Statistical and Computational Tradeoff of Regularized Dantzig-type Estimators

Xingguo Li, Yanbo Xu, Tuo Zhao, and Han Liu

Abstract

Nesterov’s smoothing technique has been widely applied to solve nonsmooth optimization problems involving high dimensional statistical models. However, existing theory focuses more on its computational properties rather than statistical properties. This paper bridges this gap by studying a family of regularized Dantzig-type estimators. For these estimators, we show that the smoothing technique obtains improved computational performance without sacrificing statistical accuracy. Thorough numerical results are provided to support our theory.

1 Introduction

High dimensionality poses significant challenges to data analysis and has motivated the development of a family of regularized estimators, including the Lasso, group Lasso, graphical Lasso, and reduced rank regression (Tibshirani, 1996; Yuan and Lin, 2005; Banerjee et al., 2008; Liu and Luo, 2012; Negahban et al., 2012). More concretely, let \( \theta^* \in \mathbb{R}^d \) be the unknown parameter under some statistical model, then a generic regularized M-estimator can be obtained by solving an unconstrained minimization problem

\[
\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} L(\theta) + \alpha R(\theta),
\]

where \( L(\cdot) \) is a convex loss function, \( R(\cdot) \) is a norm-based regularization function, and \( \alpha > 0 \) is a tuning parameter balancing the estimation bias and variance. Statistical properties have been well established for a large family of estimators (Negahban et al., 2012). Effective algorithms, such
as proximal gradient, have been proposed by exploiting the composite structure of the objective function in (1.1) (Nesterov, 2007; Beck and Teboulle, 2009a).

More recently, statisticians have proposed another family of estimators obtained by solving constrained optimization problems. Examples include the Dantzig selector, group Dantzig selector, reduced rank Dantzig selector, CLIME (Candes and Tao, 2007; Bickel et al., 2009; Liu et al., 2010; Candes and Plan, 2011; Cai et al., 2011; Gautier and Tsybakov, 2011, 2013; Cai et al., 2014). More specifically, let $R^*(\cdot)$ be the dual norm of $R(\cdot)$, i.e., for any $u \in \mathbb{R}^d$, we have

$$R^*(u) = \max_{v \in \mathbb{R}^d} u^T v \text{ subject to } R(v) \leq 1.$$ 

A generic Dantzig-type estimator can be obtained by solving the constrained minimization problem

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} R(\theta) \text{ subject to } R^*(\nabla L(\theta)) \leq \nu, \quad (1.2)$$

where $\nu > 0$ is a perturbation parameter controlling the bias and variance tradeoff. In general, $R(\cdot)$ and $R^*(\cdot)$ are both nonsmooth. Compared to (1.1), the constrained program in (1.2) has its unique advantages. For example, when extending the regularized M-estimators to the semiparametric graph estimation problem or regression problem with missing values, Zhao et al. (2014) and Loh and Wainwright (2012) show that (1.1) is possibly nonconvex, whereas the optimization problem in (1.2) is always convex.

Nevertheless, solving the Dantzig-type optimization problem in (1.2) is computationally challenging in general due to the nonsmoothness of $R(\cdot)$ and $R^*(\cdot)$. Though solvers are developed for some Dantzig-type optimization problems, they either lack non-asymptotic iteration complexity analyses or cannot scale to large problems (James et al., 2009; Boyd et al., 2011; Wang and Yuan, 2012; Li et al., 2015). On one hand, existing linear programming solvers are only applicable to sparse linear regression, i.e., $R(\cdot)$ is the $\ell_1$ norm and $R^*(\cdot)$ is the $\ell_\infty$ norm. However, our interests are far beyond the simple sparse linear model. Typical examples include structural sparse linear model, reduced rank regression model, sparse regression with missing entries, etc. These models cannot be cast as linear programs due to the existence of nonlinear terms, such as the $\ell_{\infty,2}$ norm constraint in sparse multivariate regression and the spectral norm constraint in reduced rank regression. On the other hand, existing generic solvers for nonsmooth problems, such as subgradient algorithms and alternating direction method of multipliers (ADMM) (Boyd and Vandenberghe, 2004; Combettes and Pesquet, 2007; Ouyang et al., 2013), have poor iteration complexities. Another generic solver, the interior point method, does not fit to solve nonsmooth problems and has high computational cost per iteration for large $d$, e.g., $O(d^3)$ for solving a $d$-dimensional inverse problem, even for smooth problems (Nocedal and Wright, 2006). These motivate us to develop a unified approach to solve the Dantzig-type problems to achieve both favored computational and statistical performance, via adopting the Nesterov’s smoothing technique (Nesterov, 2005).
To overcome these computational drawbacks of the Dantzig-type estimators, we propose to solve the regularized form of (1.2):

$$\tilde{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} R^\ast\left(\nabla L(\theta)\right) + \tau R(\theta), \quad (1.3)$$

where $\tau > 0$ is a regularization parameter. As an anonymous reviewer points out that the estimators in (1.2) and (1.3) are equivalent in the sense that they share the same solution for a pair of suitably chosen $\nu$ and $\tau$. But unlike (1.2), the optimization problem in (1.3) naturally fits into the Nesterov’s smoothing optimization framework for unconstrained minimization (Nesterov, 2005; Beck and Teboulle, 2009a, 2012; Chen et al., 2012): First, (1.3) is replaced by its partially smooth approximation

$$\tilde{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} R^\ast_\mu\left(\nabla L(\theta)\right) + \tau R(\theta), \quad (1.4)$$

where $R^\ast_\mu(\nabla L(\theta))$ is a smooth approximation of $R^\ast(\nabla L(\theta))$ with a smoothing parameter $\mu$; Second, an accelerated proximal gradient algorithm (Beck and Teboulle, 2009a) is applied to solve (1.4). We smooth $R^\ast(\nabla L(\theta))$ instead of $R(\theta)$ or both for the following three reasons:

(a) Smoothing $R^\ast(\theta)$ also enables us to gain additional computational efficiency by introducing a large approximation error to the original optimization, but with almost no sacrifice of statistical accuracy;

(b) The accelerated proximal gradient algorithm has a closed form solution at each iteration for a wide class of $R(\cdot)$;

(c) Keeping $R(\theta)$ yields a sparse or low rank solution at each iteration, which brings additional computational convenience.

As can be seen, (b) and (c) are straightforward and already well known to optimization researchers, but (a) is rarely noticed in existing literature. This is because the smoothing technique is usually only considered as a tradeoff between computational efficiency and approximation accuracy to the original optimization problem. From a statistical perspective, however, we are more interested in the tradeoff between the statistical accuracy of an estimator and its computational efficiency (Zhao et al., 2014). This motivates us to directly analyze the statistical properties of $\tilde{\theta}$ rather than $\hat{\theta}$. More specifically, our new results allow a larger smoothing approximation error than those in existing literature, but the resulting estimators still possess good statistical properties. Thus the smoothing technique simultaneously enjoys good computational efficiency and statistical accuracy. Our analysis of the statistical property is inspired by Gautier and Tsybakov (2011); Negahban et al. (2012), and similar analysis also appears in Gautier and Tsybakov (2013). However, our paper focuses more on the computational aspect of the problem, which is different from the focus of Gautier and Tsybakov (2013). We present numerical experiments on both simulated and real datasets to support our theory. For notational simplicity, we call $\tilde{\theta}$ in (1.4) the regularized Dantzig-type (RDE) estimators throughout the rest of the paper.
There existlems with two assumptions as follows:

Assumption 2.1. There exist $H \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}^d$, independent on $\theta$, such that $\nabla L(\theta) = H\theta - b$.

Assumption 2.2. For every $\theta \in \mathbb{R}^d$, the regularization function $R(\theta)$ satisfies $\|\theta\|_2 \leq R(\theta)$.

These assumptions are mild, and are satisfied by estimators for sparse linear regression, sparse multivariate regression, reduced rank regression, and sparse precision matrix estimation.

Example 2.3 (Sparse Linear Regression (Tibshirani, 1996)). Given a response vector $y \in \mathbb{R}^n$ and a design matrix $X \in \mathbb{R}^{n \times d}$, we consider a linear model $y = X\theta^* + w$, where $\theta^* \in \mathbb{R}^d$ is the regression coefficient vector, and $w \in \mathbb{R}^n$ is a noise vector with each entry independently sampled from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. We assume that $\theta^*$ has at most $s$ nonzero entries. To estimate $\theta^*$, we consider the following loss and regularization functions

$$L(\theta) = \frac{1}{2n}||y - X\theta||_2^2 \quad \text{and} \quad R(\theta) = ||\theta||_1.$$  

Accordingly, we have

$$H = \frac{1}{n}X^TX, \quad b = \frac{1}{n}X^Ty, \quad \text{and} \quad R^*(\nabla L(\theta)) = \frac{1}{n}||X^T(y - X\theta)||_\infty.$$  

(2.1)
Remark 2.4 (Sparse Linear Regression with Missing Data). Loh and Wainwright (2012); Wang et al. (2017) propose variants of sparse linear regression when there exists contamination in the data, including additive noise, missing data, and multiplicative noise. Taking the case with missing data as an example, they assume that the entries of $X$ are missing randomly and independently with some constant probability $\rho \in [0,1)$. In particular, they observe the matrix $Z \in \mathbb{R}^{n \times d}$ with entries

$$Z_{ij} = \begin{cases} X_{ij}, & \text{with probability } 1 - \rho, \\ 0, & \text{otherwise}, \end{cases}$$

where $|Z_{ij} = 0|$ denotes the missing entries. Given the observed matrix $Z$, they define

$$\Gamma_{\text{mis}} = \frac{Z^\top Z}{n} - \bar{\rho} \cdot \text{diag} \left( \frac{Z^\top Z}{n} \right) \quad \text{and} \quad \gamma_{\text{mis}} = \frac{1}{n} Z^\top y,$$

where $\bar{\rho} = \frac{\sum_{ij} I(Z_{ij} = 0)}{nd}$ and $Z_{ij} = Z_{ij}/(1 - \bar{\rho})$. To estimate $\theta^*$, they consider the loss and regularization functions

$$\mathcal{L}(\theta) = \frac{1}{2} \theta^\top \Gamma_{\text{mis}} \theta - \gamma_{\text{mis}}^\top \theta \quad \text{and} \quad \mathcal{R}(\theta) = \|\theta\|_1.$$ 

Accordingly, we have

$$H = \Gamma_{\text{mis}}, \quad b = \gamma_{\text{mis}}, \quad \text{and} \quad \mathcal{R}^*(\nabla \mathcal{L}(\theta)) = \|\Gamma_{\text{mis}} \theta - \gamma_{\text{mis}}\|_\infty.$$

Note that $\mathcal{R}^*(\nabla \mathcal{L}(\theta))$ in (2.3) is convex, though $\Gamma_{\text{mis}}$ may not be positive semidefinite. In contrast, directly plugging the above $\mathcal{L}(\theta)$ into (1.1) results in a possibly nonconvex optimization problem.

Example 2.5 (Sparse Multivariate Regression (Yuan and Lin, 2005)). Given a multivariate response matrix $Y \in \mathbb{R}^{n \times m}$ and a design matrix $X \in \mathbb{R}^{n \times d}$, we consider a multivariate linear model $Y = X \Theta^* + W$, where $\Theta^* \in \mathbb{R}^{d \times m}$ is the regression coefficient matrix, and $W \in \mathbb{R}^{n \times m}$ is a noise matrix with each entry independently sampled from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. We assume that $\Theta^*$ has at most $s$ rows with nonzero entries. To estimate $\Theta^*$, we consider the following loss and regularization functions

$$\mathcal{L}(\Theta) = \frac{1}{2n} \|Y - X \Theta\|_F^2 \quad \text{and} \quad \mathcal{R}(\Theta) = \|\Theta\|_{1,2}.$$ 

Accordingly, we have

$$H = \frac{1}{n} X^\top X, \quad b = \frac{1}{n} X^\top Y, \quad \text{and} \quad \mathcal{R}^*(\nabla \mathcal{L}(\Theta)) = \frac{1}{n} \Xi^\top (Y - X \Theta)_{\|0,2}.$$ 

Note that $b$ is a matrix instead of a vector, but it can be straightforwardly converted to the standard vector form as in Assumption 2.1 by the vectorization trick. See more details in Remark 2.7.
Example 2.6 (Reduced Rank Regression (Fazel, 2002)). Given a multivariate response matrix \( Y \in \mathbb{R}^{n \times m} \) and a design matrix \( X \in \mathbb{R}^{n \times d} \), we consider a multivariate linear model \( Y = X\Theta^* + W \), where \( \Theta^* \in \mathbb{R}^{d \times m} \) is the regression coefficient matrix, and \( W \in \mathbb{R}^{n \times m} \) is a noise matrix with each entry independently sampled from a Gaussian distribution \( \mathcal{N}(0, \sigma^2) \). We assume that \( \Theta^* \) has at most \( r \) nonzero singular values. To estimate \( \Theta^* \), we consider the following loss and regularization functions

\[
\mathcal{L}(\Theta) = \frac{1}{2n} \| Y - X\Theta \|_F^2 \quad \text{and} \quad \mathcal{R}(\Theta) = \| \Theta \|_*.
\]

Accordingly, we have

\[
H = \frac{1}{n}X^TX, \quad b = \frac{1}{n}X^TY, \quad \text{and} \quad \mathcal{R}'(\nabla \mathcal{L}(\Theta)) = \| \frac{1}{n}X^T(Y - X\Theta) \|_2. \tag{2.5}
\]

Remark 2.7. The parameter \( \Theta \) is a matrix for sparse multivariate regression and reduced rank regression. An alternative representation is to concatenate \( \Theta \) as a vector \( \theta = \text{vec}(\Theta) = (\Theta_{11}^T, \ldots, \Theta_{1m}^T)^T \). For notational simplicity, we also define the inverse operation as \( \Theta = \text{mat}(\theta) \). Accordingly, all \( \mathcal{L}(\Theta), \mathcal{R}(\Theta), \) and \( \mathcal{R}'(\nabla \mathcal{L}(\Theta)) \) can be rewritten using the vectorized form \( \text{vec}(\Theta) \). See more details in Negahban et al. (2012). Throughout the rest of the paper, if not clearly specified, we use \( \Theta \) to denote the vectorized form of \( \Theta \) for sparse multivariate regression and reduced rank regression.

Example 2.8 (Sparse Precision Matrix Estimation (Cai et al., 2011)). Given a data matrix \( X \in \mathbb{R}^{n \times d} \) with all \( X_{i\ast} \)'s independently sampled from a \( d \)-dimensional Gaussian distribution with mean 0 and covariance \( \Sigma^* = (\Theta^*)^{-1} \in \mathbb{R}^{d \times d} \). We assume that \( \| \Theta^* \|_1 \leq M \) and each \( \Theta_{ij}^* \) has at most \( s \) nonzero entries. To estimate \( \Theta_{ij}^* \), we consider the following loss and regularization functions

\[
\mathcal{L}(\Theta) = \frac{1}{2} \Theta_{ij}^T\tilde{S}\Theta_{ij} - \Theta_{ij}^T\Theta_{ij} \quad \text{and} \quad \mathcal{R}(\Theta_{ij}) = \| \Theta_{ij} \|_1,
\]

where \( \tilde{S} = X^TX/n \) is the empirical covariance matrix. Accordingly, we have

\[
H = \tilde{S}, \quad b = I_{ij}, \quad \text{and} \quad \mathcal{R}'(\nabla \mathcal{L}(\Theta_{ij})) = \| I_{ij} - \tilde{S}\Theta_{ij} \|_{\infty}. \tag{2.6}
\]

Note that it is possible to estimate \( \Theta^* \) directly (Zhang and Zou, 2014), but the column-wise estimation is computationally more efficient and it further guarantees the optimal statistical rate of convergence under the \( \ell_1 \) norm (Yuan, 2010).

Remark 2.9 (Sparse Latent Precision Matrix Estimation). Liu et al. (2012) extend the sparse precision matrix estimation to the nonparanormal distributions. They assume that the data follow a Gaussian distribution after a set of strictly increasing marginal transformations \( \{ f_j \}_{j=1}^d \),

\[
(f_1(X_1), \ldots, f_d(X_d))^\top \sim \mathcal{N}(0, \Sigma^*).
\]

Though the nonparanormal family is larger than the Gaussian family, they show that the conditional independence relationship of \( X \) is still encoded by the sparsity pattern of \( \Theta^* = (\Sigma^*)^{-1} \). They
call $\Sigma^*$ and $\Theta^*$ the latent covariance and precision matrices of the nonparanormal distribution. Liu et al. (2012) further show that $\Theta^*$ can be estimated by directly plugging the transformed Spearman’s rho estimator into existing sparse precision matrix estimation procedures. More specifically, let $x_1, \ldots, x_n \in \mathbb{R}^d$ be $n$ independent observations of $X$, where $x_i = (x_{i1}, \ldots, x_{id})^\top$, they first calculate the Spearman’s rho statistics between $X_k$ and $X_j$ as

$$\hat{\rho}_{kj} = \frac{\sum_{i=1}^n (r_i^j - \bar{r}_j)(r_i^k - \bar{r}_k)}{\sqrt{\sum_{i=1}^n (r_i^j - \bar{r}_j)^2} \cdot \sum_{i=1}^n (r_i^k - \bar{r}_k)^2},$$

where $r_{ij}$ denotes the rank of $x_{ij}$ among $x_{1j}, \ldots, x_{nj}$ and $\bar{r}_j = n^{-1} \sum_{i=1}^n r_{ij} = (n + 1)/2$. Then the transformed Spearman’s rho estimator can be obtained by

$$\tilde{S} = [\tilde{S}_{kj}] = \left[2\sin\left(\frac{\pi}{6} \hat{\rho}_{kj}\right)\right].$$

At last, $\Theta^*$ can be estimated by plugging the transformed Spearman’s rho estimator into the loss and regularization functions defined in Example 2.8. Note that (1.3) is convex even when $\tilde{S}$ is indefinite. However, directly plugging the above $\mathcal{L}(\theta)$ into (1.1) may result in a nonconvex problem.

### 3 Smoothing and Accelerated Proximal Gradient Algorithm

Recall that the proposed RDE estimators are obtained by solving the following partially smooth optimization problem

$$\tilde{\theta} \in \arg\min_{\theta} \mathcal{R}^*_\mu(\nabla \mathcal{L}(\theta)) + \tau \mathcal{R}(\theta).$$

We first explain how to obtain $\mathcal{R}^*_\mu(\nabla \mathcal{L}(\theta))$ using Nesterov’s smoothing technique.

#### 3.1 Smoothed Approximation

We denote the Fenchel’s dual representation of $\mathcal{R}^*(\nabla \mathcal{L}(\theta))$ as

$$\mathcal{R}^*(\nabla \mathcal{L}(\theta)) = \max_{\mathcal{R}(u) \leq 1} u^\top \nabla \mathcal{L}(\theta).$$

Given a smoothing parameter $\mu > 0$, we construct the smooth approximation of $\mathcal{R}^*(\nabla \mathcal{L}(\theta))$ as

$$\mathcal{R}^*_\mu(\nabla \mathcal{L}(\theta)) = \max_{\mathcal{R}(u) \leq 1} u^\top \nabla \mathcal{L}(\theta) - \frac{\mu}{2} \|u\|_2^2 = \max_{\mathcal{R}(u) \leq 1} -\frac{\mu}{2} \|u - \frac{1}{\mu} \nabla \mathcal{L}(\theta)\|_2^2 + \frac{1}{2\mu} \|\nabla \mathcal{L}(\theta)\|_2^2,$$

where $\|u\|_2^2$ is a proximity function. Combining (3.2) and (3.3), for any $\theta$, we have

$$\mathcal{R}^*(\nabla \mathcal{L}(\theta)) - \mu/2 \leq \mathcal{R}^*_\mu(\nabla \mathcal{L}(\theta)) \leq \mathcal{R}^*(\nabla \mathcal{L}(\theta))$$

where $\mathcal{R}(\cdot)$ satisfies $\|u\|_2^2 \leq \mathcal{R}(u)^2 \leq 1$. Particularly, we define

$$\tilde{u}^\theta = \arg\min_{u} \|u - \nabla \mathcal{L}(\theta)/\mu\|_2^2 \quad \text{subject to } \mathcal{R}(u) \leq 1.$$

Next, we provide several illustrative examples as follows.
Example 3.1 (Sparse Linear Regression). We have $\|u\|_2^2 \leq \|u\|_1^2 \leq 1$ satisfying Assumption 2.2. The optimization problem (3.3) can be rewritten as

$$\hat{u}^\theta = \arg\min_u \|u - \nabla L(\theta) / \mu\|_2^2 \text{ subject to } \|u\|_1 \leq 1,$$

(3.5)

which is an $\ell_1$-ball projection problem. An efficient algorithm for solving (3.5) with the worst-case complexity of $O(d)$ has been provided in Liu and Ye (2009).

Example 3.2 (Sparse Multivariate Regression). We have $\|U\|_F^2 \leq \|U\|_{1,2}^2 \leq 1$ satisfying Assumption 2.2. The optimization problem (3.3) can be rewritten as

$$\hat{U}^\Theta = \arg\min_U \|U - \nabla L(\Theta) / \mu\|_F^2 \text{ subject to } \|U\|_{1,2} \leq 1,$$

(3.6)

which is an $\ell_{1,2}$-ball projection problem. An algorithm for solving (3.6) is provided in Appendix A.

Example 3.3 (Reduced Rank Regression). We have $\|U\|_F^2 \leq \|U\|_*^2 \leq 1$ satisfying Assumption 2.2. The optimization problem (3.3) can be rewritten as

$$\hat{U}^\Theta_{:j} = \arg\min_u \|u - \nabla L(\Theta)_{:j} / \mu\|_2^2 \text{ subject to } \|u\|_1 \leq 1,$$

(3.7)

which is also an $\ell_1$-ball projection problem.

From (3.4), we see that the approximation error introduced by the smoothing technique can be controlled by a suitably chosen $\mu$. Figure 1 presents several two-dimensional examples of the $\ell_\infty$ norm smoothed by different values of $\mu$. Note that the $\ell_\infty$ norm is used as $R^*(\cdot)$ in the sparse linear regression and sparse precision matrix estimation problems.

The next lemma shows that $R_\mu^*(\nabla L(\theta))$ is smooth in $\theta$ with a simple form of gradient.

Lemma 3.5. Suppose Assumptions 2.1 and 2.2 hold. For any $\mu > 0$, $R_\mu^*(\nabla L(\theta))$ is a convex and continuously differentiable function in $\theta$. Moreover, $g_\mu(\theta)$ — the gradient of $R_\mu^*(\nabla L(\theta))$ with respect to $\theta$ — takes the form

$$g_\mu(\theta) = \frac{\partial \left[ \nabla L(\theta)^\top \hat{u}^\theta - \mu \|\hat{u}^\theta\|_2^2 / 2 \right]}{\partial \theta} = \nabla^2 L(\theta)^\top \hat{u}^\theta = H^\top \hat{u}^\theta,$$

(3.8)

where $H$ is defined in Assumption 2.1. Given $\phi = \|\nabla^2 L(\theta)\|_2 = \|H\|_2^2$, $g_\mu(\theta)$ is Lipschitz continuous in $\theta$ with the Lipschitz constant $\phi / \mu$. That is, for any $\theta, \theta' \in \mathbb{R}^d$, we have

$$\|g_\mu(\theta) - g_\mu(\theta')\|_2 \leq \phi \|\theta - \theta'\|_2 / \mu.$$

Lemma 3.5 is a direct result of Theorem 1 in Nesterov (2005) and enables us to exploit the accelerated proximal gradient algorithm to solve (3.1).
Figure 1: The $\ell_\infty$ norm ($\mu = 0$) and its smooth approximations with $\mu = 0.1, 0.25, 0.5$. A larger $\mu$ introduces a larger approximation error but makes the approximation smoother.

3.2 Accelerated Proximal Gradient Algorithm

The main idea of the accelerated proximal gradient algorithm is that, within each iteration, we exploit the gradient information in the previous and current iterations to find a better descent direction than the current anti-gradient direction. Here we adopt the monotone fast iterative shrinkage thresholding algorithm (MFISTA) proposed in Beck and Teboulle (2009a), because Beck and Teboulle (2009a) show that it usually outperforms FISTA (Beck and Teboulle, 2009b) in practice.

To derive the algorithm, we define two sequences of auxiliary variables $\{\vartheta(t)\}$ and $\{\beta(t)\}$, a sequence of weights $\{w_t\}$, and a nonincreasing sequence of step sizes $\{\eta_t > 0\}$. In particular, we choose the initial weight $w_1 = 1$, and then update it by $w_{t+1} = (1 + \sqrt{1 + 4w_t^2})/2$ at each iteration.

At the $t$th iteration, we consider a quadratic approximation of $R^*_\mu(\nabla L(\beta))$ as

$$Q(\beta, \vartheta(t), \eta_t) = R^*_\mu(\nabla L(\vartheta(t))) + g_\mu(\vartheta(t))\top(\beta - \vartheta(t)) + \frac{1}{2\eta_t}\|\beta - \vartheta(t)\|_2^2.$$  

Let $\tilde{\beta}(t) = \vartheta(t) - \eta_t g_\mu(\vartheta(t))$. We then take

$$\beta(t) = \arg\min_\beta Q(\beta, \vartheta(t), \eta_t) + \tau R(\beta) = \arg\min_\beta \frac{1}{2}\|\beta - \tilde{\beta}(t)\|_2^2 + \eta_t \tau R(\beta) = \mathcal{P}_{\eta_t\tau}(\tilde{\beta}(t)).$$  \hspace{1cm} (3.9)

$\mathcal{P}_{\eta_t\tau}(\cdot)$ defined in (3.9) is often referred as the proximal operator in existing literature (Parikh and Boyd, 2013). For a broad class of $R(\cdot)$, the proximal operator has a closed form expression.

**Example 3.6** (Sparse Linear Regression, Sparse Precision Matrix Estimation). From (3.9), we have

$$\hat{\beta}_j^{(t)} = \arg\min_\beta \frac{1}{2}\|\beta - \tilde{\beta}_j^{(t)}\|_2^2 + \eta_t \tau \|\beta\|_1 = \text{sign}(|\tilde{\beta}_j|) \cdot \max\{|\tilde{\beta}_j| - \eta_t \tau, 0\},$$  

which is also called the soft-thresholding operation (Friedman et al., 2007).

**Remark 3.7.** We provide some discussion on the reason of smoothing $R^*_\mu(\cdot)$ instead of $R(\cdot)$. Taking the sparse linear regression as an example, if we smooth $R(\cdot)$ instead of $R^*_\mu(\cdot)$, we will lose the solution sparsity and the closed form expression for the accelerated proximal gradient algorithm at each iteration. Particularly, we consider a smooth approximation of $R(\cdot)$ denoted by $R^*_\mu(\cdot)$, and (3.9) is replaced with

$$\beta(t) = \arg\min_\beta \|\beta - \tilde{\beta}(t)\|_2^2 + \eta_t \tau \|X^\top X\beta - X^\top y\|_{\infty},$$  \hspace{1cm} (3.10)
we take the step size as $\eta$. Suppose that Assumptions 2.1 and 2.2 hold, Theorem 3.10. The following theorem establishes the iteration complexity of the proposed algorithm.

### 3.3 Convergence Analysis

We analyze the accelerated proximal gradient algorithm in Algorithm 1.

**Example 3.8 (Sparse Multivariate Regression).** Let $\overline{B}^{(t)} = \text{mat}(\overline{\beta}^{(t)})$. For all $j$, we have

$$B_{sj}^{(t)} = \arg\min_{B_{sj}} \frac{1}{2} \| B_{sj} - \overline{B}_{sj}^{(t)} \|_2^2 + \eta_t \tau \| B_{sj} \|_2,$$

which has a closed form solution $B_{sj}^{(t)} = \max(\| B_{sj}^{(t-1)} \|_2 - \eta_t \tau, 0) \overline{B}_{sj}^{(t-1)}$. Once $B^{(t)}$ is obtained, we take $\beta^{(t)} = \text{vec}(B^{(t)})$.

**Example 3.9 (Reduced Rank Regression).** Let $\overline{B}^{(t)} = \text{mat}(\overline{\beta})$. We have

$$B^{(t)} = \arg\min_B \frac{1}{2} \| B - \overline{B}^{(t)} \|_F^2 + \eta_t \tau \| B \|_*.$$

Given a compact singular value decomposition $\overline{B}^{(t)} = \sum_{j=1}^r \sigma_j(\overline{B}^{(t)}) v_j w_j^T$, where $r$ is the rank of $\overline{B}^{(t)}$, $v_j$ and $w_j$ are singular vectors corresponding to $\sigma_j(\overline{B}^{(t)})$, we then take $B^{(t)} = \sum_{j=1}^r \max(\| \sigma_j(\overline{B}^{(t)}) - \eta_t w_j, 0) v_j w_j^T$. Once $B^{(t)}$ is obtained, we take $\beta^{(t)} = \text{vec}(B^{(t)})$.

A conservative choice of $\{\eta_t\}$ is $\eta_t = \mu / \phi$ for all $t$. To further improve the empirical performance, we can use the backtracking line search to dynamically adjust $\eta_t$. More specifically, we start with a large enough $\eta_0$, and then choose the minimum nonnegative integer $q$ within each iteration such that $Q(\beta^{(t)}, \theta^{(t)}, \eta_t) \geq R^*_h(\beta^{(t)})$, where $h \in (0, 1)$ is the shrinkage parameter and $\eta_t = h^q \eta_{t-1}$. To ensure the nonincrease of the objective value, we take

$$\theta^{(t)} = \arg\min_\theta R^*_h(\nabla L(\theta)) + \tau R(\theta) \text{ subject to } \theta \in [\beta^{(t)}, \theta^{(t-1)}]. \quad (3.11)$$

At last, we take

$$\theta^{(t+1)} = \theta^{(t)} + \frac{w_t}{w_{t+1}} (\beta^{(t)} - \theta^{(t)}) + \frac{(w_t - 1)}{w_{t+1}} (\theta^{(t)} - \theta^{(t-1)}). \quad (3.12)$$

The algorithm terminates when $\| \theta^{(t+1)} - \theta^{(t)} \|_2 \leq \xi$, where $\xi$ is the stopping precision. We summarize the accelerated proximal gradient algorithm in Algorithm 1.

### 3.3 Convergence Analysis

The following theorem establishes the iteration complexity of the proposed algorithm.

**Theorem 3.10.** Suppose that Assumptions 2.1 and 2.2 hold, $H$ is defined in Assumption 2.1, and we take the step size as $\eta_t = \mu / \| H \|_2^2$ for all $t$. Given a pre-specified accuracy $\epsilon$, we need at most

$$t = \sqrt{\frac{4 \| H \|_2^2 \cdot \| \theta^{(0)} - \overline{\theta} \|_2^2}{\epsilon \mu}} - 1 = O\left( \frac{1}{\sqrt{\epsilon} \mu} \right) \quad (3.13)$$
Algorithm 1 The Accelerated Proximal Gradient Algorithm.

**Input:** \( \vartheta^{(0)} = \beta^{(0)} = \theta^{(0)} \), \( \mu \), \( w_0 \), \( h \) and \( \xi \).

**Initialize:** \( t = 1 \) and \( \vartheta^{(1)} = \vartheta^{(0)} = \beta^{(0)} = \theta^{(0)} = 0 \).

**repeat**

1. Evaluate the gradient \( g^\mu(\vartheta^{(t)}) \) using (3.8).
2. Compute the auxiliary variable \( \beta^{(t)} \) using (3.9).
3. (Optional) Compute \( \eta_t \) by the backtracking line search.
4. Compute the weight \( w_t \).
5. Compute the solution \( \theta^{(t)} \) using (3.11).
6. Compute the auxiliary variable \( \vartheta^{(t+1)} \) using (3.12).
7. \( t = t + 1 \).

**until Convergence**

**Output:** \( \theta^{(t)} \)

iterations such that
\[
R^*_\mu(\nabla \mathcal{L}(\theta^{(t)})) + \tau R(\theta^{(t)}) \leq R^*_\mu(\nabla \mathcal{L}(\tilde{\theta})) + \tau R(\tilde{\theta}) + \frac{\varepsilon}{2} \quad \text{and} \quad R^*(\nabla \mathcal{L}(\theta^{(t)})) + \tau R(\theta^{(t)}) \leq R^*(\nabla \mathcal{L}(\tilde{\theta})) + \tau R(\tilde{\theta}) + \mu + \frac{\varepsilon}{2}.
\]
(3.14)

Moreover, if we choose \( \mu = \varepsilon/2 \), then
\[
R^*(\nabla \mathcal{L}(\theta^{(t)})) + \tau R(\theta^{(t)}) \leq R^*(\nabla \mathcal{L}(\tilde{\theta})) + \tau R(\tilde{\theta}) + \varepsilon \quad \text{and} \quad t = O\left(\frac{1}{\sqrt{\varepsilon}}\right).
\]
(3.15)

Theorem 3.10 is a direct result of Beck and Teboulle (2009a)\(^1\), and we provide the proof in Section 7.1. Existing literature considers \( \tilde{\theta} \) as an inexact solution to the original optimization problem (1.3), therefore they have to choose \( \mu = \varepsilon/2 \) to avoid a large approximation error (Nesterov, 2005; Beck and Teboulle, 2012; Chen et al., 2012; Zhao and Liu, 2012). This eventually results in an iteration complexity of \( O(1/\varepsilon) \), which is optimal over all first-order algorithms for minimizing (1.3). This is known as the tradeoff of the smoothing technique between computational efficiency and approximation error.

However, from a statistical perspective, we are more interested in the tradeoff between the statistical accuracy of the resulting estimator and the computational efficiency. This motivates us to directly analyze the statistical properties of \( \tilde{\theta} \) rather than \( \tilde{\theta} \). The next section shows that even though we choose a larger \( \mu \) than those in existing literature, \( \tilde{\theta} \) still possesses good statistical properties. Since we are directly considering a partially smooth formulation in (1.4), which is easier to solve than (1.3), the accelerated proximal gradient algorithm achieves an “improved” iteration complexity of \( O\left(\frac{1}{\sqrt{\varepsilon \mu}}\right) = O\left(\frac{1}{\sqrt{\varepsilon}}\right)^2 \), e.g., by choosing a constant \( \mu \) such that \( \mu \gg \varepsilon \).

\(^1\)Theorem 5.1 in Beck and Teboulle (2009a) also works with the backtracking line search procedure. Please refer to Remark 3.1 in Beck and Teboulle (2009a).

\(^2\)The accelerated proximal gradient algorithm achieves the optimal iteration complexity for minimizing (1.4) over
4 Statistical Properties

We start with an introduction of some necessary tools for our theoretical analysis.

Definition 4.1 (Decomposability of $R(\cdot)$, Negahban et al. (2012)). Let $S$ and $N$ be a pair of subspaces of $\mathbb{R}^d$, which are orthogonal to each other. A regularization function $R(\cdot)$ is decomposable with respect to $(S, N)$ if for any $\theta \in S$ and $\theta' \in N$, we have

$$R(\theta + \theta') = R(\theta) + R(\theta').$$

Given the decomposability of $R(\cdot)$ and some $\gamma > 0$, we define a cone shape set as follows:

$$M_\gamma = \{\delta \in \mathbb{R}^d : R(\delta_N) \leq (1 + \gamma) R(\delta_{N^\perp})\}.$$  \quad (4.1)

To establish the statistical properties of the RDE estimators, we impose the following restricted eigenvalue condition on the Hessian matrix $\nabla^2 L(\theta^*)$, where $\theta^*$ denotes the true parameter of the underlying statistical model.

Assumption 4.2 (Restricted Eigenvalue Condition). There exit two universal positive constants $\kappa$ and $\gamma$ such that

$$\min_{\delta \neq 0, \delta \in M_\gamma} \frac{\delta^\top \nabla^2 L(\theta^*) \delta}{||\delta||_2^2} \geq \kappa.$$

Assumption 4.2 has been extensively studied for analyzing the statistical properties of the regularized M-estimators and Dantzig-type estimators (Bickel et al., 2009; Raskutti et al., 2010; Negahban et al., 2012), e.g., with both deterministic designs and random designs (Lecué and Mendelson, 2014; Oliveira, 2016; Raskutti et al., 2010; Rudelson and Zhou, 2012). Now we present our main results in the next theorem.

Theorem 4.3. Suppose that Assumptions 2.1, 2.2, and 4.2 hold, and $R(\nabla L(\theta^*)) \leq \lambda$ for some $\lambda > 0$. We define

$$\psi = \max_{\delta_{N^\perp} \neq 0} \frac{R(\delta_{N^\perp})}{||\delta_{N^\perp}||_2^2}.$$  \quad (4.1)

Let $\tilde{\theta}$ be any minimizer to (3.1), and $\theta^* \in S$ be the unknown true parameter of the model. Given

$$\tau = \frac{3\kappa}{(2 + \gamma)(5\gamma + 3)\psi^2},$$

we have $\tilde{\theta} - \theta^* \in M_\gamma$ and

$$||\tilde{\theta} - \theta^*||_2 \leq \frac{(2 + \gamma)(5\gamma + 3)\psi \cdot \max[\lambda, \mu]}{2\gamma\kappa}, \quad (4.2)$$

$$R(\tilde{\theta} - \theta^*) \leq \frac{(2 + \gamma)^2(5\gamma + 3)\psi^2 \cdot \max[\lambda, \mu]}{2\gamma\kappa}. \quad (4.3)$$

all first-order algorithms.
We provide the proof of Theorem 4.3 in Section 7.2. Though the obtained bound in Theorem 4.3 is similar to that of Corollary 1 in Negahban et al. (2012), as mentioned in Remarks 2.4 and 2.9, the theoretical analysis in Negahban et al. (2012) is not applicable when $H = \nabla^2 L(\theta)$ is not positive semidefinite. This is because the computational formulation in Negahban et al. (2012), i.e., (1.1), becomes nonconvex. In contrast, our proposed computational formulation, i.e., (1.4), is convex regardless the Hessian matrix $H$ is definite or not. Therefore, given an indefinite $H$, our analysis is still applicable.

Theorem 4.3 gives a deterministic bound for a fixed $\lambda$. Since $R^*(\nabla L(\theta^*))$ is a random variable, we need to verify that $R^*(\nabla L(\theta^*)) \leq \lambda$ holds with high probability. Now we derive $\lambda$ and the concrete theoretical rate of convergence of each individual RDE estimator in our examples.

4.1 Sparse Linear Regression

Suppose that $\theta^*$ contains at most $s$ nonzero entries. We choose

$$N = \{ \theta \in \mathbb{R}^d : \theta_j = 0 \text{ for } \forall j \text{ such that } \theta_j^* \neq 0 \},$$

$$S = \{ \theta \in \mathbb{R}^d : \theta_j = 0 \text{ for } \forall j \text{ such that } \theta_j^* = 0 \}.$$ 

It is easy to verify that $\|\cdot\|_1$ satisfies the decomposability with respect to $(S, N)$. The next corollary establishes the concrete rate of convergence for sparse linear regression.

**Corollary 4.4.** Suppose that the design matrix $X$ satisfies

$$\frac{\|X\delta\|_2^2}{n} \geq \kappa \|\delta\|_2^2$$

for any $\delta$ such that $\|\delta_N\|_1 \leq (1 + \gamma)\|\delta_{N^\perp}\|_1$,

and $\|X_j\|_2 = \sqrt{n}$ for all $j = 1, \ldots, d$. If we choose

$$\tau = \frac{\kappa}{(2 + \gamma)(2\gamma + 1)\sqrt{s}} \quad \text{and} \quad \mu \leq 2\sigma \sqrt{\frac{\log d}{n}},$$

then there exists some universal constant $C$ such that we have

$$\|\tilde{\theta} - \theta^*\|_2 \leq C \cdot \sigma \sqrt{\frac{s \log d}{n}}$$

with probability at least $1 - 2 \exp(-2\log d)$.

The proof of Corollary 4.4 is provided in Appendix B. Corollary 4.4 guarantees that the RDE estimator for sparse linear regression attains the same minimax optimal statistical rate of convergence as the regularized M-estimator (Lasso) (Negahban et al., 2012; Candes and Tao, 2007; Raskutti et al., 2011).
Remark 4.5 (Sparse Linear Regression with Missing Data). Similar result can be obtained when the design matrix contains missing entries. We only need to verify the restricted eigenvalue condition and \( R^* (\nabla L(\theta^*)) \leq \lambda \). Then using Theorem 2.1 in Wang et al. (2017), we obtain

\[
\| \tilde{\theta} - \theta^* \|_2 = O_P \left( \sigma \sqrt{\frac{s \log d}{(1 - \rho)n}} \right). \tag{4.4}
\]

Note that when \( \rho = 0 \), i.e., we observe all entries of the design matrix, (4.4) gives back the result in Corollary 4.4.

Remark 4.6. We provide some further discussion on the reason of smoothing \( R^*(\cdot) \) instead of \( R(\cdot) \), as smoothing \( R(\cdot) \) may sacrifice the computational efficiency. Taking sparse linear regression as an example, where we smooth the \( \ell_1 \) regularizer. To ensure that the minimizer to the smoothed optimization problem attains the optimal statistical rate of convergence, i.e., \( \| \tilde{\theta} - \theta^* \|_2 = O_P \left( \sigma \sqrt{\frac{s \log d}{n}} \right) \), we require the smoothing parameter \( \mu \) to satisfy

\[
\mu = O \left( \sigma \frac{\sqrt{\log d}}{d} \right). \tag{4.5}
\]

For high dimensional sparse linear regression, \( d \) is usually very large. Thus (4.5) requires \( \mu \) to be a very small value, which sacrifices the computational efficiency. See more details in Appendix C.

4.2 Sparse Multivariate Regression

Suppose that \( \Theta^* \) contains at most \( s \) rows with nonzero entries. We choose

\[
N = \{ \Theta \in \mathbb{R}^{d \times m} : \Theta_{js} = 0 \text{ for } \forall j \text{ such that } \Theta^*_{js} \neq 0 \},
\]

\[
S = \{ \Theta \in \mathbb{R}^{d \times m} : \Theta_{js} = 0 \text{ for } \forall j \text{ such that } \Theta^*_{js} = 0 \}.
\]

It is easy to verify that \( ||\cdot||_{1,2} \) satisfies the decomposability with respect to \((S, N)\). The next corollary establishes the concrete rate of convergence for sparse multivariate regression.

Corollary 4.7. Suppose that the design matrix \( X \) satisfies

\[
\frac{\| X \Delta \|_F^2}{n} \geq \kappa \| \Delta \|_F^2 \quad \text{for any } \Delta \text{ such that } \| \Delta_N \|_{1,2} \leq (1 + \gamma) \| \Delta_{N^c} \|_{1,2},
\]

and \( \| X_{ij} \|_2 = \sqrt{n} \) for all \( j = 1, \ldots, d \). If we choose

\[
\tau = \frac{\kappa}{(2 + \gamma)(2\gamma + 1)\sqrt{s}} \quad \text{and} \quad \mu \leq 2\sigma \left( \frac{\sqrt{m}}{n} + \sqrt{\frac{\log d}{n}} \right),
\]

then there exists some universal constant \( C \) such that we have

\[
\frac{1}{\sqrt{m}} \| \tilde{\Theta} - \Theta^* \|_F \leq C \sigma \left( \sqrt{\frac{s}{n}} + \sqrt{\frac{s \log d}{nm}} \right)
\]

with probability at least \( 1 - 2 \exp(-21 \log d) \).
The proof of Corollary 4.4 is provided in Appendix D. Corollary 4.7 guarantees that the RDE estimator for sparse multivariate regression attains the same minimax optimal statistical rate of convergence as the Dantzig-type estimator and regularized M-estimator (Liu et al., 2010; Lounici et al., 2011).

4.3 Reduced Rank Regression

Suppose that \( \Theta^* \) contains at most \( r \) nonzero singular values with a singular value decomposition \( \Theta^* = \sum_{j=1}^{r} \sigma_j(\Theta^*) u_j v_j^\top \), where \( \sigma_j(\Theta^*) \) is the \( j \)th largest singular value, and \( u_j \)'s and \( v_j \)'s are the corresponding left and right singular vectors. We define

\[
U = \text{span}(u_1, \ldots, u_r) \subset \mathbb{R}^d \quad \text{and} \quad V = \text{span}(v_1, \ldots, v_r) \subset \mathbb{R}^m.
\]

We then choose

\[
\mathcal{N} = \{ \Theta \in \mathbb{R}^{d \times m} \mid \Theta_{:,k} \in U_\perp, \Theta_{:,j} \in V_\perp \text{ for all } j,k \}, \tag{4.6}
\]

\[
\mathcal{S} = \{ \Theta \in \mathbb{R}^{d \times m} \mid \Theta_{:,k} \in U, \Theta_{:,j} \in V \text{ for all } j,k \}. \tag{4.7}
\]

It is easy to verify that \( \| \cdot \|_* \) satisfies the decomposability with respect to \((\mathcal{S}, \mathcal{N})\). The next corollary establishes the concrete rate of convergence for reduced rank regression.

**Corollary 4.8.** Suppose that the design matrix \( X \) satisfies

\[
\frac{\|X\Delta\|^2}{n} \geq \kappa \|\Delta\|_F^2 \quad \text{for any } \Delta \text{ such that } \|\Delta_N\|_* \leq (1 + \gamma) \|\Delta_{N_\perp}\|_*,
\]

and \( \|X_{:,j}\|_2 = \sqrt{n} \) for all \( j = 1, \ldots, d \). If we choose

\[
\tau = \frac{\kappa}{(2 + \gamma)(2\gamma + 1)\sqrt{2}r} \quad \text{and} \quad \mu \leq \frac{2\sigma \|X\|_2}{\sqrt{n}} \left( \sqrt{\frac{d}{n}} + \sqrt{\frac{m}{n}} \right)
\]

there exit some universal constants \( C \) and \( c \) such that we have

\[
\frac{1}{\sqrt{m}} \|\Theta - \Theta^*\|_F \leq \frac{C\sigma \|X\|_2}{\sqrt{n}} \left( \sqrt{\frac{r}{n}} + \sqrt{\frac{rd}{nm}} \right)
\]

with probability at least \( 1 - 2 \exp(-c(d + m)) \).

The proof of Corollary 4.8 is provided in Appendix E. Corollary 4.8 guarantees that the RDE estimator for reduced rank regression attains the same minimax optimal statistical rate of convergence as the Dantzig-type estimator and regularized M-estimator (Rohde et al., 2011).
4.4 Sparse Precision Matrix Estimation

Suppose that every $\Theta^*_j$ contains at most $s$ nonzero entries. When estimating $\Theta^*_j$, we choose

$$
N_j = \{\Theta_j \in \mathbb{R}^d : \Theta_{kj} = 0 \text{ for } \forall \ k \text{ such that } \Theta^*_{kj} \neq 0\},
$$

$$
S_j = \{\Theta_j \in \mathbb{R}^d : \Theta_{kj} = 0 \text{ for } \forall \ k \text{ such that } \Theta^*_{kj} = 0\}.
$$

It is easy to verify that $\|\cdot\|_1$ satisfies the decomposability with respect to $(S_j, N_j)$. The next corollary establishes the rate of convergence for sparse precision matrix estimation.

**Corollary 4.9.** Suppose that $\Sigma^*$ and $\Theta^*$ satisfy

$$
\max_{1 \leq j \leq d} \Sigma^*_{jj} \leq K, \quad \Lambda_{\min}(\Sigma^*) \geq 8\kappa, \quad \text{and} \quad \|\Theta^*\|_1 \leq M,
$$

where $K$ and $\kappa$ are constants, and do not scale with $(n, d, s, M)$. There exist universal constants $C_0$, $C_1$, $C_2$, $c_0$, $c_1$, and $c_2$ such that if we choose

$$
\tau = \frac{\kappa}{(2 + \gamma)(2\gamma + 1)\sqrt{s}}, \quad \text{and} \quad \mu \leq C_0M\frac{\log d}{n},
$$

then for large enough $n$, we have

$$
\frac{1}{\sqrt{d}}\|\Theta - \Theta^*\|_F \leq C_1M\frac{s\log d}{n} \quad \text{and} \quad \|\Theta - \Theta^*\|_1 \leq C_2Ms\frac{\log d}{n}
$$

with probability at least $1 - 4\exp(-c_0\log d) - c_2\exp(-c_1n)$.

The proof of Corollary 4.9 is provided in Appendix F. Corollary 4.9 guarantees that the RDE estimator for sparse precision matrix estimation attains the same minimax optimal statistical rate of convergence as the Dantzig-type estimator and regularized M-estimator (Cai et al., 2011; Liu and Luo, 2012; Yuan, 2010).

**Remark 4.10** (Sparse Latent Precision Matrix Estimation). Similar results can be obtained for the nonparanormal distributions. We only need to verify the restricted eigenvalue condition and $R^*(\nabla L(\Theta^*)) \leq \lambda$ as Theorem 5.1 in Zhao and Liu (2014). Then we can obtain

$$
\frac{1}{\sqrt{d}}\|\Theta - \Theta^*\|_F = O_P\left(M\frac{s\log d}{n}\right) \quad \text{and} \quad \|\Theta - \Theta^*\|_1 = O_P\left(Ms\frac{\log d}{n}\right).
$$

Corollaries 4.4-4.9 imply that the RDE estimators attain similar statistical rates of convergence as their counterparts in the family of Dantzig-type estimators, even though our selected smoothing parameters are much larger than the desired optimization accuracy $\epsilon$. We give two concrete examples as follows.
Example 4.11 (Sparse Linear Regression). For a typical high dimensional setting with \( n = 100, \ d = 1000, \) and \( \sigma = 1, \) we choose
\[
\mu = 0.25 \leq \sigma \sqrt{\frac{\log d}{n}} \approx 0.2628.
\]

Example 4.12 (Sparse Multivariate Regression). For a typical high dimensional setting with \( n = 200, \ d = 800, \ m = 13, \) and \( \sigma = 1, \) we choose
\[
\mu = 0.25 \leq \sigma \left( \sqrt{\frac{\log d}{n}} + \sqrt{\frac{m}{n}} \right) \approx 0.4377.
\]

Remark 4.13. The choices of \( \mu \) in the examples above are much larger than the optimization accuracy in general, e.g., \( \epsilon = 10^{-6}. \) Therefore \( 1/\sqrt{\mu} \) becomes negligible compared with \( 1/\sqrt{\epsilon} = 10^3. \) This reduces the number of iterations in (3.13) to approximately \( O(1/\sqrt{\epsilon}). \) On the other hand, these choices of \( \mu \) attain the minimax optimal statistical rates, with mere potential sacrifice in terms of affordable constant factors. Therefore the smoothing technique gains computational efficiency with almost no loss of statistical accuracy (same statistical rate of convergence with the optimal rate in terms of \( s, n, \) and \( d \)).

5 Numerical Experiments

We compare the finite-sample performance of the RDE estimators with the Dantzig-type estimators. For the accelerated proximal gradient algorithm for the RDE estimators, we choose the shrinkage parameter \( h = 0.8 \) for the backtracking line search. For the Dantzig-type estimators, instead of solving the constrained form (1.2), we adopt the ADMM algorithm to solve the regularized form (1.3), which is an efficient approach in practice (See detailed derivations in Li et al. (2015); Wang and Yuan (2012); Boyd et al. (2011); He and Yuan (2012)). We run all the simulations in \( \mathbb{R} \) on a PC with Intel Core i5 3.3GHz CPU and 16GB memory. For notational simplicity, we denote all estimators by \( \hat{\theta} \) or \( \hat{\Theta}. \)

5.1 Simulated Datasets

We first compare the computational efficiency of the accelerated proximal gradient algorithm using different smoothing parameters. We consider the sparse linear regression problem in Example 2.3. The simulated data are generated as follows: We set \( n = 100, \ d = 1000. \) Each row of the design matrix is independently sampled from a 1000-dimensional Gaussian distribution \( \mathcal{N}(0, \Sigma), \) where \( \Sigma_{jj} = 1 \) and \( \Sigma_{j\ell} = 0.5^{|j-\ell|} \) for all \( \ell \neq j. \) We set the regression coefficient vector as \( \theta_1^* = 3, \theta_2^* = 2, \theta_4^* = 1.5, \) and \( \theta_j^* = 0 \) for all \( j \neq 1, 2, 4. \) Each entry of the noise vector \( w \) is independently sampled from \( \mathcal{N}(0, 1). \) We set \( \tau = 0.1, \) and vary \( \mu \) from 10 to 0.0025.

Figure 2 presents the approximation error curves averaged over 50 replications for different values of \( \mu. \) We observe that for a wide range of the values of \( \mu, \) e.g. \( \mu = 2.5 \) to 0.005, the approximation errors are very close, where larger values of \( \mu \) gain better computational efficiency than
the smaller values of $\mu$. On the other hand, when $\mu$ is too large (such as $\mu = 10$), the approximation error increases significantly, and when $\mu$ is too small (such as $\mu = 0.0025$), the convergence rate drops significantly. These are consistent with our theoretical analysis in §3 and §4 that $\mu$ plays the key role in the tradeoff. For the rest of the simulations, we fix the smoothing parameter to be $\mu = 0.1$ for a good balance between statistical and computational performances based on our empirical observation.

![Figure 2: The estimation error curves averaged over 50 replications for different values of $\mu$.](image)

We then compare the computational efficiency and statistical accuracy of the RDE estimators with the Dantzig-type estimators. We consider the sparse multivariate regression, reduced rank regression, and sparse precision matrix estimation problems in Examples 2.5-2.8. For the smoothing technique, we choose $\mu = 0.1$ and the stopping precision $\xi = 10^{-5}$. The simulated data are generated as follows:

(i) Sparse Multivariate Regression. We set $n = 200$, $d = 800$, and $m = 13$. Each row of the design matrix is independently sampled from a 800-dimensional Gaussian distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma_{jj} = 1$ and $\Sigma_{j\ell} = 0.5$ for all $\ell \neq j$. We set the regression coefficient matrix as $\Theta_{1k}^* = 3$, $\Theta_{2k}^* = 2$, $\Theta_{4k}^* = 1.5$, and $\Theta_{jk}^* = 0$ for all $j \neq 1, 2, 4$ and $k = 1, ..., 13$. Every entry of the noise matrix is independently sampled from $\mathcal{N}(0, 1)$. We also generate a validation set of 200 samples for regularization selection, and a testing set of 10000 samples for evaluating statistical accuracy.

(ii) Reduced Rank Regression. We set $n = 400$, $d = 200$, and $m = 100$. Each row of the design matrix is independently sampled from a 200-dimensional Gaussian distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma_{jj} = 1$ and $\Sigma_{j\ell} = 0.5$ for all $\ell \neq j$. We then generate the regression coefficient matrix
\(\Theta^* = LR^\top\), where \(L \in \mathbb{R}^{200 \times 3}\), \(R \in \mathbb{R}^{3 \times 100}\), and all entries of \(L\) and \(R\) are independently generated from \(\mathcal{N}(0, 0.05)\). Each entry of the noise matrix is independently sampled from \(\mathcal{N}(0, 1)\). We also generate a validation set of 400 samples for regularization selection, and a testing set of 10000 samples for evaluating statistical accuracy.

(iii) Sparse Precision Matrix Estimation. We set \(n = 100\) and \(d = 200\). We generate 100 observations independently from a 200-dimensional Gaussian distribution \(\mathcal{N}(0, (\Theta^*)^{-1})\). Here \(\Theta^*\) is generated by the adjacency matrix of a chain graph or an Erdős–Rényi graph. More specifically, each node of the chain graph is assigned an index \(j\) for \(j = 1, \ldots, 200\), and two nodes are connected by an edge if their indices are adjacent. In an Erdős–Rényi random graph, we assign edges between each pair of nodes independently with probability 0.01. Once the graph is obtained, we generate the sparse precision matrix \(\Theta^* \in \mathbb{R}^{200 \times 200}\) by

\[
\Theta^* = \left( C \left( H + (|\Lambda_{\min}(H)| + 0.2) \cdot I_{200} \right) \right)^{-1},
\]

where \(H\) is the adjacency matrix of the graph, and \(C\) is the operator for converting a symmetric positive definite matrix to a correlation matrix. We also generate a validation set of 100 samples for regularization selection.

We choose a geometrically decreasing sequence of 30 regularization parameters such that the RDE estimators and Danzig selectors obtain similar solution paths. The wall clock time for calculating solution paths is recorded for running time comparison. For sparse multivariate regression and reduced rank regression, we choose the regularization parameter from all 30 regularization parameters by minimizing the prediction error over the validation set via

\[
\hat{\lambda} = \arg\min_{\lambda \in \{\lambda_1, \ldots, \lambda_{30}\}} \| \tilde{Y} - \tilde{X} \hat{\Theta}^\lambda \|_F^2,
\]

where \(\hat{\Theta}^\lambda\) denotes the obtained RDE estimator using the regularization parameter \(\lambda\), and \(\tilde{Y}\) and \(\tilde{X}\) denote the response and design matrices of the validation set. For sparse precision matrix estimation, we choose the regularization parameter from all 30 regularization parameters by minimizing the predictive negative log-likelihood over the validation set via

\[
\hat{\lambda} = \arg\min_{\lambda \in \{\lambda_1, \ldots, \lambda_{30}\}} \log |\hat{\Theta}^\lambda| - \text{trace}(\tilde{S}^\top \hat{\Theta}^\lambda),
\]

where \(\hat{\Theta}^\lambda\) denotes the obtained RDE estimator using the regularization parameter \(\lambda\), and \(\tilde{S}\) is the empirical covariance matrix of the validation set.

For sparse multivariate regression and reduced rank regression, we evaluate the statistical accuracy of the obtained estimators using

\[
\text{Est. Err.} = \frac{1}{\sqrt{m}} \| \hat{\Theta} - \Theta^\star \|_F \quad \text{and} \quad \text{Pre. Err.} = \frac{1}{\sqrt{10000}} \| \tilde{Y} - \tilde{X} \hat{\Theta}^\lambda \|_F, \tag{5.1}
\]

where \(\tilde{Y}\) and \(\tilde{X}\) denote the response and design matrices of the testing set. For sparse precision matrix estimation, we evaluate the statistical accuracy of the obtained estimators using

\[
\text{Fro. Est. Err.} = \| \hat{\Theta} - \Theta^\star \|_F \quad \text{and} \quad \text{Spec. Est. Err.} = \| \hat{\Theta} - \Theta^\star \|_2, \tag{5.2}
\]
Tables 1 presents the experimental results averaged over 100 simulations. The RDE estimators and Dantzig-type estimators attain similar statistical performance. But the accelerated proximal gradient (APG) algorithm for RDE estimators is up to 20, 5, and 7 times faster than the ADMM algorithm for Dantzig-type estimators in sparse multivariate regression, reduced rank regression, and sparse precision matrix estimation respectively. The APG algorithm also tends to yield a sparser solution than the ADMM algorithm, though their obtained estimators have similar $\mathcal{R}(\hat{\Theta})$.

Table 1: Quantitative comparison of the computational and statistical performance between the RDE estimators and Dantzig-type estimators. The results are averaged over 100 repetitions with the average running time in seconds and the standard errors in parentheses. We see that both methods attain similar statistical performance, but the APG algorithm for obtaining RDE estimators is more computationally efficient than the ADMM algorithm for obtaining Dantzig-type estimators.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse Multivariate</td>
<td>RDE</td>
<td>27.276(4.9371)</td>
<td>32.659(6.2468)</td>
<td>0.4064(0.1755)</td>
<td>0.3940(0.0812)</td>
</tr>
<tr>
<td>Regression</td>
<td>Dantzig-type</td>
<td>535.03(235.31)</td>
<td>34.506(1.3089)</td>
<td>0.4045(0.1015)</td>
<td>0.3954(0.0803)</td>
</tr>
<tr>
<td>Reduced Rank</td>
<td>RDE</td>
<td>112.72(112.93)</td>
<td>0.8470(0.0017)</td>
<td>0.0305(0.0011)</td>
<td>0.1375(0.0124)</td>
</tr>
<tr>
<td>Regression</td>
<td>Dantzig-type</td>
<td>632.50(116.64)</td>
<td>0.8650(0.0726)</td>
<td>0.0301(0.00015)</td>
<td>0.1366(0.0072)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse Precision Matrix</td>
<td>RDE</td>
<td>209.41(11.729)</td>
<td>308.55(3.6196)</td>
<td>9.0182(0.1571)</td>
<td>1.3558(0.0325)</td>
</tr>
<tr>
<td>Estimation (Chain)</td>
<td>Dantzig-type</td>
<td>1040.4(104.65)</td>
<td>317.39(10.196)</td>
<td>9.1209(0.2353)</td>
<td>1.4308(0.0388)</td>
</tr>
<tr>
<td>Sparse Precision Matrix</td>
<td>RDE</td>
<td>394.54(14.675)</td>
<td>272.35(3.5676)</td>
<td>13.041(0.3758)</td>
<td>3.8304(0.1305)</td>
</tr>
<tr>
<td>Estimation (Random)</td>
<td>Dantzig-type</td>
<td>2982.3(477.07)</td>
<td>276.51(5.2894)</td>
<td>13.045(0.3751)</td>
<td>3.8517(0.1552)</td>
</tr>
</tbody>
</table>

5.2 Real Data Examples

We compare the RDE estimators and Dantzig-type estimators using two real datasets:

(i) The school dataset (Argyriou et al., 2008) consists of 15362 students with 27 features and examination records from 139 secondary schools in years of 1985, 1986, and 1987. We analyze the school data by sparse multivariate regression. We randomly split the data into a training set containing approximately 50% of all the samples, a validation set containing approximately 25%, and a testing set containing the remaining samples from each school (task). We use the training set to estimate sparse multivariate linear models, the validation set to select the regularization parameter by minimizing the prediction accuracy, and the testing set to evaluate the prediction accuracy. Similar to the experiments on the simulated
data, the regularization parameters are tuned over a refined sequence of 30 values, and the wall clock time for calculating solution paths is recorded for running time comparison.

(ii) The stock dataset (Zhao et al., 2012) contains closing prices from all the stocks in S&P 500 on each market open day between January 1, 2003 and January 1, 2005. By transforming the data to log-ratios of the time $t$’s price over time $(t−1)$’s price, we obtain 503 samples for 452 stocks that remained in the S&P 500 during the entire time period. We analyze the stock data by sparse precision matrix estimation. The input covariance matrix is the transformed Spearman’s rho matrix, and the resulting estimators correspond to nonparanormal graphs (see more details in Liu et al. (2012)). We tune the regularization parameters over a refined grid such that the obtained sparse precision matrices contain approximately 1000 edges (The sparsity level is about 1%).

Table 2 summarizes the experimental results for the two datasets. The results of the school data are averaged over 50 replications, and the results of the stock data are obtained using all the samples. We see that the APG algorithm for RDE estimators is more efficient than the ADMM algorithm for Dantzig-type estimators. For the school data, the RDE estimators and Dantzig-type estimators attain similar prediction performance. For the stock data, we compare the obtained graphs as follows.

We present the nonparanormal graphs obtained by both estimators for the stock data in Figure 3. For comparison purpose, we also present the graph obtained by the RDE estimator using the empirical covariance matrix as the input covariance matrix. The nodes in graphs are colored according to the GICS sector of the corresponding stock. We highlight a region in the nonparanormal graph obtained by the RDE estimator. By color coding, we see that the nodes in this region belong to the same sector of the market. A similar pattern can also be found in the nonparanormal graph obtained by the Dantzig-type estimator. In contrast, this region is shown to be sparse in the Gaussian graph obtained by the Dantzig-type estimator. Therefore we can see that the RDE estimator is capable of generating similar refined structures to the Dantzig-type estimator when estimating the nonparanormal graph, but is more computationally efficient.

Table 2: Quantitive comparison of the computational and statistical performance between RDE estimators and Dantzig-type estimators on the school and stock datasets. The results on the school data are averaged over 50 replications, with the standard errors in parentheses. The results on the stock data are based on all samples.
We consider the following decomposition

\[ R^*(\nabla L(\theta^{(t)})) + \tau R(\theta^{(t)}) - R^*(\nabla L(\hat{\theta})) - \tau R(\hat{\theta}) = R^*(\nabla L(\theta^{(t)})) + \tau^2 R(\theta^{(t)}) - R^*(\nabla L(\hat{\theta})) - \tau R(\hat{\theta}). \]  

By (3.4), we have

\[ R^*(\nabla L(\theta^{(t)})) \leq \mu + R^*_\mu(\nabla L(\theta^{(t)})) \quad \text{and} \quad R^*(\nabla L(\hat{\theta})) \geq R^*_\mu(\nabla L(\hat{\theta})). \]  

Figure 3: Stock graphs with all singletons removed. The nonparanormal graphs reveal more refined structures than the Gaussian graph.

6 Discussions

Our proposed methodology and theory can be extended to other statistical learning methods, such as LAD Lasso, square-root Lasso, tuning-insensitive graph estimation, and calibrated multivariate regression (Wang, 2013; Belloni et al., 2011; Liu and Wang, 2012; Liu et al., 2014). These methods adopt the nonsmooth loss functions to gain estimation robustness, modeling flexibility, and tuning insensitiveness. We can replace their loss functions with the corresponding smooth approximations. By similar theoretical analysis in this paper, we can show that these resulting estimators also possess good statistical properties.

7 Proof of the Main Theorems

We provide the proofs of our main computational and statistical results in this Section.

7.1 Proof of Theorem 3.10

We consider the following decomposition

\[ R^*(\nabla L(\theta^{(t)})) + \tau R(\theta^{(t)}) - R^*(\nabla L(\hat{\theta})) - \tau R(\hat{\theta}) = R^*(\nabla L(\theta^{(t)})) + \tau R(\theta^{(t)}) \]

\[ - R^*_\mu(\nabla L(\hat{\theta})) - \tau R(\hat{\theta}) + R^*_\mu(\nabla L(\hat{\theta})) + \tau R(\hat{\theta}) - R^*(\nabla L(\hat{\theta})) - \tau R(\hat{\theta}). \]  

By (3.4), we have

\[ R^*(\nabla L(\theta^{(t)})) \leq \mu + R^*_\mu(\nabla L(\theta^{(t)})) \quad \text{and} \quad R^*(\nabla L(\hat{\theta})) \geq R^*_\mu(\nabla L(\hat{\theta})). \]  

22
Combining (7.1) and (7.2), we have
\[ R^*(\nabla L(\theta^{(t)})) + \tau R(\theta^{(t)}) - R^*(\nabla L(\tilde{\theta})) - \tau R(\tilde{\theta}) \leq \mu + R^*_\mu(\nabla L(\theta^{(t)})) + \tau R(\theta^{(t)}) - R^*_\mu(\nabla L(\tilde{\theta})) - \tau R(\tilde{\theta}). \] (7.3)

Since \( \tilde{\theta} \) is the minimizer of (3.1), we have
\[ R^*_\mu(\nabla L(\tilde{\theta})) + \tau R(\tilde{\theta}) \leq R^*_\mu(\nabla L(\tilde{\theta})) + \tau R(\tilde{\theta}). \] (7.4)

By Theorem 5.1 in Beck and Teboulle (2009a), we have the following convergence rate of the fast proximal gradient algorithm for minimizing (3.1),
\[ R^*_\mu(\nabla L(\theta^{(t)})) + \tau R(\theta^{(t)}) - R^*(\nabla L(\tilde{\theta})) - \tau R(\tilde{\theta}) \leq \frac{2\phi\|\theta^{(0)} - \tilde{\theta}\|^2_2}{\mu(t + 1)^2}, \] (7.5)

where \( \phi = \|\nabla^2 L(\theta)\|^2_2 = \|H\|^2_2 \). If we set the of (7.5) to be no larger than \( \frac{\epsilon}{2} \), we have the desired results (3.13) and (3.14). By combining (7.3), (7.4) and (7.5), we have
\[ R^*(\nabla L(\theta^{(t)})) + \tau R(\theta^{(t)}) - R^*(\nabla L(\tilde{\theta})) - \tau R(\tilde{\theta}) \leq \mu + \frac{2\phi\|\theta^{(0)} - \tilde{\theta}\|^2_2}{\mu(t + 1)^2}. \] (7.6)

By the choice of \( t \) as (3.13), we have the desired result (3.15). If we further set \( \mu = \epsilon/2 \), to make the left-hand side of (7.6) no larger than \( \epsilon \), we need
\[ \frac{4\phi\|\theta^{(0)} - \tilde{\theta}\|^2_2}{\epsilon(t + 1)^2} \leq \epsilon. \]

By solving the inequality above, we obtain
\[ t \geq \frac{2\sqrt{\phi}\|\theta^{(0)} - \tilde{\theta}\|_2}{\epsilon} - 1, \]
which completes the proof.

### 7.2 Proof of Theorem 4.3

For simplicity, we define \( \tilde{\delta} = \tilde{\theta} - \theta^* \). By triangle inequality, we have
\[ R(\tilde{\theta}) = R(\theta^* + \tilde{\delta}) = R(\theta^*_S + \theta^*_N + \tilde{\delta}_S + \tilde{\delta}_N) \]
\[ \geq R(\theta^*_S + \tilde{\delta}_N) - R(\theta^*_S + \tilde{\delta}_N) \geq R(\theta^*_S + \tilde{\delta}_N) - R(\theta^*_S) - R(\tilde{\delta}_N). \] (7.7)

By the decomposability of \( R(\cdot) \), (7.7) implies
\[ R(\tilde{\theta}) \geq R(\theta^*_S) + R(\tilde{\delta}_N) - R(\theta^*_S) = R(\tilde{\delta}_N), \] (7.8)

where the last equality comes from \( R(\tilde{\delta}_S) = 0 \). From \( R(\theta^*) = R(\theta^*_S), (7.8) \) further implies
\[ R(\tilde{\theta}) - R(\theta^*) \geq R(\tilde{\delta}_N) - R(\theta^*_S) = R(\tilde{\delta}_N) - R(\theta^*_S). \] (7.9)
Since $\tilde{\theta}$ is a minimum to (3.1), we have

$$R^*\left(\nabla L(\tilde{\theta})\right) + \tau R(\tilde{\theta}) - \frac{\mu}{2} \leq R^*_\mu(\nabla L(\tilde{\theta})) + \tau R(\tilde{\theta}) \leq \frac{1}{\tau} R^*(\nabla L(\theta^*)) + \frac{\mu}{\tau} \leq \frac{1}{\tau} R^*(\nabla L(\theta^*)) + \frac{\mu}{2\tau} \leq \frac{3\max\{\lambda, \mu\}}{2\tau},$$

(7.10)

where (i) and (ii) come from (3.4). By simple manipulation, (7.10) implies

$$R(\tilde{\theta}) - R(\theta^*) \leq \frac{1}{\tau} R^*(\nabla L(\theta^*)) - \frac{1}{\tau} R^*(\nabla L(\tilde{\theta})) + \frac{\mu}{2\tau} \leq \frac{1}{\tau} R^*(\nabla L(\theta^*)) + \frac{\mu}{2\tau} \leq \frac{3\max\{\lambda, \mu\}}{2\tau},$$

(7.11)

where the last inequality comes from the assumption $R^*(\nabla L(\theta^*)) \leq \lambda$. Denote $M_{\lambda, \mu} = \max\{\lambda, \mu\}$ for notational simplicity. Combining (7.9) with (7.11), we obtain

$$R(\tilde{\delta}_{N}) - R(\tilde{\delta}_{N^*}) \leq \frac{3M_{\lambda, \mu}}{2\tau}. \tag{7.12}$$

We consider two cases in the subsequent analysis.

**Case 1**: $3M_{\lambda, \mu} \leq 2\tau \gamma R(\tilde{\delta}_{N^*})$. Then (7.12) implies $\tilde{\delta} \in M_{\gamma}$, i.e.,

$$R(\tilde{\delta}_{N}) \leq (1 + \gamma)R(\tilde{\delta}_{N^*}). \tag{7.13}$$

Then we have

$$\frac{\kappa R(\tilde{\delta}_{N^*})}{\psi^2} \leq \frac{\kappa||\tilde{\delta}_{N^*}||^2}{\psi^2} \leq \frac{\kappa||\tilde{\delta}||^2}{\psi^2} \leq \delta^T \nabla^2 L(\theta^*) \delta \leq R(\tilde{\delta}) \cdot R^*(\nabla^2 L(\theta^*) \delta) \leq \frac{(2 + \gamma) R(\tilde{\delta}_{N^*}) \cdot R^*(\nabla^2 L(\theta^*) \delta)}{\kappa}, \tag{7.14}$$

where (i) comes from the definition of $\psi$, (ii) comes from Assumption 4.2, (iii) comes from the Hölder’s inequality, and (iv) comes from (7.13). Since Assumption 2.1 implies $\nabla^2 L(\theta^*) \delta = \nabla L(\tilde{\theta}) - \nabla L(\theta^*)$, by simple manipulation, we rewrite (7.14) as

$$R(\tilde{\delta}_{N^*}) \leq \frac{(2 + \gamma) \psi^2 R^*(\nabla L(\tilde{\theta}) - \nabla L(\theta^*))}{\kappa}. \tag{7.15}$$

Note that (7.10) also implies

$$R^*(\nabla L(\tilde{\theta})) + R^*(\nabla L(\theta^*)) + \tau R(\tilde{\theta}) \leq 2R^*(\nabla L(\theta^*)) + \tau R(\theta^*) + \frac{\mu}{2}.$$ \tag{7.16}

Then by the triangle inequality, we have

$$R^*(\nabla L(\tilde{\theta}) - \nabla L(\theta^*)) \leq R^*(\nabla L(\tilde{\theta})) + R^*(\nabla L(\theta^*)) \leq 2R^*(\nabla L(\theta^*)) + \tau R(\theta^*) + \frac{\mu}{2} \leq \frac{5M_{\lambda, \mu}}{2} - \tau R(\tilde{\delta}_{N^*}) + \tau R(\tilde{\delta}_{N^*}) \leq \frac{5M_{\lambda, \mu}}{2} + \tau R(\tilde{\delta}_{N^*}). \tag{7.17}$$

24
where (i) comes from (7.16), and (ii) comes from (7.9) and the assumption \( \mathcal{R}^* \left( \nabla L(\theta^*) \right) \leq M_{\lambda,\mu} \).

Combining (7.15) with (7.17), we have

\[
\mathcal{R}(\delta_{N_\perp}) \leq \frac{(2 + \gamma)\psi^2 \left( 2.5M_{\lambda,\mu} + \tau \mathcal{R}(\delta_{N_\perp}) \right)}{\kappa}.
\]  

(7.18)

For \((2 + \gamma)\psi^2 \tau \geq \kappa\), we have that (7.18) holds for any \( \mathcal{R}(\delta_{N_\perp}) \). For \((2 + \gamma)\psi^2 \tau < \kappa\), (7.18) implies

\[
\mathcal{R}(\delta_{N_\perp}) \leq \frac{2.5(2 + \gamma)\psi^2 M_{\lambda,\mu}}{\kappa - (2 + \gamma)\tau \psi^2}.
\]  

(7.19)

**Case 2:** \(3M_{\lambda,\mu} \geq 2\tau \gamma \mathcal{R}(\delta_{N_\perp})\). Combining (7.19) with Case 1, we obtain

\[
\mathcal{R}(\delta_{N_\perp}) \leq \max \left\{ \frac{2.5(2 + \gamma)\psi^2 M_{\lambda,\mu}}{\kappa - (2 + \gamma)\tau \psi^2}, \frac{3M_{\lambda,\mu}}{2\tau \gamma} \right\}.
\]  

(7.20)

Minimizing the right-hand side of (7.20) with respect to \(\tau\) by setting two terms equal, we obtain

\[
\tau = \frac{3\kappa}{(5\gamma + 3)(2 + \gamma)\psi^2}
\]

which makes two cases equivalent, i.e., \(2\tau \gamma \mathcal{R}(\delta_{N_\perp}) = 3M_{\lambda,\mu}\). Combining this with (7.12), we have \(\mathcal{R}(\delta_N) \leq (1 + \gamma)\mathcal{R}(\delta_{N_\perp})\), i.e., \(\delta \in \mathcal{M}_y\) holds for both cases. Since \(\gamma > 0\), we have \(\tau\) satisfying

\[
(2 + \gamma)\psi^2 \tau = \frac{3\kappa}{(5\gamma + 3)} < \kappa.
\]

Now we proceed to derive the error bounds. Since we have

\[
\mathcal{R}(\delta_{N_\perp}) = \frac{3M_{\lambda,\mu}}{2\tau \gamma} = \frac{(2 + \gamma)(5\gamma + 3)\psi^2 M_{\lambda,\mu}}{2\gamma \kappa},
\]  

(7.21)

by (7.13) again, we obtain

\[
\mathcal{R}(\delta) \leq (2 + \gamma)\mathcal{R}(\delta_{N_\perp}) = \frac{(2 + \gamma)^2(5\gamma + 3)\psi^2 M_{\lambda,\mu}}{2\gamma \kappa}.
\]

Since \(\nabla^2 L(\theta^*)\delta = \nabla L(\theta) - \nabla L(\theta^*)\), plugging (7.21) back to (7.17), we obtain

\[
\mathcal{R}^* \left( \nabla^2 L(\theta^*)\delta \right) = \mathcal{R}^* \left( \nabla L(\theta) - \nabla L(\theta^*) \right) \leq \frac{5M_{\lambda,\mu}}{2} + \frac{3M_{\lambda,\mu}}{2\gamma} = \frac{(5\gamma + 3)M_{\lambda,\mu}}{2\gamma}.
\]  

(7.22)

Combining (7.14), (7.21) and (7.22), we have

\[
\|\delta\|^2 \leq \frac{(2 + \gamma)\mathcal{R}(\delta_{N_\perp}) \cdot \mathcal{R}^* \left( \nabla^2 L(\theta^*)\delta \right)}{\kappa} \leq \frac{(2 + \gamma)^2(5\gamma + 3)^2\psi^2 M_{\lambda,\mu}^2}{4\gamma^2 \kappa^2}.
\]
A Efficient Projections

The optimization problem

$$\hat{U}^\Theta = \arg\min_{\|U\|_{F} \leq 1} \|U - \nabla L(\Theta) / \mu\|^2_F$$  \hspace{1cm} (A.1)$$
can be solved by a simple variant of the root finding-based algorithm in Liu and Ye (2009). In particular, we denote $V = \nabla L(\Theta) / \mu$. Let $v = (v_1, \ldots, v_d)^\top$ with $v_j = \|V_{j*}\|_2$, we solve

$$\hat{u} = \arg\min_{\|u\|_1 \leq 1} \|u - v\|^2_2.$$  \hspace{1cm} (A.2)$$
Note that (A.2) is an $\ell_1$-ball projection problem, and can be efficiently solved by the algorithm in Liu and Ye (2009). Once $\hat{u}$ is obtained, we can recover $\hat{U}^\Theta$ by

$$\hat{U}^\Theta_{j*} = (\hat{u}_j / v_j) V_{j*}.$$  \hspace{1cm} (A.3)$$

Moreover, the optimization problem

$$\hat{U}^\Theta = \arg\min_{\|U\|_{F} \leq 1} \|U - \nabla L(\Theta) / \mu\|^2_F$$  \hspace{1cm} (A.4)$$
can also be solved in a similar fashion. In particular, we denote $V = \nabla L(\Theta) / \mu$, and we consider the singular value decomposition

$$V = \sum_{k=1}^{r} \sigma_k(V) v_k w_k^\top,$$
where $r$ is the rank of $V$, $\sigma_k(V)$’s are singular values of $V$, and $v_k$’s and $w_k$’s are corresponding singular vectors. Then we solve

$$\{\hat{\rho}_k\}_{k=1}^r = \arg\min_{\sum_{k=1}^r |\rho_k| \leq 1} \sum_{k=1}^r (\rho_k - \sigma_k(V))^2.$$  \hspace{1cm} (A.4)$$
Note that (A.4) is also an $\ell_1$-ball projection problem. Once $\{\hat{\rho}_k\}_{k=1}^r$ are obtained, we can recover $\hat{U}^\Theta$ by

$$\hat{U}^\Theta = \sum_{k=1}^{r} \hat{\rho}_k v_k w_k^\top.$$

B Proof of Corollary 4.4

Proof. Since $\|X_{ij}\|_2 = \sqrt{n}$ for all $j = 1, \ldots, d$ and each entry of $w$ is independently sampled from $\mathcal{N}(0, \sigma^2)$, we have

$$\mathbb{P}\left(\mathcal{R}^* (\nabla L(\theta^*)) = \frac{1}{n} X^\top w \right) \leq 2\sigma \sqrt{\frac{\log d}{n}} \leq 2 \frac{\log d}{d},$$  \hspace{1cm} (B.1)$$
The detailed derivation of (B.1) can be found in Negahban et al. (2012). Since $\theta^*$ has at most $s$ nonzero entries, we have $\psi = \sqrt{s}$. Plugging $\lambda = 2\sigma \sqrt{\log d/n}$ into Theorem 4.3, we obtain

$$\mathbb{P}\left(\|\hat{\theta} - \theta^*\|_2 \leq \frac{4(2 + \gamma)(5\gamma + 3)\sigma}{\gamma^\kappa} \sqrt{\frac{s \log d}{n}} \right) \geq 1 - \frac{2}{d}.$$  \hspace{1cm} (B.2)$$

26
C  Details of Remark 4.6

Proof. We denote \( ||\tilde{\theta}_{1,\mu}|| \) as the smooth approximation of \( ||\tilde{\theta}||_1 \). Then following the construction of smooth approximation, \( ||\tilde{\theta}_{1,\mu}|| \) is obtained by solving the following maximization problem

\[
||\tilde{\theta}_{1,\mu}|| = \max_{||u||_\infty \leq 1} u^T \tilde{\theta} - \frac{\mu}{2} ||u||_2^2.
\]

This implies for any \( u \) satisfying \( ||u||_\infty \leq 1 \), we have

\[
||\tilde{\theta}||_1 - \frac{\mu}{2} ||u||_2^2 \leq \max_{||v||_\infty \leq 1} v^T \tilde{\theta} - \frac{\mu}{2} ||u||_2^2 \leq \max_{||u||_\infty \leq 1} u^T \tilde{\theta} - \frac{\mu}{2} ||u||_2^2 = ||\tilde{\theta}_{1,\mu}||
\]

We further have the bound

\[
||u||_2^2 \leq d ||u||_\infty^2 \leq d,
\]

Combining the results above, we have the following universal bound

\[
||\tilde{\theta}||_1 - \frac{d \mu}{2} \leq ||\tilde{\theta}_{1,\mu}||.
\]

To ensure the same minimax optimal statistical rate of convergence \( ||\tilde{\theta} - \theta^*||_2 = C\sigma \sqrt{\frac{s \log d}{n}} \) for some universal constant \( C \) with high probability, we require

\[
\frac{d \mu}{2} \leq 2C\sigma \sqrt{\frac{\log d}{n}},
\]

which completes the argument.

D  Proof of Corollary 4.7

Proof. Since \( ||X_j||_2 = \sqrt{n} \) for all \( j = 1, \ldots, d \) and each entry of \( W \) is independently sampled from \( N(0, \sigma^2) \), we have

\[
P\left( \mathcal{R}^*(\nabla L(\theta^*)) = \frac{1}{n} W^T X \right) \leq 2\sigma \left( \sqrt{\frac{m}{n}} + \sqrt{\frac{\log d}{n}} \right) \geq 1 - \frac{2}{d^2}. \tag{D.1}
\]

The detailed derivation of (D.1) can be found in Lounici et al. (2011). Since \( \Theta^* \) has at most \( s \) rows with nonzero entries, we have \( \psi = \sqrt{s} \). Plugging \( \lambda = 2\sigma \left( \sqrt{\frac{m}{n}} + \sqrt{\frac{\log d}{n}} \right) \) into Theorem 4.3, we obtain

\[
P\left( ||\tilde{\Theta} - \Theta^*||_F \leq \frac{4(2+\gamma)(5\gamma+3)^2\sigma}{\gamma \kappa} \left( \sqrt{\frac{sm}{n}} + \sqrt{\frac{s \log d}{n}} \right) \right) \geq 1 - \frac{2}{d^2}.
\]

\( \square \)
Proof of Corollary 4.9

Proof. We first prove that Assumption 4.2 holds with high probability. For notational simplicity, we denote $\tilde{\theta} = \tilde{\Theta}_{ij}$, $\theta^* = \Theta^*_{ij}$, and $\tilde{\delta} = \tilde{\Delta}_{ij} = \tilde{\Theta}_{ij} - \Theta^*_{ij}$. We have

$$\frac{\tilde{\delta}^T \nabla^2 \mathcal{L}(\theta^*) \tilde{\delta}}{||\tilde{\delta}||^2} = \frac{\tilde{\delta}^T S \tilde{\delta}}{n ||\tilde{\delta}||^2} = \frac{\tilde{\delta}^T X^T X \tilde{\delta}}{n ||\tilde{\delta}||^2} = \frac{||X \tilde{\delta}||_2^2}{n ||\tilde{\delta}||^2} = \frac{||X \tilde{\delta}||_2^2}{\sqrt{n} ||\tilde{\delta}||_2}.$$

(F.1)

Since each $X_i$ is independently sampled from $\mathcal{N}(0, \Sigma^*)$ with $\Lambda_{\min}(\Sigma^*) \geq 8 \kappa$, there exist universal constants $c_1$ and $c_2$ such that we have

$$P \left( \frac{||X \tilde{\delta}||_2^2}{\sqrt{n}} \geq 2 \sqrt{\kappa} ||\tilde{\delta}||_2 - 9 ||\Sigma^*||_{\text{max}} ||\delta||_1 \sqrt{\frac{\log d}{n}} \right) \geq 1 - c_2 \exp(-c_1 n).$$

(F.2)

The detailed derivation of (F.1) can be found in Raskutti et al. (2011). We consider a large enough $n$ such that

$$9 ||\Sigma^*||_{\text{max}} ||\delta||_1 \sqrt{\frac{\log d}{n}} \leq 9 (2 + \gamma) K ||\tilde{\delta}_{S_j^c}||_1 \sqrt{\frac{\log d}{n}} \leq 9 (2 + \gamma) K ||\tilde{\delta}||_2 \sqrt{\frac{s \log d}{n}} \leq \sqrt{\kappa} ||\tilde{\delta}||_2,$$

where (i) comes from $\tilde{\delta} \in M_\gamma$ and $||\Sigma^*||_{\text{max}} = \max_j \Sigma^*_{jj} \leq K$, and (ii) comes from $||\tilde{\delta}_{S_j^c}||_2 \leq ||\tilde{\delta}||_2$ and the fact that $\tilde{\delta}_{S_j}$ contains at most $s$ nonzero entries. Then (F.1) and (F.2) imply that for all $j = 1, \ldots, d$, we have

$$P \left( \frac{\tilde{\delta}^T \nabla^2 \mathcal{L}(\theta^*) \tilde{\delta} \geq \kappa ||\tilde{\delta}||_2^2 \right) \geq 1 - c_2 \exp(-c_1 n).$$

Now we proceed to derive the error bounds. By the Cauchy-Schwarz inequality, we have

$$\mathcal{R}'(\nabla \mathcal{L}(\Theta^*_{ij})) = ||I_j - S \Theta^*_{ij}||_{\infty} = ||(\Sigma^* - S) \Theta^*_{ij}||_{\infty} \leq ||\Theta^*_{ij}||_1 ||\Sigma^* - S||_\infty.$$

(F.3)
Since each \( X_i \) is independently sampled from \( N(0, \Sigma^*) \), there exist some universal constants \( C_0 \) and \( c_0 \) such that we have
\[
P \left( \| \Sigma^* - S \|_{\max} \leq C_0 \sqrt{\frac{\log d}{n}} \right) \geq 1 - \frac{4}{d^{c_0}}. \tag{F.4}
\]
The detailed derivation can be found in Cai et al. (2011). Since \( \| \Theta^*_j \|_1 \leq \| \Theta^* \|_1 \leq M \), (F.3) and (F.4) imply
\[
P \left( R^* (\nabla L(\Theta^*_j)) \leq C_0 M \sqrt{\frac{\log d}{n}} \right) \geq 1 - \frac{4}{d^{c_0}}.
\]
Since \( \Theta^*_j \) has at most \( s \) nonzero entries, we have \( \psi = \sqrt{s} \). Plugging \( \lambda = C_0 M \sqrt{\log d/n} \) into Theorem 4.3, we have
\[
\| \Theta^*_j - \Theta^*_{s,j} \|_2 \leq \frac{2(2 + \gamma)(5\gamma + 3)C_0 M}{\gamma \kappa} \sqrt{\frac{s \log d}{n}}, \tag{F.5}
\]
\[
\| \Theta^*_j - \Theta^*_{s,j} \|_1 \leq \frac{2(2 + \gamma)^2(5\gamma + 3)C_0 Ms}{\gamma \kappa} \sqrt{\frac{\log d}{n}} \tag{F.6}
\]
with probability at least \( 1 - 4 \exp(-c_0 \log d) - c_2 \exp(-c_1 n) \). Since (F.5) and (F.6) hold for all \( j = 1, \ldots, d \), by simple manipulation, we have
\[
\frac{1}{\sqrt{d}} \| \Theta^* - \Theta^* \|_F = \frac{1}{\sqrt{d}} \sqrt{\sum_{j=1}^{d} \| \Theta^*_j - \Theta^*_{s,j} \|_2^2} \leq \frac{2(2 + \gamma)(5\gamma + 3)C_0 M}{\gamma \kappa} \sqrt{\frac{s \log d}{n}},
\]
\[
\| \Theta^* - \Theta^* \|_1 \leq \max_{1 \leq j \leq d} \| \Theta^*_j - \Theta^*_{s,j} \|_1 \leq \frac{2(2 + \gamma)^2(5\gamma + 3)C_0 Ms}{\gamma \kappa} \sqrt{\frac{\log d}{n}}
\]
with probability at least \( 1 - 4 \exp(-c_0 \log d) - c_2 \exp(-c_1 n) \). \qed

References


