_2^2} \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|^2_{\boldsymbol{X}^T\boldsymbol{X}} \le (\rho^2 - \|\boldsymbol{X}\|_2^2)\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2^2 \le \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|^2_{(\rho^2\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X})}.$$

Accordingly, there is no need to run TISP till convergence—one can terminate the algorithm earlier, at, say, $t_{\max} = \log\{\rho^2\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|^2/(K\sigma^2\lambda^2 J^*)\} / \log(1/\kappa)$, without sacrificing much statistical accuracy. The formula also reflects that the quality of the initial point affects the required iteration number.

There are some related results in the literature. (i) As mentioned previously, in a broad convex setting [1] proved the geometric decay of the optimization error $\|\boldsymbol{\beta}^{(t)} - \hat{\boldsymbol{\beta}}\|$ to the desired statistical precision, where $\hat{\boldsymbol{\beta}}$ is the convergent point. [8] extended the conclusion to a family of nononvex optimization problems, and they showed that when some regularity conditions hold, *every* local minimum point is close to the authentic $\boldsymbol{\beta}^*$. In comparison, our results are derived toward the statistical error between $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\beta}^*$ directly, without requiring all local minimum points to be statistically accurate. (ii) [21] showed a similar fast-converging statistical error bound for an elegant multi-stage capped-ℓ_1 regularization procedure. However, the procedure carries out an expensive ℓ_1 optimization at each step. Instead, (10) involves a simple and cheap thresholding, and our analysis covers any Θ.

4. Proofs

Throughout the proofs, we use C, c, L to denote universal non-negative constants. They are not necessarily the same at each occurrence. Given any matrix $\boldsymbol{A}$, we use $\mathcal{R}(\boldsymbol{A})$ to denote its column space. Denote by $\mathbf{P}_{\boldsymbol{A}}$ the orthogonal projection matrix onto $\mathcal{R}(\boldsymbol{A})$, i.e., $\mathbf{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^+\boldsymbol{A}^T$, where $^+$ stands for the Moore-Penrose pseudoinverse. Let $[p] := \{1, \cdots, p\}$. Given $\mathcal{J} \subset [p]$, we use $\boldsymbol{X}_{\mathcal{J}}$ to denote a column submatrix of $\boldsymbol{X}$ indexed by $\mathcal{J}$.

Definition 2. *ξ is called a sub-Gaussian random variable if there exist constants $C, c > 0$ such that $\mathbb{P}\{|\xi| \ge t\} \le Ce^{-ct^2}, \forall t > 0$. The scale ($\psi_2$-norm) for ξ is defined as $\sigma(\xi) = \inf\{\sigma > 0 : \mathbb{E}\exp(\xi^2/\sigma^2) \le 2\}$. $\boldsymbol{\xi} \in \mathbb{R}^p$ is called a sub-Gaussian random vector with scale bounded by σ if all one-dimensional marginals $\langle \boldsymbol{\xi}, \boldsymbol{\alpha}\rangle$ are sub-Gaussian satisfying $\|\langle \boldsymbol{\xi}, \boldsymbol{\alpha}\rangle\|_{\psi_2} \le \sigma\|\boldsymbol{\alpha}\|_2$, $\forall \boldsymbol{\alpha} \in R^p$.*

Examples include Gaussian random variables and bounded random variables such as Bernoulli. Note that the assumption that $\mathrm{vec}\,(\boldsymbol{\epsilon})$ is sub-Gaussian does not imply that the components of $\boldsymbol{\epsilon}$ must be i.i.d.

We begin with two basic facts. Because they are special cases of Lemma 1 and Lemma 2 in [13], respectively, we state them without proofs.

Lemma 1. *Given an arbitrary thresholding rule Θ, let P be any function satisfying $P(\theta;\lambda) - P(0;\lambda) = P_\Theta(\theta;\lambda) + q(\theta;\lambda)$ where $P_\Theta(\theta;\lambda) \triangleq \int_0^{|\theta|}(\sup\{s : \Theta(s;\lambda) \le u\} - u)\,\mathrm{d}u$, $q(\theta;\lambda)$ is nonnegative and $q(\Theta(t;\lambda)) = 0$ for all t. Then, $\hat{\beta} = \Theta(y;\lambda)$ is always a globally optimal solution to $\min_\beta \frac{1}{2}\|y - \beta\|_2^2 + P(|\boldsymbol{\beta}|;\lambda)$. It is the unique optimal solution provided $\Theta(\cdot;\lambda)$ is continuous at $|y|$.*

Lemma 2. *Let $Q_0(\beta) = \|y - \beta\|_2^2/2 + P_\Theta(|\beta|;\lambda)$. Denote by $\hat{\beta}$ the unique minimizer of $Q_0(\beta)$. Then for any δ, $Q_0(\hat{\beta} + \delta) - Q_0(\hat{\beta}) \ge (1 - \mathcal{L}_\Theta)\|\delta\|_2^2/2$.*

4.1. Proof of Theorem 1

Let $s(u;\lambda) := \Theta^{-1}(u;\lambda) - u$ for $u \geq 0$. Assume $\hat{\boldsymbol{\beta}}$ is a local minimum point (the proof for a coordinate-wise minimum point follows the same lines). We write f_Θ as f for simplicity. Let $\delta f(\boldsymbol{\beta};\boldsymbol{h})$ denote the Gateaux differential of f at $\boldsymbol{\beta}$ with increment $\boldsymbol{h}$: $\delta f(\boldsymbol{\beta};\boldsymbol{h}) = \lim_{\epsilon\to 0+} \frac{f(\boldsymbol{\beta}+\epsilon\boldsymbol{h})-f(\boldsymbol{\beta})}{\epsilon}$. By the definition of P_Θ, $\delta f(\boldsymbol{\beta},\boldsymbol{h})$ exists for any $\boldsymbol{h} \in \mathbb{R}^p$. Let $l(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2$. We consider the following directional vectors: $\boldsymbol{d}_j = [d_1, \cdots, d_p]^T$ with $d_j = \pm 1$ and $d_{j'} = 0, \forall j' \neq j$. Then for any j,

$$\delta l(\boldsymbol{\beta};\boldsymbol{d}_j) = d_j \boldsymbol{x}_j^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}), \tag{31}$$

$$\delta P_\Theta(\boldsymbol{\beta};\boldsymbol{d}_j) = \begin{cases} s(|\beta_j|)\mathrm{sgn}(\beta_j)d_j, & \text{if } \beta_j \neq 0, \\ s(|\beta_j|), & \text{if } \beta_j = 0. \end{cases} \tag{32}$$

Due to the local optimality of $\hat{\boldsymbol{\beta}}$, $\delta f(\hat{\boldsymbol{\beta}};\boldsymbol{d}_j) \geq 0$, $\forall j$. When $\hat{\beta}_1 \neq 0$, we obtain $\boldsymbol{x}_1^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}) + s(|\hat{\beta}_1|;\lambda)\mathrm{sgn}(\hat{\beta}_1) = 0$. When $\hat{\beta}_1 = 0$, $\boldsymbol{x}_1^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}) + s(|\hat{\beta}_1|;\lambda) \geq 0$ and $-\boldsymbol{x}_1^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}) + s(|\hat{\beta}_1|;\lambda) \geq 0$, i.e., $|\boldsymbol{x}_1^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y})| \leq s(|\hat{\beta}_1|;\lambda) = \Theta^{-1}(0;\lambda)$. To summarize, when f achieves a local minimum or a coordinate-wise minimum (or more generally, a *local* coordinate-wise minimum) at $\hat{\boldsymbol{\beta}}$, we have

$$\hat{\beta}_j \neq 0 \Rightarrow \Theta^{-1}(|\hat{\beta}_j|;\lambda)\mathrm{sgn}(\hat{\beta}_j) = \hat{\beta}_j - \boldsymbol{x}_j^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}) \tag{33}$$

$$\hat{\beta}_j = 0 \Rightarrow \Theta(\boldsymbol{x}_j^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y});\lambda) = 0 \tag{34}$$

When Θ is continuous at $\hat{\beta}_j - \boldsymbol{x}_j^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y})$, (33) implies that $\hat{\beta}_j = \Theta(\hat{\beta}_j - \boldsymbol{x}_j^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y});\lambda)$. Hence $\hat{\boldsymbol{\beta}}$ must be a Θ-estimator satisfying $\boldsymbol{\beta} = \Theta(\boldsymbol{\beta} + \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta};\lambda)$.

4.2. Proofs of Theorem 2 and Theorem 3

Given Θ, let $\hat{\boldsymbol{\beta}}$ be any Θ-estimator, $\boldsymbol{\beta}$ be any p-dimensional vector (non-random) and $\boldsymbol{\Delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. The first result constructs a useful criterion for $\hat{\boldsymbol{\beta}}$ on basis of Lemma 1 and Lemma 2.

Lemma 3. *Any Θ-estimator $\hat{\boldsymbol{\beta}}$ satisfies the following inequality for* any $\boldsymbol{\beta} \in \mathbb{R}^p$

$$\begin{aligned} &\frac{1}{2}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + \frac{1}{2}\boldsymbol{\Delta}^T(\boldsymbol{X}^T\boldsymbol{X} - \mathcal{L}_\Theta\boldsymbol{I})\boldsymbol{\Delta} \\ \leq\ &\frac{1}{2}\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2^2 + P_\Theta(\boldsymbol{\beta};\lambda) - P_\Theta(\hat{\boldsymbol{\beta}};\lambda) + \langle\boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\Delta}\rangle, \end{aligned} \tag{35}$$

where $\boldsymbol{\Delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$.

To handle $\langle\boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\Delta}\rangle$, we introduce another lemma.

Lemma 4. *Suppose* $\|\boldsymbol{X}\|_2 \leq 1$ *and let* $\lambda^o = \sigma\sqrt{\log(ep)}$. *Then there exist universal constants* $A_1, C, c > 0$ *such that for any constants* $a \geq 2b > 0$, *the following event*

$$\sup_{\boldsymbol{\beta}\in\mathbb{R}^p}\{2\langle\boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\beta}\rangle - \frac{1}{a}\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2 - \frac{1}{b}[P_H(\boldsymbol{\beta}; \sqrt{ab}A_1\lambda^o)]\} \geq a\sigma^2 t \tag{36}$$

occurs with probability at most $C\exp(-ct)p^{-cA_1^2}$, *where* $t \geq 0$.

The lemma plays an important role in bounding the last stochastic term in (35). Its proof is based on the following results.

Lemma 5. *Suppose* $\|\boldsymbol{X}\|_2 \leq 1$. *There exists a globally optimal solution* $\boldsymbol{\beta}^o$ *to* $\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + P_H(\boldsymbol{\beta}; \lambda)$ *such that for any* $j : 1 \leq j \leq p$, *either* $\beta_j^o = 0$ *or* $|\beta_j^o| \geq \lambda$.

Lemma 6. *Given* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ *and* $J : 1 \leq J \leq p$, *define* $\Gamma'_J = \{\boldsymbol{\alpha} \in \mathbb{R}^p : \|\boldsymbol{\alpha}\|_2 \leq 1, \boldsymbol{\alpha} \in \mathcal{R}(\boldsymbol{X}_{\mathcal{J}})$ *for some* $\mathcal{J} : |\mathcal{J}| = J\}$. *Let* $P'_o(J) = \sigma^2\{J + \log\binom{p}{J}\}$. *Then for any* $t \geq 0$,

$$\mathbb{P}\left(\sup_{\boldsymbol{\alpha}\in\Gamma'_J}\langle\boldsymbol{\epsilon}, \boldsymbol{\alpha}\rangle \geq t\sigma + \sqrt{LP'_o(J)}\right) \leq C\exp(-ct^2), \tag{37}$$

where $L, C, c > 0$ *are universal constants.*

Let $R = \sup_{1\leq J\leq p}\sup_{\boldsymbol{\Delta}\in\Gamma_J}\{\langle\boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\Delta}\rangle - \frac{1}{2b}P_H(\boldsymbol{\Delta}; \sqrt{ab}A_1\lambda^o) - \frac{1}{2a}\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2\}$, with λ^o, A_1 given in Lemma 4. (The starting value of J is 1 because when $J(\boldsymbol{\Delta}) = 0$, $\langle\boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\Delta}\rangle = 0$.) Substituting it into (35) gives

$$\begin{aligned}
&\frac{1}{2}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + \frac{1}{2}\boldsymbol{\Delta}^T(2\boldsymbol{X}^T\boldsymbol{X} - \mathcal{L}_\Theta\boldsymbol{I})\boldsymbol{\Delta}\\
\leq&\frac{1}{2}\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2^2 + P_\Theta(\boldsymbol{\beta}; \lambda) - P_\Theta(\hat{\boldsymbol{\beta}}; \lambda) + \frac{1}{2b}P_H(\boldsymbol{\Delta}; \sqrt{ab}A_1\lambda^o)\\
&+ \frac{1}{2a}\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2 + \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2 + R\\
\leq&\frac{1}{2}\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2^2 + P_\Theta(\boldsymbol{\beta}; \lambda) - P_\Theta(\hat{\boldsymbol{\beta}}; \lambda) + \frac{1}{2b}P_H(\boldsymbol{\Delta}; \sqrt{ab}A_1\lambda^o)\\
&+ \frac{1}{2}(1 + \frac{1}{a})\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2 + R.
\end{aligned}$$

Because $\mathbb{P}(R \geq a\sigma^2 t) \leq C\exp(-ct)$, we know $\mathbb{E}[R] \lesssim a\sigma^2$.

Let $\lambda = A\lambda^o$ with $A = A_1\sqrt{ab}$ and set $b \geq 1/(2\vartheta)$. The regularity condition $\mathcal{R}_0(\delta, \vartheta, K, \boldsymbol{\beta}, \lambda)$ implies that

$$\frac{1}{2b}P_H(\boldsymbol{\Delta}; \lambda) + \frac{\mathcal{L}_\Theta}{2}\|\boldsymbol{\Delta}\|_2^2 \leq \frac{2-\delta}{2}\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2 + P_\Theta(\hat{\boldsymbol{\beta}}; \lambda) + KP_\Theta(\boldsymbol{\beta}; \lambda). \tag{38}$$

Choose a to satisfy $a > 1/\delta$, $a \geq 2b$. Combining the last two inequalities gives

$$\mathbb{E}[\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2]$$

$$\begin{aligned}
&\leq \|\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\|_2^2 + 2(K+1)P_\Theta(\boldsymbol{\beta};\lambda) + \mathbb{E}[(1+\frac{1}{a}-\delta)\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2] + 2\,\mathbb{E}[R]\\
&\lesssim \|\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\|_2^2 + P_\Theta(\boldsymbol{\beta};\lambda) + \sigma^2, \qquad (39)
\end{aligned}$$

with the last inequality due to $\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2 \leq (1+1/c)\|\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\|_2^2 + (1+c)\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2^2$ for any $c>0$.

The proof of Theorem 3 follows the lines of the proof of Theorem 2, with (38) replaced by

$$\frac{1}{2b}P_H(\boldsymbol{\Delta};\lambda) + \frac{\mathcal{L}_\Theta}{2}\|\boldsymbol{\Delta}\|_2^2 + P_\Theta(\boldsymbol{\beta};\lambda) \leq \frac{2-\delta}{2}\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2 + P_\Theta(\hat{\boldsymbol{\beta}};\lambda) + K\lambda^2 J(\boldsymbol{\beta}),$$

and (39) replaced by

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2^2]\\
&\leq \|\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\|_2^2 + 2K\lambda^2 J(\boldsymbol{\beta}) + \mathbb{E}[(1+\frac{1}{a}-\delta)\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2] + 2\,\mathbb{E}[R]\\
&\lesssim \|\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\|_2^2 + \lambda^2 J(\boldsymbol{\beta}) + \sigma^2.
\end{aligned}$$

The details are omitted.

4.3. Proof of Theorem 4

From the proof of Lemma 5, there exists a Θ-estimator $\hat{\boldsymbol{\beta}}$ which minimizes $f(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + P_\Theta(\boldsymbol{\beta};\lambda)$. This means that the term $\frac{1}{2}\boldsymbol{\Delta}^T(\boldsymbol{X}^T\boldsymbol{X} - \mathcal{L}_\Theta\boldsymbol{I})\boldsymbol{\Delta}$ can be dropped from (35). Following the lines of Section 4.2, (17) holds under a modified version of $\mathcal{R}_0(\delta,\vartheta,K,\boldsymbol{\beta},\lambda)$, which replaces (15) with

$$\vartheta P_H(\boldsymbol{\beta}'-\boldsymbol{\beta};\lambda) \leq \frac{1-\delta}{2}\|\boldsymbol{X}(\boldsymbol{\beta}'-\boldsymbol{\beta})\|_2^2 + P_\Theta(\boldsymbol{\beta}';\lambda) + KP_\Theta(\boldsymbol{\beta};\lambda). \qquad (40)$$

Using the sub-additivity of P_H, we know that any design matrix satisfies (40) for any $0<\vartheta\leq 1$, $\delta\leq 1$, $K\geq\vartheta$.

4.4. Proof of Theorem 5 and Corollary 3

Let $f(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + P_\Theta(\boldsymbol{\beta};\lambda)$ where $l(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y}\|_2^2$.

Lemma 7. *Let $\boldsymbol{\beta}^{(t+1)} = \Theta(\boldsymbol{\beta}^{(t)} + \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}^{(t)};\lambda)$. Then the following 'triangle inequality' holds for any $\boldsymbol{\beta}\in\mathbb{R}^p$*

$$\begin{aligned}
&\frac{1-\mathcal{L}_\Theta}{2}\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}\|_2^2 + \frac{1}{2}\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\|_{\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X}}^2\\
&\leq \frac{1}{2}\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}\|_{\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X}}^2 + f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^{(t+1)}).
\end{aligned}$$

Letting $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ in the lemma, we have

$$\begin{aligned}&\frac{1}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|^2_{\boldsymbol{X}^T\boldsymbol{X}+(1-\mathcal{L}_\Theta)\boldsymbol{I}} + \frac{1}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|^2_{\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X}} + P_\Theta(\boldsymbol{\beta}^{(t+1)};\lambda)\\ \le&\frac{1}{2}\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|^2_{\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X}} + \langle \boldsymbol{\epsilon}, \boldsymbol{X}(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*)\rangle + P_\Theta(\boldsymbol{\beta}^*;\lambda).\end{aligned}$$

Moreover, under $\mathcal{S}_0(\delta, \vartheta, K, \boldsymbol{\beta}^*, \boldsymbol{\beta}', \lambda)$ with $\boldsymbol{\beta}' = \boldsymbol{\beta}^{(t+1)}$,

$$\begin{aligned}&\vartheta P_H(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*;\lambda) + \frac{1+\delta}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_2^2 - KP_\Theta(\boldsymbol{\beta}^*;\lambda)\\ \le&\frac{1}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|^2_{\boldsymbol{X}^T\boldsymbol{X}+(1-\mathcal{L}_\Theta)\boldsymbol{I}} + P_\Theta(\boldsymbol{\beta}^{(t+1)};\lambda) + \frac{1}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|^2_{\boldsymbol{X}^T\boldsymbol{X}}.\end{aligned}$$

Combining the last two inequalities gives

$$\begin{aligned}&\frac{1+\delta}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|^2_{\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X}} + \frac{1}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|^2_{\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X}}\\ &+ \frac{\delta}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|^2_{\boldsymbol{X}^T\boldsymbol{X}} + \vartheta P_H(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*;\lambda)\\ \le&\frac{1}{2}\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|^2_{\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X}} + (K+1)P_\Theta(\boldsymbol{\beta}^*;\lambda) + \langle \boldsymbol{\epsilon}, \boldsymbol{X}(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*)\rangle.\end{aligned}$$

Let $\Gamma_J = \{\boldsymbol{\beta} \in \mathbb{R}^p : J(\boldsymbol{\beta}) = J\}$, $\lambda^o = \sigma\sqrt{\log(ep)}$. We define an event $\mathcal{E}$ with its complement given by

$$\mathcal{E}^c \triangleq \{\sup_{\boldsymbol{\beta}}\{2\langle \boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\beta}\rangle - \frac{1}{a}\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2 - \frac{1}{b}[P_H(\boldsymbol{\beta}; \sqrt{ab}A_1\lambda^o)]\} \ge 0\}.$$

By Lemma 4, there exists a universal constant L such that for any $A_1^2 \ge L$, $a \ge 2b > 0$, $P(\mathcal{E}^c) \le Cp^{-cA_1^2}$. Clearly, $\mathcal{E}$ implies

$$\langle \boldsymbol{\epsilon}, \boldsymbol{X}(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*)\rangle \le \frac{1}{2a}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|^2_{\boldsymbol{X}^T\boldsymbol{X}} + \frac{1}{2b}P_H(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*; \sqrt{ab}A_1\lambda^o). \tag{41}$$

Take $b = 1/(2\vartheta)$, $a = 1/(\delta \wedge \vartheta)$, $A_1 \ge \sqrt{L}$, and $\lambda = A_1\sqrt{ab}\lambda^o$. Then, on $\mathcal{E}$ we get the desired statistical accuracy bound

$$\frac{1+\delta}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|^2_{\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X}} \le \frac{1}{2}\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|^2_{\boldsymbol{I}-\boldsymbol{X}^T\boldsymbol{X}} + (K+1)P_\Theta(\boldsymbol{\beta}^*;\lambda).$$

The bound under $\mathcal{S}_1$ can be similarly proved. Noticing that (41) holds for any t, Corollary 3 is immediately true.

4.5. Proofs of Lemmas

4.5.1. Proof of Lemma 3

Let $f(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + P_\Theta(\boldsymbol{\beta};\lambda)$ with $l(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2$. Define

$$g(\boldsymbol{\beta}, \boldsymbol{\gamma}) = l(\boldsymbol{\beta}) + \langle \nabla l(\boldsymbol{\beta}), \boldsymbol{\gamma} - \boldsymbol{\beta}\rangle + \frac{1}{2}\|\boldsymbol{\gamma} - \boldsymbol{\beta}\|_2^2 + P_\Theta(\boldsymbol{\gamma};\lambda). \tag{42}$$

Given $\boldsymbol{\beta}$, $g(\boldsymbol{\beta}, \boldsymbol{\gamma})$ can be expressed as

$$\frac{1}{2}\|\boldsymbol{\gamma} - (\boldsymbol{\beta} - \nabla l(\boldsymbol{\beta}))\|_2^2 + P_\Theta(\boldsymbol{\gamma}; \lambda) + c(\boldsymbol{\beta}),$$

where $c(\boldsymbol{\beta})$ depends on $\boldsymbol{\beta}$ only.

Let $\hat{\boldsymbol{\beta}}$ be a Θ-estimator satisfying $\hat{\boldsymbol{\beta}} = \Theta(\hat{\boldsymbol{\beta}} - \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}^T\boldsymbol{y}; \lambda)$. Based on Lemma 1 and Lemma 2, we have

$$g(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}} + \boldsymbol{\Delta}) - g(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}) \geq \frac{1 - \mathcal{L}_\Theta}{2}\|\boldsymbol{\Delta}\|_2^2,$$

from which it follows that

$$f(\hat{\boldsymbol{\beta}} + \boldsymbol{\Delta}) - f(\hat{\boldsymbol{\beta}}) \geq \frac{1}{2}\boldsymbol{\Delta}^T(\boldsymbol{X}^T\boldsymbol{X} - \mathcal{L}_\Theta\boldsymbol{I})\boldsymbol{\Delta}.$$

This holds for any $\boldsymbol{\Delta} \in \mathbb{R}^p$. □

4.5.2. Proof of Lemma 4.

Let

$$\begin{aligned} l_H(\boldsymbol{\beta}) &= 2\langle \boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\beta}\rangle - \frac{1}{a}\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2 - \frac{1}{b}[P_H(\boldsymbol{\beta}; \sqrt{ab}A_0\lambda^o)] \\ l_0(\boldsymbol{\beta}) &= 2\langle \boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\beta}\rangle - \frac{1}{a}\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2 - \frac{1}{b}[P_0(\boldsymbol{\beta}; \sqrt{ab}A_0\lambda^o)], \end{aligned}$$

and $\mathcal{E}_H = \{\sup_{\boldsymbol{\beta}\in\mathbb{R}^p} l_H(\boldsymbol{\beta}) \geq at\sigma^2\}$, and $\mathcal{E}_0 = \{\sup_{\boldsymbol{\beta}\in\mathbb{R}^p} l_0(\boldsymbol{\beta}) \geq at\sigma^2\}$. Because $P_0 \geq P_H$, $\mathcal{E}_0 \subset \mathcal{E}_H$. We prove that $\mathcal{E}_H = \mathcal{E}_0$. The occurrence of $\mathcal{E}_H$ implies that $l_H(\boldsymbol{\beta}^o) \geq at\sigma^2$ for any $\boldsymbol{\beta}^o$ defined by

$$\boldsymbol{\beta}^o \in \arg\min_{\boldsymbol{\beta}} \frac{1}{a}\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2 - 2\langle \boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\beta}\rangle + \frac{1}{b}[P_H(\boldsymbol{\beta}; \sqrt{ab}A_0\lambda^o)],$$

With $a \geq 2b > 0$, Lemma 5 states that there exists at least one global minimizer $\boldsymbol{\beta}^{oo}$ satisfying $P_H(\boldsymbol{\beta}^{oo}; \sqrt{ab}A_1\lambda^o) = P_0(\boldsymbol{\beta}^{oo}; \sqrt{ab}A_1\lambda^o)$ and thus $l_H(\boldsymbol{\beta}^{oo}) = l_0(\boldsymbol{\beta}^{oo})$. This means that $\sup l_0(\boldsymbol{\beta}) \geq l_0(\boldsymbol{\beta}^{oo}) = l_H(\boldsymbol{\beta}^{oo}) \geq at\sigma^2$. So $\mathcal{E}_H \subset \mathcal{E}_0$, and it suffices to prove $\mathcal{E}_0^c$ occurs with high probability, or more specifically, $\mathbb{P}(\mathcal{E}_0) \leq C\exp(-ct)p^{-cA_1^2}$.

Given $1 \leq J \leq p$, define $\Gamma_J = \{\boldsymbol{\beta} \in \mathbb{R}^p : J(\boldsymbol{\beta}) = J\}$. Let $R = \sup_{1\leq J\leq p}\sup_{\boldsymbol{\beta}\in\Gamma_J}\{\langle \boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\beta}\rangle - \frac{1}{2b}P_0(\boldsymbol{\beta}; \sqrt{ab}A_1\lambda^o) - \frac{1}{2a}\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2\}$. We will use Lemma 6 to bound its tail probability.

Let $P_o'(J) = \sigma^2\{J + \log\binom{p}{J}\}$. We claim that

$$\mathbb{P}[\sup_{\boldsymbol{\beta}\in\Gamma_J}\{\langle \boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\beta}\rangle - \frac{1}{2a}\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2 - aLP_o'(J)\} > at\sigma^2] \leq C\exp(-ct). \tag{43}$$

Indeed,

$$\begin{aligned}
&2\langle \boldsymbol{\epsilon}, \boldsymbol{X\beta}\rangle - \frac{1}{a}\|\boldsymbol{X\beta}\|_2^2 - 2aLP_o'(J) \\
\leq &2\langle \boldsymbol{\epsilon}, \boldsymbol{X\beta}/\|\boldsymbol{X\beta}\|_2\rangle \|\boldsymbol{X\beta}\|_2 - 2\|\boldsymbol{X\beta}\|_2\sqrt{LP_o'(J)} - \frac{1}{2a}\|\boldsymbol{X\beta}\|_2^2 \\
= &2\|\boldsymbol{X\beta}\|_2\left(\langle \boldsymbol{\epsilon}, \boldsymbol{X\beta}/\|\boldsymbol{X\beta}\|_2\rangle - \sqrt{LP_o'(J)}\right) - \frac{1}{2a}\|\boldsymbol{X\beta}\|_2^2 \qquad (44)\\
\leq &2\|\boldsymbol{X\beta}\|_2\left(\langle \boldsymbol{\epsilon}, \boldsymbol{X\beta}/\|\boldsymbol{X\beta}\|_2\rangle - \sqrt{LP_o'(J)}\right)_+ - \frac{1}{2a}\|\boldsymbol{X\beta}\|_2^2 \\
\leq &2a\left(\langle \boldsymbol{\epsilon}, \boldsymbol{X\beta}/\|\boldsymbol{X\beta}\|_2\rangle - \sqrt{LP_o'(J)}\right)_+^2,
\end{aligned}$$

where the last inequality is due to Cauchy-Schwarz inequality. (43) now follows from Lemma 6.

Set $A_1 \geq 4\sqrt{L}$. We write $P_0(\boldsymbol{\beta};\lambda^o)$ with $\boldsymbol{\beta} \in \Gamma_J$ as $P_0(J;\lambda^o)$. Noticing some basic facts that (i) $P_o'(J) \leq CJ\log(ep) \leq CP_0(J;\lambda^o)$ due to Stirling's approximation, (ii) $\sqrt{(A_1^2/2)P_0(J;\lambda^o)} \geq \sqrt{LP_o'(J)} + \sqrt{cA_1^2P_0(J;\lambda^o)}$ for some $c>0$, and (iii) $J\log(ep) \geq \log p + J$ for any $J \geq 1$, we get

$$\begin{aligned}
&\mathbb{P}(R \geq a\sigma^2 t) \\
\leq &\sum_{J=1}^{p} \mathbb{P}\left(a \sup_{\boldsymbol{\beta}\in\Gamma_J}\left(\langle \boldsymbol{\epsilon}, \boldsymbol{X\beta}/\|\boldsymbol{X\beta}\|_2\rangle - \sqrt{(A_1^2/2)P_0(J;\lambda^o)}\right)_+^2 \geq a\sigma^2 t\right) \\
= &\sum_{J=1}^{p} \mathbb{P}(\sup_{\boldsymbol{\alpha}\in\Gamma'_J} \langle \boldsymbol{\epsilon}, \boldsymbol{\alpha}\rangle - \sqrt{(A_1^2/2)P_0(J;\lambda^o)} \geq \sigma\sqrt{t}) \\
\leq &\sum_{J=1}^{p} \mathbb{P}(\sup_{\boldsymbol{\alpha}\in\Gamma'_J} \langle \boldsymbol{\epsilon}, \boldsymbol{\alpha}\rangle - \sqrt{LP_o'(J)} \geq \sqrt{t}\sigma + \sqrt{cA_1^2P_0(J;\lambda^o)}) \\
\leq &\sum_{J=1}^{p} C\exp(-ct)\exp\{-cA_1^2(J+\log(p))\} \\
\leq &C\exp(-ct)\sum_{J=1}^{p}\exp(-cA_1^2\log p)\exp(-cA_1^2 J) \\
\leq &C\exp(-ct)p^{-cA_1^2},
\end{aligned}$$

where the last inequality due to the sum of geometric series. □

4.5.3. Proof of Lemma 5.

Similar to the proof of Lemma 3, we set $f_H(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + P_H(\boldsymbol{\beta};\lambda)$ with $l(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{X\beta} - \boldsymbol{y}\|_2^2$ and construct $g_H(\boldsymbol{\beta},\boldsymbol{\gamma}) = f_H(\boldsymbol{\gamma}) + \frac{1}{2}\|\boldsymbol{\gamma} - \boldsymbol{\beta}\|_2^2 - (l(\boldsymbol{\gamma}) - l(\boldsymbol{\beta}) - \langle\nabla l(\boldsymbol{\beta}), \boldsymbol{\gamma} - \boldsymbol{\beta}\rangle)$. Under $\|\boldsymbol{X}\|_2 \leq 1$, for any $(\boldsymbol{\beta},\boldsymbol{\gamma})$,

$$g_H(\boldsymbol{\beta},\boldsymbol{\gamma}) - f_H(\boldsymbol{\gamma}) = \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T(\boldsymbol{I} - \boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\gamma} - \boldsymbol{\beta}) \geq 0.$$

Let $\boldsymbol{\beta}^o$ be a globally optimal solution to $\min_{\boldsymbol{\beta}} f_H(\boldsymbol{\beta})$. Then $\boldsymbol{\gamma}^o := \Theta_H(\boldsymbol{\beta}^o - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}^o + \boldsymbol{X}^T\boldsymbol{y}; \lambda)$ gives

$$f_H(\boldsymbol{\gamma}^o) \le g_H(\boldsymbol{\beta}^o, \boldsymbol{\gamma}^o) \le g_H(\boldsymbol{\beta}^o, \boldsymbol{\beta}^o) = f_H(\boldsymbol{\beta}^o),$$

with the second inequality due to Lemma 1. Therefore, $\boldsymbol{\gamma}^o$ must also be a global minimizer of f_H, and by definition, $\boldsymbol{\gamma}^o$ demonstrates a threshold gap as desired. □

4.5.4. Proof of Lemma 6.

By definition, $\{\langle \boldsymbol{\epsilon}, \boldsymbol{\alpha}\rangle : \boldsymbol{\alpha} \in \Gamma'_J\}$ is a stochastic process with sub-Gaussian increments. The induced metric on Γ'_J is Euclidean: $d(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = \sigma\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_2$.

To bound the metric entropy $\log \mathcal{N}(\varepsilon, \Gamma'_J, d)$, where $\mathcal{N}(\varepsilon, \Gamma'_J, d)$ is the smallest cardinality of an ε-net that covers Γ'_J under d, we notice that $\boldsymbol{\alpha}$ is in a J-dimensional ball in $\mathbb{R}^p$. The number of such balls $\{\mathbf{P}_{\boldsymbol{X}_{\mathcal{J}}} \cap B_p(0,1) : \mathcal{J} \subset [p]\}$ is at most $\binom{p}{J}$, where $B_p(0,1)$ denotes the unit ball in $\mathbb{R}^p$. By a standard volume argument (see, e.g., [17]),

$$\log \mathcal{N}(\varepsilon, \Gamma'_{r,J}, d) \le \log \binom{p}{J} (\frac{C\sigma}{\varepsilon})^J = \log \binom{p}{J} + J \log(C\sigma/\varepsilon), \tag{45}$$

where C is a universal constant. The conclusion follows from Dudley's integral bound [15]. □

4.5.5. Proof of Lemma 7

We use the notation in the proof of Lemma 3 with g defined in (42). By Lemma 1 and Lemma 2, we obtain $g(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}) - g(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \ge \frac{1-\mathcal{L}_\Theta}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}\|_2^2$, namely,

$$\begin{aligned}\langle \nabla l(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(t+1)}\rangle + P_\Theta(\boldsymbol{\beta}) - P_\Theta(\boldsymbol{\beta}^{(t+1)}) + \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\|_2^2 \\ -\frac{1}{2}\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t+1)}\|_2^2 \ge \frac{1-\mathcal{L}_\Theta}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}\|_2^2.\end{aligned}$$

To cancel the first-order term, we give two other inequalities based on second-order lower/upper bounds:

$$\begin{aligned}&l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}^{(t)}) - \langle \nabla l(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\rangle \ge \frac{1}{2}\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}\|_{\boldsymbol{X}^T\boldsymbol{X}}^2, \\ &l(\boldsymbol{\beta}^{(t)}) + \langle \nabla l(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\rangle - l(\boldsymbol{\beta}^{(t+1)}) \ge -\frac{1}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_{\boldsymbol{X}^T\boldsymbol{X}}^2.\end{aligned}$$

Adding the three inequalities together gives the triangle inequality. □

Acknowledgement

The author would like to thank the editor, the associated editor and two anonymous referees for their careful comments and useful suggestions that improve the quality of the paper. The author also appreciates Florentina Bunea for the encouragement. This work was supported in part by NSF grant DMS-1352259.

References

[1] Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.*, 40(5):2452–2482. MR3097609

[2] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732. MR2533469

[3] Bunea, F., Tsybakov, A. B., and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194. MR2312149

[4] Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation via wavelet shrinkages. *Biometrika*, 81:425–455. MR1311089

[5] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360. MR1946581

[6] He, Y., She, Y., and Wu, D. (2013). Stationary sparse causality network learning. *J. Mach. Learn. Res.*, 14:3073–3104. MR3138910

[7] Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329. MR2906869

[8] Loh, P.-L. and Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.*, 16(1):559–616. MR3335800

[9] Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. (2011). Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, 39:2164–2204. MR2893865

[10] Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Prediction and Discovery (Contemporary Mathematics)*, 443:59–71. MR2433285

[11] Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239.

[12] She, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of Statistics*, 3:384–415. MR2501318

[13] She, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics and Data Analysis*, 9:2976–2990. MR2929353

[14] She, Y. (2014). Selective factor extraction in high dimensions. *arXiv preprint arXiv:1403.6212*.

[15] Talagrand, M. (2005). *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. Springer Monographs in Mathematics. Springer. MR2133757

[16] van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392. MR2576316

[17] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing*. MR2963170

[18] Zhang, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942. MR2604701

[19] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist*, 36:1567–1594. MR2435448

[20] Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high dimensional sparse estimation problems. *Statist. Sci.*, 27(4):576–593. MR3025135

[21] Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.*, 11:1081–1107. MR2629825

[22] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *JRSSB*, 67(2):301–320. MR2137327