

Slow Kill for Big Data Learning

Yiyuan She, Jiahui Shen, and Adrian Barbu

Abstract—Big-data applications often involve a vast number of observations and features, creating new challenges for variable selection and parameter estimation. This paper presents a novel technique called “slow kill,” which utilizes nonconvex constrained optimization, adaptive ℓ_2 -shrinkage, and increasing learning rates. The fact that the problem size can decrease during the slow kill iterations makes it particularly effective for large-scale variable screening. The interaction between statistics and optimization provides valuable insights into controlling quantiles, stepsize, and shrinkage parameters in order to relax the regularity conditions required to achieve the desired level of statistical accuracy. Experimental results on real and synthetic data show that slow kill outperforms state-of-the-art algorithms in various situations while being computationally efficient for large-scale data.

Index Terms—Top-down algorithms, sparsity, nonconvex optimization, nonasymptotic analysis, sub-Nyquist spectrum sensing

I. INTRODUCTION

This paper studies how to build a parsimonious and predictive model in big data applications, where both the number of predictors and the number of observations can be extremely large. Let $y \in \mathbb{R}^n$ be a response vector with n samples and $X = [x_1, \dots, x_p] \in \mathbb{R}^{n \times p}$ be a design matrix consisting of p features or predictors. Consider a general learning problem with loss $l_0(X\beta; y)$ to measure the discrepancy between $X\beta$ and y . As p can be much larger than n , a sparsity-promoting regularizer is often used to capture model parsimony

$$\min_{\beta \in \mathbb{R}^p} l_0(X\beta; y) + P(\beta; \lambda), \quad (1)$$

where λ is a regularization parameter. There are numerous options for l_0 and P , neither of which are necessarily convex. In many cases, l_0 may be a negative log-likelihood function, but we will consider a more general setup that may not be based on likelihood.

Over the past decade, there have been significant advancements in statistical theory for the minimizers of the penalized problem (1). However, modern scientists often encounter challenges with big data, making it impractical to obtain globally optimal estimators even when convexity is present. This paper aims to incorporate computational considerations into statistical modeling, resulting in a new big-data learning framework with theoretical guarantees. When tackling these challenges

This work is supported in part by the National Science Foundation. The authors thank Dr. Crainiceanu for suggesting the term slow kill. (Corresponding author: Yiyuan She.)

Y. She is with the Department of Statistics, Florida State University, Tallahassee, Florida 32306, U.S.A. e-mail: yshe@stat.fsu.edu.

J. Shen was with the Department of Statistics, Florida State University, Tallahassee, Florida 32306, U.S.A. e-mail: jshen@fsu.edu.

A. Barbu is with the Department of Statistics, Florida State University, Tallahassee, Florida 32306, U.S.A. e-mail: abarbu@stat.fsu.edu.

in large-scale variable selection, the desired algorithms should possess the following traits:

(a) Ease in tuning. It is common in practice to seek a solution with a *prescribed* cardinality (or a specific number of variable, denoted by q). However, using an algorithm designed for the penalized problem (1) may require excessive computation, and the regularization parameter λ may not be as intuitive when attempting to achieve this objective. Many practitioners perform a grid search for λ . However, when dealing with big data, the grid must be fine enough to encompass potentially useful candidate models, resulting in a substantial computational burden.

(b) Scalability. In addition to being efficient, an ideal algorithm should be easy to implement. Since ad-hoc procedures can be unreliable, it is preferable to employ an algorithm based on *optimization* rather than relying on heuristics. It would also be advantageous if the algorithm could adapt its parameters according to the available computational resources, which necessitates an understanding of the algorithm’s iteration complexity and per-iteration cost.

(c) Statistical guarantee. It is widely recognized that the lasso is effective for variable selection when the design matrix exhibits low coherence and the signal is sufficiently strong [1, 2]. Some simpler and faster methods, such as those for variable screening [3], are based on the assumption of independent (or only mildly correlated) features. While these weak-correlation assumptions allow for aggressive feature elimination, they are often restrictive for real-world high-dimensional data. Evaluating a globally optimal solution to (1) with an ℓ_0 -type penalty [4] does have a statistically sound guarantee regardless of coherence, but is only computationally feasible for small datasets. Therefore, a more pressing challenge is to design an iterative process that can relax the stringent regularity conditions required for attaining optimal statistical accuracy.

This work proposes a new approach called *slow kill* to tackle the aforementioned challenges. The main features of the algorithm are as follows.

- Interestingly, slow kill works in the opposite direction of forward pathwise methods and boosting algorithms, which all build up a model from the null [5–9].
- Slow kill incorporates adaptive ℓ_2 -shrinkage and growing learning rates to handle coherent designs and reduce computational burden. Its roots in optimization make it computationally scalable and easy to tune parameters.
- Theoretically, slow kill enjoys rigorous, provable guarantees of accuracy and linear convergence in a statistical sense. In particular, our theory supports backward quantile control and fast learning.

The rest of the paper is organized as follows. Section II investigates a hybrid regularized estimation in the regres-

sion setting to motivate some basic elements of slow kill and compares it to related works. Section III introduces the general slow kill procedure for a differentiable loss function and analyzes how the statistical error changes as the cycles progress. Section IV performs extensive simulations and real data experiments to compare slow kill to some state-of-the-art methods in terms of both efficiency and accuracy. We summarize our findings in Section V. More technical details are provided in the appendix.

Notations and symbols. The following notations and symbols will be used. Let $[n] = \{1, \dots, n\}$ and $\lfloor x \rfloor$ be the largest integer smaller than or equal to x . Define $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. We use $a \lesssim b$ to denote $a \leq cb$ for some positive constant c , and the constants denoted by c or C may not be the same at each occurrence. Given any $\beta \in \mathbb{R}^p$, we use $\mathcal{J}(\beta) \subset [p]$ to denote its support, i.e., $\mathcal{J}(\beta) = \{j : \beta_j \neq 0\}$, and $J(\beta) = |\mathcal{J}(\beta)| = \|\beta\|_0 = \sum_{j=1}^p 1_{\beta_j \neq 0}$. Given $I \subset [p]$, we use X_I to denote the sub-matrix of X formed with the columns in I , and β_I the subvector associated with I . In particular, x_j denotes the j th column of X for any $j \in [p]$. When A is a symmetric matrix, we use A_I to denote the submatrix of A formed with the columns and rows indexed by I , and $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ to denote its largest and smallest eigenvalues, respectively.

Given $X \in \mathbb{R}^{n \times p}$, the restricted isometry numbers $\rho_+(s)$, $\rho_-(s)$ [10] are the smallest and largest numbers, respectively, that satisfy

$$\rho_-(s)\|\beta\|_2^2 \leq \|X\beta\|_2^2 \leq \rho_+(s)\|\beta\|_2^2, \quad \forall \beta \in \mathbb{R}^p : \|\beta\|_0 \leq s, \quad (2)$$

and their dependence on X is omitted. Obviously, $0 \leq \rho_-(s) \leq \rho_+(s) \leq \rho_+(p) = \|X\|_2^2$, where $\|X\|_2$ denotes the spectral norm of X .

For ease of presentation, we introduce a quantile-thresholding operator $\Theta^\#$ which performs simultaneous thresholding and ℓ_2 -shrinkage [11]. Given any $s = [s_1, \dots, s_p]^T \in \mathbb{R}^p$, $\Theta^\#(s; q, \eta) = [t_1, \dots, t_p]^T$ satisfying $t_{(j)} = s_{(j)} / (1 + \eta)$ if $1 \leq j \leq q$, and 0 otherwise, where $s_{(1)}, \dots, s_{(p)}$ are the order statistics of s_1, \dots, s_p satisfying $|s_{(1)}| \geq \dots \geq |s_{(p)}|$, and $t_{(1)}, \dots, t_{(p)}$ are defined similarly. To avoid ambiguity, we make a $\Theta^\#$ -uniqueness assumption in performing $\Theta^\#(s; q, \eta)$ throughout the paper: either $|s_{(q)}| > |s_{(q+1)}|$ or $s_{(q)} = s_{(q+1)} = 0$ occurs. The multivariate quantile thresholding function $\vec{\Theta}^\#(S; q, \eta)$ for any $S = [s_1, \dots, s_p]^T \in \mathbb{R}^{p \times m}$ is defined as a $p \times m$ matrix $T = [t_1, \dots, t_p]^T$ with $t_j = s_j / (1 + \eta)$ if $\|s_j\|_2$ is among the q largest elements in $\{\|s_j\|_2 : 1 \leq j \leq p\}$, and 0 otherwise.

II. WHY BACKWARD SELECTION?

This section is to motivate a “top-down” algorithm design in the fundamental regression setting. The quadratic loss is an important case of strongly convex losses and examining this case will provide a foundation for more general studies under restricted strong convexity.

Assume $y = X\beta^* + \epsilon$, where $\beta^* \in \mathbb{R}^p$, $\|\beta^*\|_0 \leq s$ with $s \leq p \wedge n$. To begin with, we consider an ℓ_0 -constrained,

ℓ_2 -penalized optimization problem to estimate the coefficient vector in high dimensions,

$$\min_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + \frac{\eta_0}{2} \|\beta\|_2^2 \equiv f(\beta) \quad \text{s.t. } \|\beta\|_0 \leq q. \quad (3)$$

When X, y are not centered, an intercept term 1α should be added in the loss, and α is subject to no regularization. The hybrid regularization in (1) differs from the commonly used linear combination of ℓ_1 and ℓ_2 penalties in the elastic net [12]. Compared to the regular ℓ_1 penalty and other nonconvex penalties, $\|\cdot\|_0$ is arguably an ideal choice for enforcing sparsity and does not incur any unwanted bias. The constraint parameter q ($\leq p$) directly controls the number of variables in the resulting model, making it more convenient to use than a penalty parameter λ . The simultaneous ℓ_2 -penalty is to compensate for collinearity and large noise, and is later used to overcome some obstacles in backward elimination. The associated regularization parameter η_0 can be easily tuned and is not highly sensitive in experiments. Our theoretical analysis will reveal the benefits of a carefully designed shrinkage sequence for both numerical stability and statistical accuracy.

Problem (3) is nonconvex and includes a discrete constraint. While it can be challenging to computationally solve problems of this nature, it is possible to find a local minimum using a scalable iterative optimization algorithm. Moreover, in the era of big data, it may not be necessary to fully solve (3) in order to achieve good statistical performance for “regular” problems and analyzing algorithm-driven non-global estimators is crucial to discovering new and cost-effective methods for improving the statistical performance of nonconvex optimization. Concretely, to introduce a prototype algorithm, we first construct a surrogate function $g(\beta, \beta^-)$ for (3),

$$g(\beta, \beta^-) = \frac{1}{2} \|y - X\beta^-\|_2^2 + \langle X^T(X\beta^- - y), \beta - \beta^- \rangle + \frac{\rho}{2} \|\beta - \beta^-\|_2^2 + \frac{\eta_0}{2} \|\beta\|_2^2,$$

with $\rho > 0$ to be chosen later, and then define a sequence of iterates by

$$\beta^{(t+1)} = \arg \min_{\beta : \|\beta\|_0 \leq q} g(\beta, \beta^{(t)}). \quad (4)$$

Recall the quantile-thresholding operator $\Theta^\#$ defined at the end of Section I. With some simple algebra (details omitted), we obtain an iterative quantile-thresholding algorithm

$$\beta^{(t+1)} = \Theta^\# \left\{ \beta^{(t)} - \frac{1}{\rho} X^T(X\beta^{(t)} - y); q, \frac{\eta_0}{\rho} \right\}. \quad (5)$$

The first step amounts to the sure independence screening [3] when $\beta^{(0)} = 0$. However, (5) iterates to lessen greediness with a low per-iteration cost.

The update rule in (5) possesses some desirable computational properties. For instance, if ρ is large enough (more specifically, $\rho \geq \rho_+(2q)$ with $\rho_+(\cdot)$ defined in (2)), then the algorithm shows a worst-case sublinear convergence rate, regardless of the problem’s dimensions, coherence, and signal strength. The obtained solutions (though not necessarily optimal) can be characterized as *fixed points* of the algorithm

mapping defined in (4). For more results and technical details, please refer to Theorem A.1.

This class of procedures has been used in signal and information processing [11, 13], and in the special case of $\eta_0 = 0$, the plain update rule of (5) falls under the category of iterative hard-thresholding (IHT) algorithms [14, 15] which only exhibit mediocre performance (cf. Remark 3 and Section IV). In fact, there is much potential for improvement by adaptively adjusting the three key parameters ρ, η_0, q in (5), which has not been systematically explored in the literature.

A. Statistical error analysis: power and limitations

While optimization error is important for analyzing an algorithm, our main focus is on *statistical error*. This subsection investigates the prototype algorithm (5) to motivate new techniques in later sections. In order to obtain sharp nonasymptotic results for this algorithm, it is important to note that the thresholds vary from iteration to iteration and the final estimator may not be globally optimal.

Recall $y = X\beta^* + \epsilon$ with $\|\beta^*\|_0 \leq s$. Let

$$\vartheta := q/s$$

with $\vartheta > 1$ throughout the paper. A fixed point $\hat{\beta}$ associated with (5) that satisfies the following equation is called a $\Theta^\#$ -estimator,

$$\hat{\beta} = \Theta^\# \left\{ \hat{\beta} - \frac{1}{\rho} X^T (X\hat{\beta} - y); q, \bar{\eta}_0 \right\}, \text{ with } \bar{\eta}_0 = \eta_0/\rho. \quad (6)$$

Theorem 1 studies the statistical accuracy of these estimators.

Theorem 1. *Assume that ϵ is a sub-Gaussian random vector with mean zero and scale bounded by σ (cf. Definition A.1 in the appendix). Let $\hat{\beta}$ be any estimator satisfying (6) for some $\eta_0 \geq 0$ with $\|\hat{\beta}\|_0 = q$, and $\rho > 0$ be chosen such that*

$$\frac{\rho - \{(2 - \varepsilon)\sqrt{\vartheta} - 1\}\eta_0}{\sqrt{\vartheta}} \|\beta\|_2^2 \leq (2 - \delta) \|X\beta\|_2^2, \quad (7)$$

$$\forall \beta : \|\beta\|_0 \leq (1 + \vartheta)s$$

for some $\varepsilon, \delta > 0$. Then with probability at least $1 - Cp^{-c}$,

$$\|X(\hat{\beta} - \beta^*)\|_2^2 \sqrt{\frac{\eta_0 \varepsilon}{\delta}} \|\hat{\beta} - \beta^*\|_2^2 \lesssim \frac{1}{\delta^2} \sigma^2 \vartheta s \log \frac{ep}{\vartheta s} + \frac{\eta_0}{\delta \varepsilon} \|\beta^*\|_2^2, \quad (8)$$

where $C, c > 0$ are constants.

From the error bound, (5) can achieve the minimax optimal error rate of $\mathcal{O}(\sigma^2 s \log(ep/s))$ [16], under the assumption of (7) and when $\vartheta, \delta, \varepsilon$ are treated as constants. The result does not need η_0 to be exactly zero. In fact, a positive η_0 can actually be beneficial in satisfying the condition of (7) (e.g., $\rho = (1.9\sqrt{\vartheta} - 1)\eta_0 + 1.9\sqrt{\vartheta}\rho_-(q + s)$ and $\varepsilon = \delta = 0.1$, applicable to $q > n$). Another interesting observation is that ρ should be chosen to be properly small to achieve good statistical accuracy, which is in contrasts to the bound $\rho \geq \rho_+(2q)$ mentioned earlier for numerical convergence. The remarks below make some further extensions and comparisons.

Remark 1 (Estimation error bounds and faithful variable selection). *The ℓ_2 -recovery result of Theorem 1 is fundamental,*

and can be used to derive estimation error bounds in other norms under proper regularity conditions.

Theorem 2. *In the setup of Theorem 1, suppose the regularity condition (7) is replaced by*

$$\left\{ \frac{\rho - (2\sqrt{\vartheta} - 1)\eta_0}{\sqrt{\vartheta}} + \delta \rho_+((1 + \vartheta)s) \right\} \|\beta\|_2^2 \leq 2 \|X\beta\|_2^2, \quad (9)$$

$$\forall \beta : \|\beta\|_0 \leq (1 + \vartheta)s$$

for some $\delta > 0$. Then

$$\begin{aligned} \|\hat{\beta} - \beta^*\|_2^2 &\lesssim \frac{1}{\delta^2 \rho_+((1 + \vartheta)s)} \sigma^2 \vartheta s \log \frac{ep}{\vartheta s} \\ &\quad + \frac{\eta_0^2}{\delta^2 \rho_+((1 + \vartheta)s)} \|\beta^*\|_2^2 \end{aligned} \quad (10)$$

holds with probability at least $1 - Cp^{-c}$, for some $C, c > 0$. Moreover, under

$$\nu \|\beta\|_\infty \leq \|X^T X + \eta_0 I\|_\infty / n, \quad \beta : \|\beta\|_0 \leq (1 + \vartheta)s \quad (11)$$

for some $\nu > 0$, any fixed-point $\hat{\beta}$ satisfies

$$\|\hat{\beta} - \beta^*\|_\infty \leq \frac{(\rho + \eta_0)}{n\nu\sqrt{\vartheta - 1}} \frac{\|\hat{\beta} - \beta^*\|_2}{\sqrt{s}} + \frac{\|X^T \epsilon\|_\infty}{n\nu} + \frac{\eta_0}{n\nu} \|\beta^*\|_\infty, \quad (12)$$

$$\begin{aligned} \|(\hat{\beta} - \beta^*)_{\mathcal{J}^*}\|_\infty + (1 - \frac{\rho + \eta_0}{n\nu}) \|(\hat{\beta} - \beta^*)_{\hat{\mathcal{J}} \setminus \mathcal{J}^*}\|_\infty \\ \leq \frac{\|X^T \epsilon\|_\infty}{n\nu} + \frac{\eta_0}{n\nu} \|\beta^*\|_\infty, \end{aligned} \quad (13)$$

where $\mathcal{J}^* = \mathcal{J}(\beta^*)$, $\hat{\mathcal{J}} = \mathcal{J}(\hat{\beta})$.

Compared with (7), the condition of (9) replaces $\delta \|X\beta\|_2^2$ by $\delta \rho_+((1 + \vartheta)s) \|\beta\|_2^2$. When q and s are small, $\rho_+((1 + \vartheta)s)$ is of the order $\mathcal{O}(n)$. Therefore, (10) becomes $\|\hat{\beta} - \beta^*\|_2^2 \lesssim \{\sigma^2 s \log(ep/s)\}/n$, assuming δ, ϑ are constants and η_0 is properly small.

Moreover, the element-wise error bound (12) implies faithful variable selection under regularity condition (11) (which, like previous regularity conditions, favors low coherence, i.e., the off-diagonal entries of $X^T X/n$ should be relatively small in magnitude). Specifically, assuming ϑ, ν, δ are constants, $\|x_j\|_2 \lesssim \sqrt{n}$, $\rho + \eta_0 \lesssim n$ and the beta-min condition $\min_{j \in \mathcal{J}^*} |\beta_j^*| > c\sigma \{\log(ep)/n\}^{1/2}$ with a sufficient large constant c , (12) indicates that the s largest elements in $|\hat{\beta}_j|$ correspond to $\mathcal{J}^* = \{j : \beta_j^* \neq 0\}$ with high probability.

Remark 2 (Fixed points vs. globally optimal solutions). *The statistical accuracy results (8), (10), and (12) are proved for all nonglobal fixed-point estimators defined by (6). Our proof can be slightly modified to show that if a globally optimal solution can be computed, the statistical error rate remains unchanged but the left-hand side of (7) becomes 0, indicating that the regularity condition always holds for any $\delta \leq 2$. However, relying on multiple starting points to obtain a globally optimal solution and thus improve statistical performance can be inefficient for large datasets.*

Remark 3 (Comparison with some theoretical works). *The aforementioned class of IHT algorithms may refer to the use of hard-thresholding $\Theta_H(s; \lambda) = [s_i 1_{|s_i| \geq \lambda}]$ with a fixed*

threshold λ , or a varying threshold as the q/p -th quantile of $|s_i|$ ($1 \leq i \leq p$) by fixing q [14, 15]. In comparison, the ℓ_2 component in (5) should not be ignored, and it may result in a different sparsity pattern in the presence of high coherence and large p . Fairly speaking, the performance of IHT is not on par with some standard statistical methods and packages (such as the lasso). This is why we performed theoretical analysis in the hopes of discovering and developing new techniques.

In a theoretical study, [17] obtained a convergence result in terms of function value under

$$\vartheta > \rho_+^2(2q)/\rho_-^2(2q),$$

which improves the condition in [18]

$$\vartheta > 32\rho_+^2(2q)/\rho_-^2(2q).$$

Our condition in Theorem 1 is even less restrictive. For example, a sufficient condition for (7) is

$$\vartheta > \{\rho_+(2q) + \eta_0\}^2/[4\{\rho_-(q+s) + \eta_0\}^2],$$

or

$$\vartheta > \{\rho_+(2q) + \eta_0\}^2/[4\{\rho_-(q+s) + \eta_0\}^2]$$

because $\rho_-(q+s) \geq \rho_-(2q)$, which becomes $\vartheta > \rho_+^2(2q)/\{4\rho_-^2(2q)\}$ in the worst case of $\eta_0 = 0$. In conclusion, $32\rho_+^2(2q)/\rho_-^2(2q) \geq \rho_+^2(2q)/\rho_-^2(2q) \geq \rho_+^2(2q)/\{4\rho_-^2(2q)\} \geq \rho_+^2(2q)/[4\{\rho_-(q+s)\}^2] \geq \{\rho_+(2q) + \eta_0\}^2/[4\{\rho_-(q+s) + \eta_0\}^2]$, and our obtained error rate of $\sigma^2 s \log(ep/s)$ is minimax optimal.

Interested readers may also refer to [19–21, 9, 17, 22], for example, for the analyses of various penalties and mixed thresholding rules, with an error rate of $\sigma^2 s \log(ep)$. Since our purpose is to design a new backward selection algorithm for problems with a predetermined number of features, we will not discuss their technical assumptions. The experiments in Section IV make a comprehensive comparison of different methods in various scenarios.

B. New means of improvement for large-scale data

Providing provable guarantees for prediction, estimation, and variable selection is reassuring. But the real challenge lies in finding innovative techniques that can relax the required regularity conditions to ensure good statistical accuracy, while being more cost-effective than using multiple random starts. To gain further insights, we can use the restricted isometry numbers (as defined in (2)) to provide a sufficient condition for (7):

$$\rho < 2\sqrt{\vartheta}\rho_-(q+s) + (2\sqrt{\vartheta}-1)\eta_0 \text{ or } 4\vartheta > \frac{(\rho + \eta_0)^2}{(\rho_-(q+s) + \eta_0)^2}. \quad (14)$$

1) “Fast” learning: One key takeaway from the results presented in Section II-A is the importance of the inverse learning rate, ρ . In the field of machine learning, it is commonly advised to use a “slow” learning rate when training a nonconvex model. This can ensure good computational performance, as evidenced by the lower bound of ρ in Theorem A.1. However, it is important to note that according to (14),

using an excessively large value for ρ may compromise the statistical guarantee of the model.

In fact, (7) suggests that smaller values of ρ are preferred, and combining statistical and numerical analysis leads to the following range for ρ :

$$\rho_+(2q) \leq \rho \leq 2\sqrt{\vartheta}\rho_-(q+s) + (2\sqrt{\vartheta}-1)\eta_0. \quad (15)$$

In convex programming, the choice of stepsize does not affect the optimality of the solution as long as the algorithm converges. However, in our case of nonconvex constrained optimization, it is important to choose a large enough value for $1/\rho$ not only to gain fast convergence, but also to ensure statistical accuracy. To the best of our knowledge, this is a novel finding. Since it may not be easy to determine the theoretical restricted isometry numbers in practice, a routine line search for the step size can be used. Specifically, according to the proof in Appendix A, one can use the majorization condition $f(\beta^{(t+1)}) \leq g(\beta^{(t+1)}, \beta^{(t)})$ or $\|X(\beta^{(t+1)} - \beta^{(t)})\|_2^2 \leq \rho\|\beta^{(t+1)} - \beta^{(t)}\|_2^2$ to prevent ρ from becoming too large while still preserving the convergence properties stated in Theorem A.1. The concept of using an iteration-varying sequence ρ_t will be important in the next section.

2) “Backward” selection: Another important discovery is the influence of cardinality control. If we use a conservative inverse learning rate of $\rho = \rho_+(2q)$, then (14) imposes a limit on the restricted condition number of the design matrix:

$$\vartheta > [\rho_+(2q) + \eta_0]^2/[4\{\rho_-(q+s) + \eta_0\}^2]. \quad (16)$$

This suggests a promising approach to relax the regularity condition by increasing the value of ϑ .

Figure 1 confirms the point assuming random designs: the larger the value of q is, the more likely it is for (14) to hold on large-scale data. Random matrix theory also supports this idea.

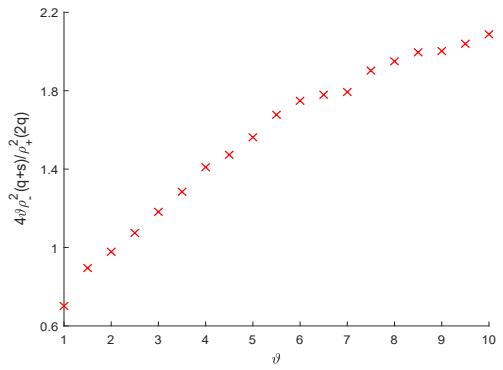


Fig. 1: An illustration of how $4\vartheta\rho_-^2(q+s)/\rho_+^2(2q)$ varies as ϑ increases. Here, the rows of X are independently drawn from a multivariate Gaussian distribution with zero mean and the covariance $\Sigma = [0.5^{|i-j|}]$, $n = 2,000, p = 4,000, s = 4$. To determine ρ_{\pm} for a given matrix X , we perform a random sampling. The results are averaged over 100 independent X ’s that are generated from the same distribution.

Theorem 3. Assume that the rows of the random matrix $X \in \mathbb{R}^{n \times p}$ are independent and identically distributed as $N(0, \Sigma)$,

where $\Sigma_{ii} \leq 1$. Let $\lambda_{\max}^{(2q)}$ be the largest eigenvalue of Σ_I for all $I \subset [p]$ with $|I| \leq 2q$, and $\lambda_{\min}^{(q+s)}$ be the smallest eigenvalue of Σ_I for all $I \subset [p]$ with $|I| \leq q+s$. Then for any $0 < c < 1$,

$$\frac{\rho_+(2q)}{\rho_-(q+s)} \leq \left\{ \frac{(1+c)\sqrt{\lambda_{\max}^{(2q)}} + \sqrt{2\lambda_{\max}^{(2q)}q \log(ep/q)/n} + \sqrt{2q/n}}{(1-c)\sqrt{\lambda_{\min}^{(q+s)}} - \sqrt{(\lambda_{\min}^{(q+s)}(q+s) \log(ep/q)/n) - \sqrt{(q+s)/n}}} \right\}^2 \quad (17)$$

with probability at least $1 - 2\exp(-nc^2/2)$, assuming $n \geq \{2(q+s)/(1-c)^2\}\{1/\lambda_{\min}^{(q+s)} + \log(ep/q)\}$.

The results can be extended to sub-Gaussian designs (by using, for example, Theorem 6.2 of [23] and Weyl's theorem). Let us consider the Toeplitz design $\Sigma = [\tau^{|i-j|}]$ with $0 \leq \tau < 1$. By the interlacing theorem,

$$\frac{1-\tau}{1+\tau} = \lambda_{\min}(\Sigma) \leq \lambda_{\min}^{(q+s)} \leq \lambda_{\max}^{(2q)} \leq \lambda_{\max}(\Sigma) = \frac{1+\tau}{1-\tau},$$

and so the right-hand side of (17) is bounded by a constant with high probability as $n \gg q \log(ep/q)$. Accordingly, the regularity condition can be satisfied with a properly large ϑ .

Of course, the error bound in (8) also increases with larger values of q . To address this issue, we propose employing a *decreasing* sequence of q_t to progressively tighten the cardinality constraint. Based on previous discussions, it is thus advisable to use *increasing* learning rates $1/\rho_t$ (such as $1/\rho_+(2q_t)$) in the iterative process. It may also be beneficial to adjust the shrinkage parameter to a sequence η_t , particularly when $q_t > n$. This resulting algorithm, which combines progressive quantiles, ℓ_2 -shrinkage, and learning rates, will be referred to as "slow kill." It differs from the pure optimization algorithm (5) with a fixed q and from various bottom-up boosting and greedy algorithms that are commonly used in the literature.

The purpose of this section is to provide a compelling rationale for certain aspects of slow kill techniques. We will present results in a more general setting, including fast convergence of the iterates and how slow kill improves the quality of the initial estimate as q_t approaches q , further relaxing the regularity conditions.

III. ADAPTIVE CONTROL OF QUANTILES, LEARNING RATES, AND ℓ_2 -SHRINKAGE

Given a general loss, based on the discussions in the last section, we pursue sparsity in β via

$$\min_{\beta \in \mathbb{R}^p} l_0(X\beta; y) + \frac{\eta_0}{2} \|\beta\|_2^2 \equiv l(\beta) + \frac{\eta_0}{2} \|\beta\|_2^2 \equiv f(\beta) \quad (18)$$

s.t. $\|\beta\|_0 \leq q$,

where for notational ease, $l_0(X\beta; y)$ is often abbreviated as $l(\beta)$. Again, the use of hybrid regularization is intended to address collinearity and large p . We assume that the regularization parameters q, η_0 are given in the algorithm design and theoretical analysis. (Of course, given q , one can easily tune the value of η_0 using methods such as AIC; as for the selection of q , an information criterion is provided in the Appendix H.) The generalized Bregman function for a differentiable l is one of the main tools we use to handle a variety of losses:

$$\Delta_l(\beta_1, \beta_2) := l(\beta_1) - l(\beta_2) - \langle \nabla l(\beta_2), \beta_1 - \beta_2 \rangle, \quad (19)$$

where the differentiability can be replaced by directional differentiability to analyze a wide range of algorithms in statistical computation [22]. If l is also strictly convex, Δ_l becomes the standard Bregman divergence [24, 25]. When $l(\cdot) = \|\cdot\|_2^2/2$, $\Delta_l(\beta_1, \beta_2) = \|\beta_1 - \beta_2\|_2^2/2$, which is symmetric, and we abbreviate it to $\mathbf{D}_2(\beta_1, \beta_2)$. Define the symmetrized version of $\Delta_l(\beta_1, \beta_2)$ by $\tilde{\Delta}_l(\beta_1, \beta_2) := \{\Delta_l(\beta_1, \beta_2) + \Delta_l(\beta_2, \beta_1)\}/2$, where $\tilde{\Delta}_l(\beta_1, \beta_2) = \Delta_l(\beta_2, \beta_1)$. As an extension of (2), we introduce two generalized restricted isometry numbers $\rho_+^l(s_1, s_2), \rho_-^l(s_1, s_2)$ that satisfy

$$\Delta_l(\beta_1, \beta_2) \leq \rho_+^l(s_1, s_2) \mathbf{D}_2(\beta_1, \beta_2), \forall \beta_i : \|\beta_i\|_0 \leq s_i, i = 1, 2 \quad (20)$$

$$\Delta_l(\beta_1, \beta_2) \geq \rho_-^l(s_1, s_2) \mathbf{D}_2(\beta_1, \beta_2), \forall \beta_i : \|\beta_i\|_0 \leq s_i, i = 1, 2. \quad (21)$$

We differentiate s_1, s_2 because Δ_l may not be symmetric. These numbers will be convenient and useful for theoretical purposes; for example, Theorem 4 and Theorem 5 will use positive $\rho_+^l(q, q)$ and $\rho_+^l(q, s)$, respectively, while Theorem 6 will use nonnegative ρ_-^l . When $l(\beta) = \|X\beta - y\|_2^2/2$, $\Delta_l(\beta_1, \beta_2) = \|X\beta_1 - X\beta_2\|_2^2/2$ and $\rho_+^l(s_1, s_2) = \rho_+(s_1 + s_2)$. More generally, if the gradient of $l_0(\cdot; y)$ is L -Lipschitz continuous, as is the case in regression or logistic regression,

$$\|\nabla l_0(\xi_1; y) - \nabla l_0(\xi_2; y)\|_2 \leq L\|\xi_1 - \xi_2\|_2, \quad (22)$$

for all $\xi_1, \xi_2 \in \mathbb{R}^n$, then it is easy to show that

$$\rho_+^l(s_1, s_2) \leq L\rho_+(s_1 + s_2) (\leq L\|X\|_2^2). \quad (23)$$

A. Numerical convergence and statistical accuracy for the general optimization algorithm

First, we extend the previous iterative quantile-thresholding algorithm to handle losses that may not be quadratic. Construct the following surrogate function

$$g(\beta, \beta^-) = l_0(X\beta; y) + \frac{\eta_0}{2} \|\beta\|_2^2 + (\rho \mathbf{D}_2 - \Delta_l)(\beta, \beta^-), \quad (24)$$

which is by linearizing the loss (only). Then, similar to the derivation in Section II, (24) leads to an algorithm

$$\beta^{(t+1)} = \Theta^{\#} \left\{ \beta^{(t)} - \frac{1}{\rho} X^T \nabla l_0(X\beta^{(t)}; y); q, \frac{\eta_0}{\rho} \right\}. \quad (25)$$

Some basic numerical properties are summarized as follows.

Theorem 4. Assume that $\inf_{\xi, y} l_0(\xi; y) > -\infty$. Consider (25) starting from an arbitrary feasible $\beta^{(0)}$. Then $\rho \geq \rho_+^l(q, q)$ guarantees that for all $t \geq 0$, $f(\beta^{(t+1)}) \leq g(\beta^{(t+1)}, \beta^{(t)})$ and $(\rho - \rho_+^l(q, q))\mathbf{D}_2(\beta^{(t+1)}, \beta^{(t)}) \leq f(\beta^{(t)}) - f(\beta^{(t+1)})$, and so the objective function values converge as $t \rightarrow \infty$. Assume $\rho > \rho_+^l(q, q)$, $\eta_0 > 0$ and ∇l_0 is continuous. Then every accumulation point $\hat{\beta}$ of $\beta^{(t)}$ satisfies the fixed-point equation

$$\hat{\beta} = \Theta^{\#} \{ \hat{\beta} - X^T \nabla l_0(X\hat{\beta}; y)/\rho; q, \eta_0/\rho \}. \quad (26)$$

Furthermore, if $l_0(\cdot; y)$ is convex, $\lim_{t \rightarrow \infty} \beta^{(t)} = \hat{\beta}$, and under $\|\hat{\beta}\|_0 = q$, $\hat{\beta}$ is a local minimizer to problem (18) and the support of $\beta^{(t)}$ stabilizes in finitely many iterations.

Next, we turn to the statistical accuracy of the estimators that are defined by (26). To overcome the obstacle that the

loss is not necessarily associated with a probability density function, we define the concept of *effective noise* with respect to the statistical truth β^* as

$$\epsilon = -\nabla l_0(X\beta^*; y), \quad (27)$$

where we treat X as fixed and y as random in this section. The definition of effective noise in (27) does not depend on the regularizer. In the special case of a generalized linear model with cumulant function b and canonical link function $g = (b')^{-1}$, the loss is $l(\beta) = l_0(X\beta; y) = -\langle y, X\beta \rangle + \langle 1, b(X\beta) \rangle$, and so $\epsilon = y - g^{-1}(X\beta^*) = y - \mathbb{E}y$. For regression, the effective noise term ϵ is equivalent to the raw noise, which is usually assumed to be Gaussian. In the case of classification using the logistic deviance, ϵ is bounded, making it sub-Gaussian. In fact, any loss function with a bounded derivative, such as Huber's loss, Hampel's loss, or the hinge loss, will always result in a sub-Gaussian ϵ , regardless of the distribution of y . In this section, we assume that the effective noise is a sub-Gaussian random vector with mean zero and scale bounded by σ . However, our proof techniques can be applied more generally. The following theorem provides a risk bound for the estimators obtained by (25), and also demonstrates the impact of the quality of the starting point on the regularity condition.

Theorem 5. Let $\hat{\beta} : \|\hat{\beta}\|_0 = q$ be an estimate obtained from (25) with a feasible starting point $\beta^{(0)}$, namely, $\hat{\beta} \in \min_{\|\beta\|_0 \leq q} g(\beta, \hat{\beta})$ and $f(\hat{\beta}) \leq f(\beta^{(0)})$ with $\|\beta^{(0)}\|_0 \leq q$. Define

$$P_o(q) = q \log(ep/q). \quad (28)$$

Suppose that $\beta^{(0)}$ satisfies

$$\mathbb{E}\mathbf{D}_2(\beta^{(0)}, \beta^*) = \mathcal{O}(M) \frac{\sigma^2 P_o(q) + \sigma^2}{n} \text{ for some } 1 \leq M \leq +\infty. \quad (29)$$

Let $Q = \{\rho_+(q+s)M/n\}^{1/2} + \{\rho_+^l(q, s) + \eta_0\}M/n$. Assume for some $\delta > 0$, $0 < \varepsilon \leq 1$ and large $K \geq 0$,

$$\begin{aligned} & K\sigma^2 P_o(\vartheta s) + \left\{ 2(1 - \frac{1}{M})\bar{\Delta}_{l_0} + \frac{C}{M(Q\delta \vee 1)}\Delta_{l_0} - \delta\mathbf{D}_2 \right\} (X\beta, X\beta') \\ & \geq \frac{1 - 1/M}{\sqrt{\vartheta}} [\rho - \{(2 - \varepsilon)\sqrt{\vartheta} - 1\}\eta_0] \mathbf{D}_2(\beta, \beta') \end{aligned} \quad (30)$$

$\forall \beta, \beta' : \|\beta\|_0 \leq \vartheta s, \|\beta'\|_0 \leq s,$

where C is some positive constant. Then

$$\begin{aligned} & \mathbb{E}\{\mathbf{D}_2(X\hat{\beta}, X\beta^*) \vee \frac{\eta_0 \varepsilon}{\delta} \mathbf{D}_2(\hat{\beta}, \beta^*)\} \\ & \lesssim \frac{K\delta \vee 1}{\delta^2} \left\{ \sigma^2 \vartheta s \log\left(\frac{ep}{\vartheta s}\right) + \sigma^2 \right\} + \frac{\eta_0}{\delta \varepsilon} \|\beta^*\|_2^2. \end{aligned} \quad (31)$$

Therefore, we can achieve the desired level of statistical accuracy as long as K, δ, ϑ are constants and η_0 is not excessively large. When $M = +\infty$ (no requirement on $\beta^{(0)}$), the regularity condition (30) becomes

$$\begin{aligned} & \frac{\rho - \{(2 - \varepsilon)\sqrt{\vartheta} - 1\}\eta_0}{\sqrt{\vartheta}} \mathbf{D}_2(\beta, \beta') \leq (2\bar{\Delta}_{l_0} - \delta\mathbf{D}_2)(X\beta, X\beta') \\ & + K\sigma^2 P_o(\vartheta s), \quad \forall \beta, \beta' : \|\beta\|_0 \leq \vartheta s, \|\beta'\|_0 \leq s, \end{aligned}$$

which includes (7) as a special case. But when one uses a decent starting point, (30) is much more relaxed. In the

extreme case where $M = 1$, the right-hand side of (30) becomes 0, and so with μ -restricted strong convexity $(\bar{\Delta}_{l_0} - \delta\mathbf{D}_2)(X\beta, X\beta') \geq 0$ for $\|\beta\|_0 \leq \vartheta s, \|\beta'\|_0 \leq s$, (30) is always satisfied.

B. Slow kill: algorithm design & sequential analysis

Using a multi-start strategy to select a high-quality initial value for $\beta^{(0)}$ may be computationally infeasible for large-scale data. Fortunately, we will see that designing iteration-varying thresholding and shrinkage can effectively relax the statistical regularity conditions and improve the statistical accuracy of the sequence of iterates.

More concretely, slow kill modifies the optimization algorithm (25) by introducing three auxiliary sequences $\rho_{t+1}, q_{t+1}, \eta_{t+1}$

$$\begin{aligned} \beta^{(t+1)} &= \Theta^{\#}\{\beta^{(t)} - \rho_{t+1}^{-1} X^T \nabla l_0(X\beta^{(t)}; y); q_{t+1}, \bar{\eta}_{t+1}\} \\ & \text{with } \bar{\eta}_{t+1} = \eta_{t+1}/\rho_{t+1} \end{aligned} \quad (32)$$

where $q_t \rightarrow q, \eta_t \rightarrow \eta_0$. The scaled shrinkage sequence $\bar{\eta}_t$ will be more convenient to use than the raw sequence η_t in later analysis. We want to understand whether adapting the inverse learning rate, cardinality, and ℓ_2 -shrinkage parameters during the iteration can lead to improved performance. Specifically, we aim to investigate how the statistical accuracy of $\beta^{(t)}$ changes as t increases, and under what conditions the error converges geometrically fast. The focus of Theorem 6 is on the statistical error of $\beta^{(t)}$ with respect to the statistical truth β^* , rather than on their optimization errors relative to a specific minimizer β^o . We will see that in principle, slow kill benefits from decreasing q_t and ρ_t . It is also worth noting that the error bound in (35) places no requirements on $\vartheta_t, \rho_t, \eta_t$.

Theorem 6. Let the sequence of iterates $\beta^{(t)} : \|\beta^{(t)}\|_0 = q_t$ be generated from (32) with a feasible $\beta^{(0)}$. Given any $t \geq 1$, define

$$h_t^{-1} = (1 - \frac{1}{\sqrt{\vartheta_t}})(\rho_t + \eta_t) + (1 - \varepsilon)(\rho_-^l(q_t, s) + \eta_t) \quad (33)$$

$$\kappa_t = (\rho_t - \rho_-^l(s, q_t))h_t, \quad (34)$$

where ε is an arbitrary number in $(0, 1]$. Then the following recursive statistical error bound

$$\begin{aligned} & \mathbf{D}_2(\beta^*, \beta^{(T+1)}) + \sum_{t=0}^T (\Pi_{\tau=t}^T h_{\tau+1}) (\rho_{t+1} \mathbf{D}_2 - \Delta_l)(\beta^{(t+1)}, \beta^{(t)}) \\ & \leq \sum_{t=0}^T (\kappa_{t+1} \cdots \kappa_{T+1}) \cdot \\ & \quad \left\{ \frac{A\sigma^2}{\varepsilon} \frac{\rho_+(q_{t+1} + s)}{\left(\frac{\rho_-^l(q_{t+1}, s)}{\rho_{t+1}} \vee \bar{\eta}_{t+1}\right) \left(1 - \frac{\rho_-^l(s, q_{t+1})}{\rho_{t+1}}\right) \rho_{t+1}^2} \cdot \vartheta_{t+1} s \log\left(\frac{ep}{\vartheta_{t+1} s}\right) \right. \\ & \quad \left. + \frac{\bar{\eta}_{t+1}}{\left(1 - \frac{\rho_-^l(s, q_{t+1})}{\rho_{t+1}}\right) \varepsilon} \|\beta^*\|_2^2 \right\} + \left(\Pi_{t=0}^T \kappa_{t+1} \right) \mathbf{D}_2(\beta^*, \beta^{(0)}). \end{aligned} \quad (35)$$

holds for all $T \geq 0$, with probability at least $1 - Cp^{-cA}$, where C, c are positive constants.

The corollary below showcases the usefulness of the theorem on algorithm configuration.

Corollary 1. In the setup of Theorem 6, given any $\varepsilon \in (0, 1]$, if ρ_t and η_t are chosen to satisfy

$$\rho_{t+1} \geq \rho_+^l(q_{t+1}, q_t) \quad (36)$$

$$\bar{\eta}_t \geq 0 \vee \frac{(1/\sqrt{\vartheta_t} + \varepsilon) - 2(\rho_-^l(s, q_t) \wedge \rho_-^l(q_t, s))/\rho_t}{2 - 1/\sqrt{\vartheta_t} - \varepsilon} \quad (37)$$

so that $(\rho_{t+1}\mathbf{D}_2 - \Delta_l)(\beta^{(t+1)}, \beta^{(t)}) \geq 0$ and $\kappa_t \leq (1 + \varepsilon)^{-1}$, then with probability at least $1 - Cp^{-cA}$ the statistical error of $\{\beta^{(t)}\}$ decays geometrically fast,

$$\begin{aligned} & \mathbf{D}_2(\beta^*, \beta^{(T+1)}) \\ & \leq \left(\frac{1}{1 + \varepsilon} \right)^{T+1} \mathbf{D}_2(\beta^*, \beta^{(0)}) + \frac{1}{\varepsilon} \sum_{t=0}^T \left(\frac{1}{1 + \varepsilon} \right)^{T-t+1} E_{t+1} \end{aligned} \quad (38)$$

for all $T \geq 0$, where

$$\begin{aligned} E_{t+1} = & \left\{ 1 - \frac{\rho_-^l(s, q_{t+1})}{\rho_+^l(q_{t+1}, q_t)} \right\}^{-1} \left\{ \frac{A\sigma^2}{\frac{\rho_-^l(q_{t+1}, s)}{\rho_+^l(q_{t+1}, q_t)} \vee \bar{\eta}_{t+1}} \right. \\ & \left. - \frac{\rho_+(q_{t+1} + s)}{(\rho_+^l(q_{t+1}, q_t))^2} \vartheta_{t+1}s \log \left(\frac{ep}{\vartheta_{t+1}s} \right) + \bar{\eta}_{t+1} \|\beta^*\|_2^2 \right\}. \end{aligned} \quad (39)$$

The theoretical results provide valuable insights into the design of the three main elements of slow kill. Let's first apply Theorem 6 to analyze the basic optimization algorithm with fixed quantiles $q_t \equiv q$ and universal values $\rho_t \equiv \rho$, $\bar{\eta}_t \equiv \bar{\eta}$. (38) then shows linear convergence of the statistical error, with the first term on the right-hand side indicating the impact of the initial point. Because $\sum_{t=0}^T \{1/(1 + \varepsilon)\}^{T-t+1} \leq 1/\varepsilon$, the final error is of the order

$$\frac{\rho_+(q + s)}{(\rho_+^l(q, q))^2} \sigma^2 \vartheta s \log \left(\frac{ep}{\vartheta s} \right) + \bar{\eta} \|\beta^*\|_2^2, \quad (40)$$

where the restricted condition number $\rho_+^l(q, q)/\{\rho_-^l(s, q) \wedge \rho_-^l(q, s)\}$ and ε are assumed to be constants. The lower bound derived in (37) can help reduce the bias, and suggests the benefit of using a large quantile in this regard.

On the other hand, large quantiles can lead to an inflated variance term $\vartheta_{t+1}s \log\{ep/(\vartheta_{t+1}s)\}$ in (39), which motivates the use of decreasing quantiles, the most distinctive feature of slow kill. Indeed, a more careful examination of (38) shows that the factor $1/(1 + \varepsilon)^{T-t+1}$ allows for much larger q_t to be used in earlier iteration steps. This is because for small t , the associated error E_{t+1} will be more heavily shrunk in the final bound. Although it can be difficult to theoretically derive the optimal cooling scheme for the sequence q_t , various schemes seem to perform well in practice, such as $q_{t+1} = \lfloor q + (T-t)/(aTt + bT) \rfloor$ (inverse) or $\lfloor q + (p - q)/1 + a \exp(bt/T)^c \rfloor$ (sigmoidal), among others.

After q_t is given, the choice of ρ_{t+1} can be determined theoretically using (36): $\rho_{t+1} \geq \rho_+^l(q_{t+1}, q_t)$, which gives an upper bound of the stepsize to prevent slow kill from diverging. In implementation, $\rho_+^l(q_{t+1}, q_t)$ is often unknown. With regular design matrices (such as Toeplitz), a constant multiple of $L\{n + q_{t+1} \log(ep/q_{t+1})\}$ can be employed based on (A.17) in the proof of Appendix D, assuming that ∇l_0 is L -Lipschitz continuous. More generally, seen from the second term on the left-hand side of (35), we can use a line search with criterion

$$(\rho_{t+1}\mathbf{D}_2 - \Delta_l)(\beta^{(t+1)}, \beta^{(t)}) \geq 0. \quad (41)$$

See Appendix I for some implementation details of the line search. (41) enforces the majorization condition at $(\beta^{(t+1)}, \beta^{(t)})$, and so the resulting ρ_{t+1} can be even smaller than $\rho_+^l(q_{t+1}, q_t)$. The importance of limiting the size of ρ_t was previously discussed in Section II-B for ℓ_0 -constrained regression. Similarly, having a smaller ρ_{t+1} can help achieve a larger ε , which in turn leads to faster convergence and smaller error, as demonstrated in (33) and (37).

The lower bound for the scaled ℓ_2 -shrinkage sequence $\bar{\eta}_t$ in Corollary 1 can be rewritten as

$$2\sqrt{\vartheta_t} > \frac{\rho_{t+1} + \bar{\eta}_t \rho_t}{\rho_-^l(s, q_t) \wedge \rho_-^l(q_t, s) + \bar{\eta}_t \rho_t}. \quad (42)$$

It is similar to a restricted condition number condition, and extends (16) to a general loss. Specifically, when $2q_t > n$, (37) implies $\bar{\eta}_t > (1/\sqrt{\vartheta_t})/(2 - 1/\sqrt{\vartheta_t}) = 1/(2\sqrt{\vartheta_t} - 1)$, and as a result, we recommend using a scaled shrinkage sequence defined by

$$\bar{\eta}_t = 1/(2\sqrt{q_t/\bar{s}} - 1), \quad (43)$$

where $\bar{s} = q \wedge nL^2/\log(ep)$ (a surrogate for s , according to Appendix F) and L is the Lipschitz parameter of ∇l_0 . (43) plays an important role in early slow kill iterations and is independent of the learning rate.

Our analyses support the use of the ℓ_2 -assisted backward quantile control to gradually tighten the constraint. The update formula (32) used in slow kill has a strong foundation in optimization, which gives it an advantage over heuristics based multi-stage procedures. The fast geometric convergence established in Theorem 6, together with a strong signal strength, indicates that the zeros in $\beta^{(t)}$ represent irrelevant predictors with high probability (cf. Remark 1 and Appendix G). This allows us to occasionally squeeze the design matrix using $\mathcal{J}(\beta^{(t+1)})$ (e.g., when q_{t+1} reaches $p/2^k$) to reduce the problem size (Appendix I). The apparent junk features are thus removed at an early stage, saving computational cost, while the more difficult to identify irrelevant features are addressed only when we are close to finding an optimal solution. This trait makes slow kill particularly well-suited for big data learning. Slow kill offers similar advantages in group variable selection [11] and low-rank matrix estimation [26].

In contrast, forward pathwise and boosting algorithms [5, 27, 6–8, 19, 9] grow a model from the null in a *bottom-up* fashion. Such algorithms must consider almost all features at each iteration, making them computationally intensive, as they often require hundreds or thousands of boosting iterations. Motivated by the ℓ_0 -optimization perspective, we can also investigate a class of “steady grow” procedures in which q_t increases from 0 to q in (32). Compared with boosting, the update and selection would incorporate the effect of the previous estimate in addition to the gradient. A retaining option can be introduced in steady grow that works in the opposite way to the squeezing operation in slow kill. The investigation of retaining and squeezing, as well as a combination of slow kill and steady grow, is left for future research.

Finally, how to obtain a sparse model with a prescribed cardinality is the problem of interest throughout the paper. But if one wants to determine the best value for q , we suggest using

a predictive information criterion [28] that can guarantee the optimal prediction error rate in a nonasymptotic sense (which is presented in Appendix H).

IV. EXPERIMENTS

A. Simulations

In this part, we conduct simulation studies to compare the performance of slow kill (abbreviated as SK in tables and figures below) with some popular sparse learning methods in terms of prediction accuracy, selection consistency, and computational efficiency. Unless otherwise mentioned, the rows \tilde{x}_i^T of the predictor matrix $X = [\tilde{x}_1, \dots, \tilde{x}_n]^T \in \mathbb{R}^{n \times p}$ are independently generated from a multivariate normal distribution with covariance matrix Σ , where Σ either has a Toeplitz structure $[\tau^{|i-j|}]$ or has equal correlations $[\tau_{1_{i \neq j}}]$. High correlation strengths such as $\tau = 0.9$ will be included in our experiments. We consider both regression and classification with a sparse β^* : $\beta_j^* = 1$, if $j = 10k + 1, 0 \leq k < s$ and so $s = \|\beta^*\|_0$. In the regression experiments, $y = X\beta^* + \epsilon$ with $\epsilon_i \sim N(0, 1)$, and for the classification experiments, $y_i = 1$ if $\tilde{x}_i^T \beta^* > 0$ and 0 otherwise.

In addition to slow kill, the following methods are included for comparison: lasso [29], elastic net (ENET) [12], MCP [4], SCAD [30], and IHT and NIHT ([15, 31], for regression only). (We also evaluated the performance of picasso [9] in simulations as an improved version of [19]. However, its pathwise computation resulted in worse error rates and missing rates than standard nonconvex optimization on the synthetic data. Therefore, we did not present the results. We will include the algorithm in our experiments with real data in later sections.) The quadratic loss is used in regression and the logistic deviance is used in classification. For slow kill, we take a simple single starting point $\beta^{(0)} = 0$ and $\eta_0 = 50$; an inverse cooling schedule $q_{t+1} = \lfloor q + (T - t)/\{tT/(p - q) + 2T/(p - 2q)\} \rfloor$ ($0 \leq t \leq T$) is used so that $q_T = q$ and $q_1 = p/2$, and we set $T = 100$ in all experiments for convenience and efficiency. We use the R package `glmnet` to implement lasso and elastic net, the package `ncpen` [32] for the aforementioned nonconvex penalties, and the package `sparsify` for IHT methods. (The core of `glmnet` is implemented using Fortran subroutines, while `ncpen` is mainly based on C++. Our implementation of slow kill could potentially be made more efficient and require less memory by using C or Fortran, but it already performs comparably or better than the other methods, as shown in later tables and figures.) To ensure a fair comparison and eliminate the influence of different parameter tuning schemes, we select the estimate with 1.5s nonzeros for each method. To calibrate the bias, we refit each obtained model using only the selected variables. All other algorithmic parameters are set to their default values.

Given each simulation setup, we repeat the experiment for 50 times and evaluate the performance of each algorithm according to the measures defined below: the missing rate $\times 100\%$ and the prediction error. Concretely, the missing rate is the fraction of undetected true variables, and in regression, the prediction error is calculated by 10 times $(\hat{\beta} - \beta^*)^T \Sigma (\hat{\beta} - \beta^*)$ using the true signal, while in classification, it refers to the

misclassification error rate $\times 100\%$ on a separate test set containing the same number of observations as the training dataset. The total computational time (in seconds) is also included to describe the computational cost. Since the implementation of a penalized method often uses warm starts, we terminate the algorithm once it reaches an estimate with the prescribed cardinality.

Table I shows some experiment results in the regression setup. Figure 2 plots more results of some representative methods when varying the sparsity level s and the correlation strength τ (excluding elastic net and IHT, because their performance is similar to that of lasso and poor, respectively). It can be seen that slow kill outperforms the other methods in terms of both statistical accuracy and computational time, particularly in more challenging situations with more relevant features and coherent designs.

TABLE I: Regression: performance comparison in terms of prediction error, missing rate and computational time with different correlation structures. In more details, $p = 5,000$, $n = 150$, $s = 10$ and $\Sigma = [\tau^{|i-j|}]$ or $[\tau_{1_{i \neq j}}]$ with $\tau = 0.9$

	Toeplitz structure			Equal correlation		
	Error	Miss	Time	Error	Miss	Time
LASSO	16	32	5	15	83	13
ENET	16	31	13	14	82	34
IHT	85	68	55	16	88	57
NIHT	12	22	4	17	80	18
MCP	12	23	34	18	78	24
SCAD	12	23	13	16	85	6
SK	2	2	1	12	50	1

For classification, Table II and Figure 3 make a comparison between different methods with various correlation structures and problem dimensions, and similar conclusions can be drawn. It is important to note that the excellent statistical accuracy of slow kill is *not* accompanied by a sacrifice in computational time compared to other methods. In fact, as seen in Figure 3, slow kill offers substantial time savings especially when n is large, while being very successful at selection and prediction.

Next, we present some experiments in which the signal strength is varied. Recall that in the regression setup, we set $\beta_j^* = 1$ for $j \in \mathcal{J}(\beta^*)$. For $n = 100$, $p = 5000$, $\sigma = 1$, the minimax optimal rate is approximately $\sigma \sqrt{(\log p)/n} (\approx 0.292)$ (ignoring the constant factor for which a sharp value

TABLE II: Classification: performance comparison in terms of prediction error, missing rate and computational time with different correlation structures. In more details, $p = 2,000$, $n = 500$, $s = 10$ and $\Sigma = [\tau^{|i-j|}]$ or $[\tau_{1_{i \neq j}}]$ with $\tau = 0.9$

	Toeplitz structure			Equal correlation		
	Error	Miss	Time	Error	Miss	Time
LASSO	8.0	24	10	5.1	95	49
ENET	8.0	25	31	4.7	95	135
MCP	6.9	23	15	5.0	93	20
SCAD	7.0	22	22	5.1	94	16
SK	2.2	2	4	3.9	78	4

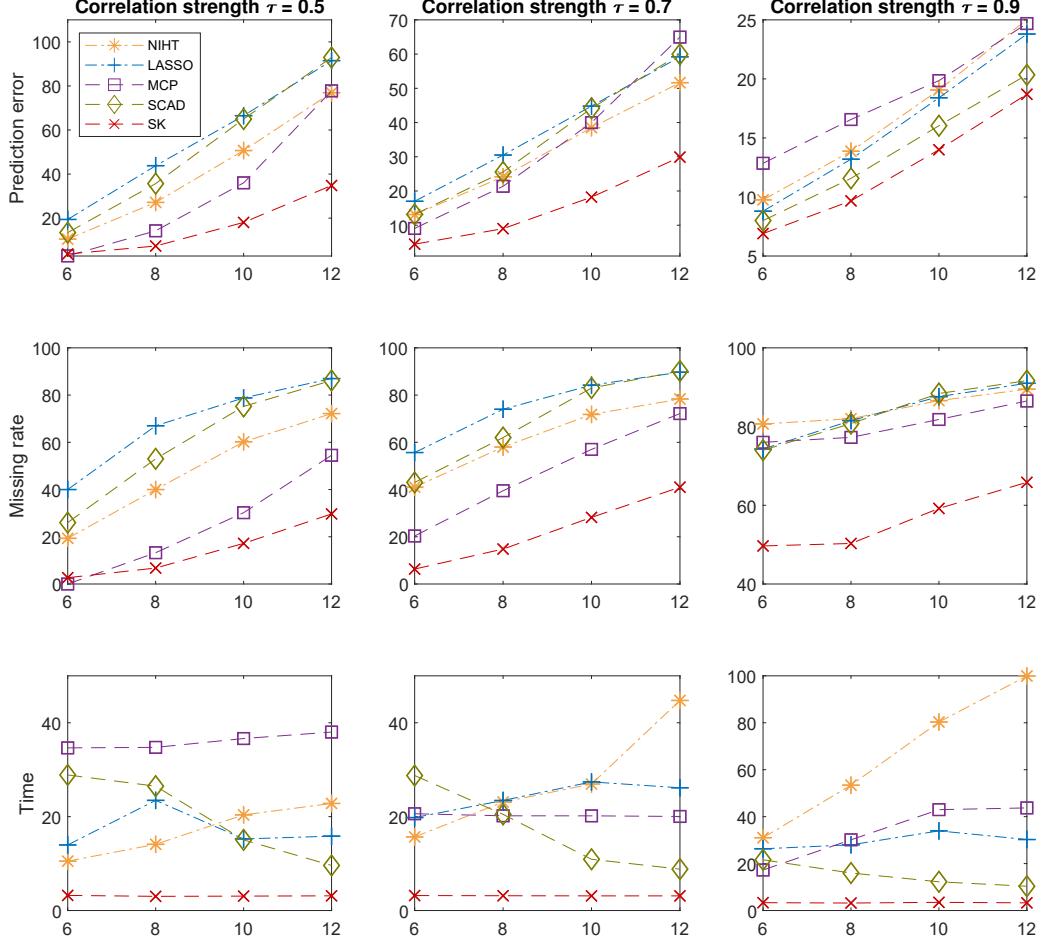


Fig. 2: Regression: performance comparison in terms of prediction error, missing rate and computational time when varying the sparsity and the correlation strength of the model. In more details, $p = 10,000$, $n = 150$, $s = 6, 8, 10, 12$ and $\Sigma = [\tau \mathbf{1}_{i \neq j}]$ with $\tau = 0.5, 0.7, 0.9$.

may be difficult to derive). We conducted additional experiments by setting $\beta_j^* = 0.8, 0.6, 0.4, 0.2$. The comparison results for different methods are demonstrated in Figure 4. As the signal strength was low (e.g., $\beta_j^* = 0.2, 0.4$), all methods performed poorly. For higher values, slow kill outperformed the other methods by a large margin.

We conducted another experiment to explore larger values of $\|\beta^*\|_2^2$. (As a reminder, in the previous setting where $s = 10$ and $\beta_j^* = 1, \forall j \in \mathcal{J}(\beta^*)$, we had $\|\beta^*\|_2^2 = 10$.) We tested $\|\beta^*\|_2^2 = 50, 100, 150, 200$ by scaling up each β_j^* by a corresponding factor. The results of this experiment are shown in Figure 5. As $\|\beta^*\|_2^2$ increases, NIHT, MCP, and slow kill exhibit clear advantages, with the latter two showing similar prediction errors and missing rates.

B. Handwritten digits classification

The Gisette dataset [33] was created to classify the highly confusing digits 4 and 9 for handwritten digit recognition. There are 5,000 predictors, including various pixel constructed features as well as some ‘probes’ with little predictive power. Because the exact number of relevant features is unknown, we assess the performance of different methods given the same model cardinality to make a fair comparison. We randomly split the 7,000 samples into a training subset with 3,000

samples and a test subset with 4,000 samples for 20 times to report the average misclassification error rate and total computational time.

Due to the relatively large size of the data, computational efficiency is a major concern. Many statistical packages were unable to deliver meaningful results in a reasonable amount of time. Here, we compare the glmnet [34], logitboost [35, 36], picasso with the MCP option [37], and slow kill with different numbers of selected features.

According to Figure 6, logitboost and picasso achieved better misclassification error rates on the dataset than glmnet, but slow kill consistently performed the best. In terms of computational cost, glmnet and slow kill were extremely scalable; logitboost was quite expensive even for just $q = 40$, and picasso suffered a similar issue when $q \geq 60$.

C. Breast cancer microarray data

The breast-cancer microarray dataset [38] from the Curated Microarray Database contains 35,981 gene expression levels of 143 tumor samples of patients with breast cancer and 146 paired adjacent normal breast tissue samples. The goal is to identify some differentially expressed genes to help the classification of normal and tumor tissues. We randomly split the dataset into a training subset (60%) and a test subset (40%)

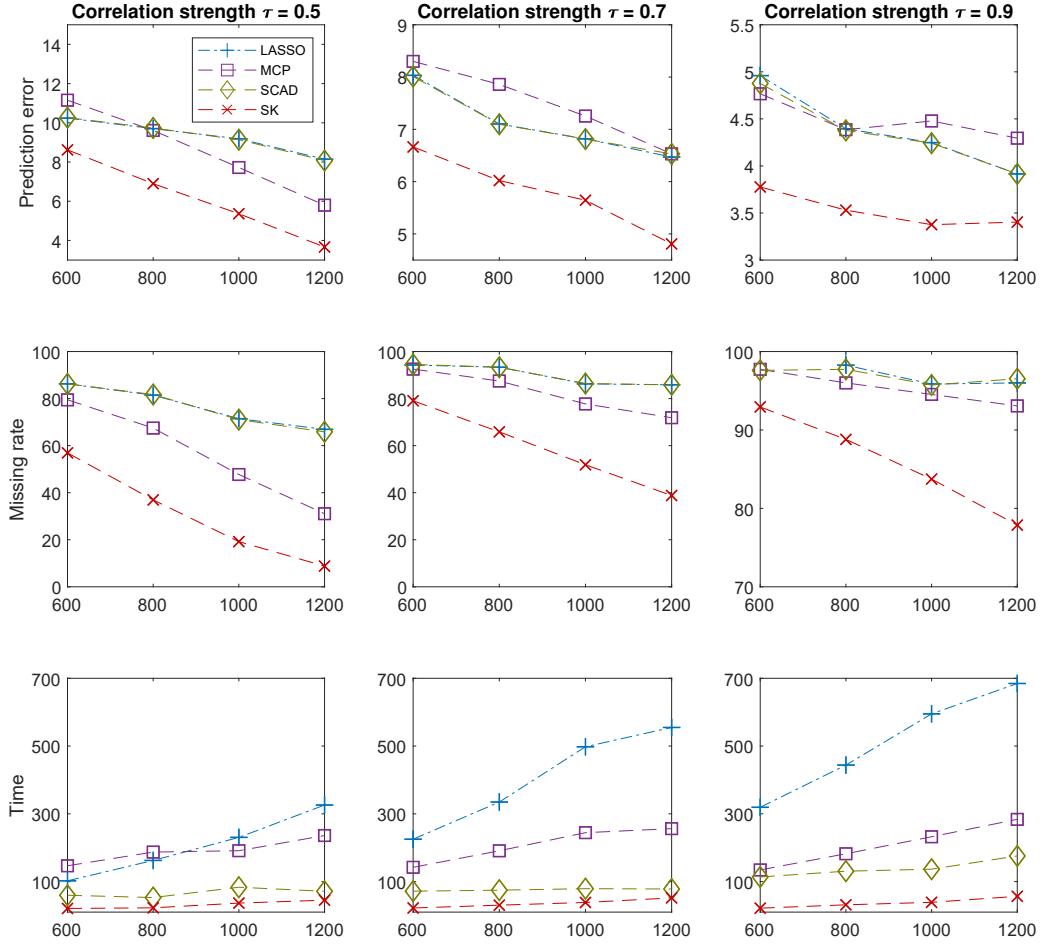


Fig. 3: Classification: performance comparison in terms of prediction error, missing rate and computational time with different correlation structures and sample sizes. In more details, $p = 10,000$, $n = 600, 800, 1000, 1200$, $s = 15$ and $[\tau_{1 \neq j}]$ with $\tau = 0.5, 0.7, 0.9$.

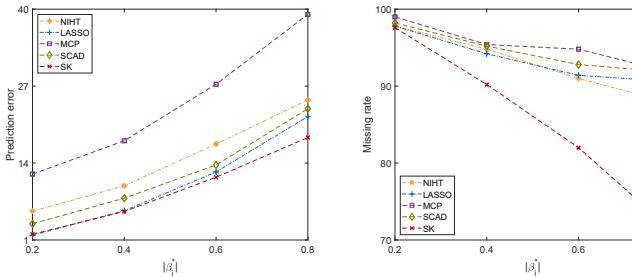


Fig. 4: Comparison of prediction errors (left) and missing rates (right) of different methods under different signal strengths. The details of the regression setup are given in Section IV-A, and we set $p = 5,000$, $n = 100$, $s = 10$, $\tau = 0.8$, and $\beta_j^* = 0.2, 0.4, 0.6, 0.8$ for $j \in \mathcal{J}(\beta^*)$.

for 20 times and report the misclassification error rates and total computational time of different methods in Table III.

According to Tables III, logitboost has the highest computational complexity, and picasso shows the worst overall classification performance on this dataset. In contrast, glmnet and slow kill can achieve lower misclassification error rates, and the latter is much more cost-effective according to our experiments.

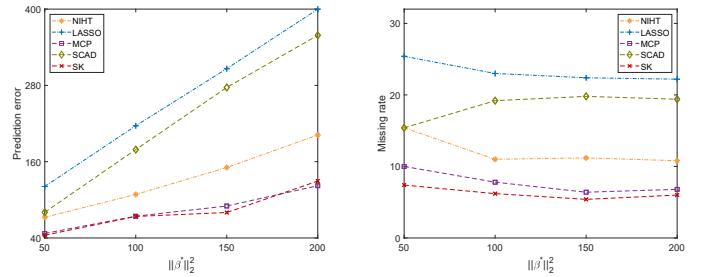


Fig. 5: Comparison of prediction errors (left) and missing rates (right) of different methods for large signals. The details of the regression setup are given in Section IV-A, and we set $p = 5,000$, $n = 100$, $s = 10$, $\tau = 0.8$, and $\|\beta^*\|_2^2 = 50, 100, 150, 200$ (by scaling up each β_j^*).

D. Sub-Nyquist spectrum sensing and learning

Sub-Nyquist sampling-based wideband spectrum sensing for millimeter wave is an important topic for next-generation wireless communication systems. With a multi-co-set sampler [39], a multiple-measurement-vector model in signal processing can be formulated as $Y = XB^* + \mathcal{E}$, where the goal is to exploit the joint (row-wise) weak sparsity of B^* to reconstruct the spectrum. Here, all the matrices are complex (e.g., $Y \in \mathbb{C}^{n \times m}$, $X \in \mathbb{C}^{n \times p}$), and the size of the predictor

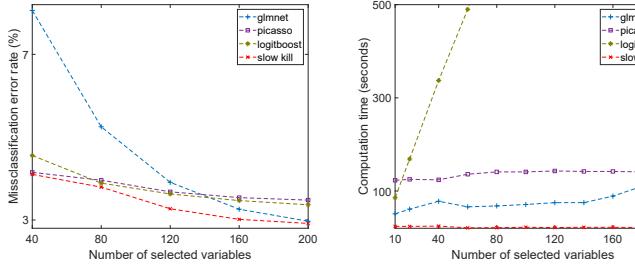


Fig. 6: Gisette data. Left panel: mean misclassification error rate, panel: total computational time, with different numbers of selected tures. Picasso is too costly compared with the other methods and only of its cost curve is shown.

TABLE III: Breast cancer microarray data: misclassification error ($\times 100\%$) and total computational time (in seconds)

	$q = 60$		$q = 80$		$q = 100$		$q = 120$		$q = 140$	
	Error	Time	Error	Time	Error	Time	Error	Time	Error	Time
GLMNET	10.9	19	10.7	19	10.5	50	10.2	50	10.2	50
PICASSO	11.4	43	11.3	43	11.1	48	11.3	48	11.2	42
LogitBoost	11.2	500	11.2	680	10.9	860	10.6	1080	10.8	1220
SK	10.8	10	10.2	11	10.2	11	10.1	11	9.8	11

matrix X is determined by the number of cosets and the number of channels; interested reader may refer to [40] for more detail. Nicely, with the Hermitian inner product $\langle A, B \rangle \triangleq \text{tr}\{A^H B\}$ in place of the real inner product, and the generalized Bregman function redefined as $\Delta_l(B_1, B_2) = l(B_1) - l(B_2) - \langle \nabla l(B_2), B_1 - B_2 \rangle / 2 - \langle B_1 - B_2, \nabla l(B_2) \rangle / 2$, all of our theorems and algorithms can be extended to the complex group sparsity pursuit.

We compared our method with two popular methods, SOMP [41] and JB-HTP [42], on a benchmark time-domain dataset in [43]. Table IV shows the normalized mean square error $\|\hat{B} - B^*\|_F / \|B^*\|_F$ of each method as we vary q (the number of selected channels). A demonstration of spectral recovery is plotted in Figure 7, where the predictive information criterion in Appendix H was used for model selection in slow kill.

TABLE IV: Spectrum reconstruction error in terms of normalized mean square error

	$q = 3$	$q = 4$	$q = 5$	$q = 6$	$q = 7$	$q = 8$
SOMP	0.83	0.93	0.82	0.91	0.92	0.94
JB-HTP	0.94	1.00	0.99	0.95	1.07	0.96
SK	0.74	0.65	0.53	0.38	0.42	0.50

V. SUMMARY

This paper proposed a new slow kill method for large-scale variable selection. It is a scalable optimization-based algorithm that uses three carefully designed and theoretically justified sequences of thresholds, shrinkage, and learning rates.

Intuitively, slow kill uses a novel backward quantile control with adaptive ℓ_2 shrinkage and increasing learning rates to relax regularity conditions and overcome obstacles in backward elimination. This method is significantly different from

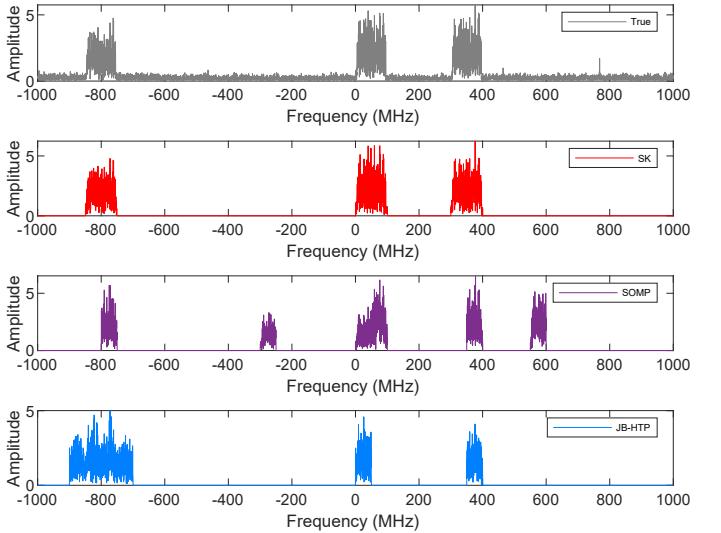


Fig. 7: Spectrum sensing results by different methods.

boosting and many forward stagewise procedures in the existing literature. Our theoretical studies led to insights on how to design a progressive hybrid regularization to achieve the optimal error rate and fast convergence. The technique is applicable to a general loss that is not necessarily a negative log-likelihood function, and its ability to reduce the problem size throughout the iteration makes it attractive for big data.

APPENDIX

The definition of a sub-Gaussian random variable or a sub-Gaussian random vector is standard in the literature.

Definition A.1. We call ξ a sub-Gaussian random variable if it has mean zero and the scale (ψ_2 -norm) for ξ , defined as $\inf\{\sigma > 0 : \mathbb{E}[\exp(\xi^2/\sigma^2)] \leq 2\}$, is finite. We call $\xi \in \mathbb{R}^p$ a sub-Gaussian random vector with scale bounded by σ if all one-dimensional marginals $\langle \xi, \alpha \rangle$ are sub-Gaussian satisfying $\|\langle \xi, \alpha \rangle\|_{\psi_2} \leq \sigma \|\alpha\|_2$, for any $\alpha \in \mathbb{R}^p$. Similarly, a random matrix ξ is called sub-Gaussian if $\text{vec}(\xi)$ is sub-Gaussian.

A. Theorem A.1 and Theorem 4

First, for the algorithm (5) defined in the setup of Section II, we have the following numerical properties.

Theorem A.1. Given any X, y and $\beta^{(0)}$, the sequence of iterates $\beta^{(t)}$ generated by (5) satisfies $f(\beta^{(t)}) - f(\beta^{(t+1)}) \geq \rho \|\beta^{(t+1)} - \beta^{(t)}\|_2^2 / 2 - \|X(\beta^{(t+1)} - \beta^{(t)})\|_2^2 / 2$, $\forall t \geq 0$ and so when $\rho \geq \rho_+(2q)$, $f(\beta^{(t)})$ converges, and $\beta^{(t)}$ satisfies

$$\min_{0 \leq t \leq T} \|\beta^{(t+1)} - \beta^{(t)}\|_2^2 \leq \frac{1}{T+1} \frac{2f(\beta^{(0)})}{\rho - \rho_+(2q)}.$$

Moreover, as long as $\rho > \rho_+(2q)$ and $\eta_0 > 0$, $\beta^{(t)}$ has a unique limit point $\hat{\beta}$ that satisfies the ‘‘fixed-point’’ equation

$$\beta = \Theta^\# \{ \beta - X^T(X\beta - y)/\rho; q, \eta_0/\rho \},$$

and when $\|\hat{\beta}\|_0 = q$, $\hat{\beta}$ is also a local minimizer of problem (3).

To prove the first conclusion in Theorem A.1, notice that in the regression setting,

$$\begin{aligned} & g(\beta^{(t+1)}, \beta^{(t)}) - f(\beta^{(t+1)}) \\ &= \rho \|\beta^{(t+1)} - \beta^{(t)}\|_2^2/2 - \|X(\beta^{(t+1)} - \beta^{(t)})\|_2^2/2, \end{aligned}$$

and thus

$$\begin{aligned} & f(\beta^{(t)}) - f(\beta^{(t+1)}) \\ &\geq \frac{\rho}{2} \|\beta^{(t+1)} - \beta^{(t)}\|_2^2 - \frac{1}{2} \|X(\beta^{(t+1)} - \beta^{(t)})\|_2^2, \quad \forall t \geq 0. \end{aligned}$$

Taking the summation from $t = 0$ to $t = T$ and using the fact that $\|X(\beta^{(t+1)} - \beta^{(t)})\|_2^2 \leq \rho_+(2q) \|\beta^{(t+1)} - \beta^{(t)}\|_2^2$, we have

$$\frac{(\rho - \rho_+(2q))}{2} \sum_{t=0}^T \|\beta^{(t+1)} - \beta^{(t)}\|_2^2 \leq f(\beta^{(0)}) - f(\beta^{(T+1)}),$$

which leads to

$$\min_{0 \leq t \leq T} \|\beta^{(t+1)} - \beta^{(t)}\|_2^2 \leq \frac{2}{(T+1)(\rho - \rho_+(2q))} f(\beta^{(0)}).$$

Next, we consider the general problem and prove Theorem 4, which implies the second part of Theorem A.1. From $\inf_{\xi, y} l_0(\xi; y) > -\infty$, we assume without loss of generality that $l_0(\xi; y) \geq 0$. Recall $l_0(X\beta; y)$ is abbreviated as $l(\beta)$ and thus $\nabla l(\beta) = X^T \nabla l_0(X\beta)$ by the chain rule.

From the construction $g(\beta, \beta^{(t)}) = f(\beta) + (\rho \mathbf{D}_2 - \Delta_{l_0})(\beta, \beta^{(t)})$, we get

$$(\rho \mathbf{D}_2 - \Delta_{l_0})(\beta^{(t+1)}, \beta^{(t)}) + f(\beta^{(t+1)}) \leq g(\beta^{(t)}, \beta^{(t)}) = f(\beta^{(t)}).$$

When $\rho \geq \rho_+(q, q)$, $(\rho \mathbf{D}_2 - \Delta_{l_0})(\beta^{(t+1)}, \beta^{(t)}) \geq 0$, from which it follows that the sequence of $f(\beta^{(t)})$ is non-increasing and convergent. In fact, one just needs

$$f(\beta^{(t+1)}) \leq g(\beta^{(t+1)}, \beta^{(t)}) \quad (\text{A.1})$$

to enjoy the function value convergence, which can be used for line search.

In addition, we obtain

$$(\rho - \rho_+(q, q)) \mathbf{D}_2(\beta^{(t+1)}, \beta^{(t)}) \leq f(\beta^{(t)}) - f(\beta^{(t+1)}).$$

Finally, let us study the limit points of the sequence of iterates. We first notice that $\{\beta^{(t)}\}_{t=0}^\infty$ is uniformly bounded under $\eta_0 > 0$, since

$$\eta_0 \|\beta^{(t)}\|_2^2/2 \leq f(\beta^{(t)}) \leq f(\beta^{(0)}).$$

From $\lim_{t \rightarrow \infty} \{f(\beta^{(t)}) - f(\beta^{(t+1)})\} = 0$, $\lim_{t \rightarrow \infty} (\rho \mathbf{D}_2 - \Delta_{l_0})(\beta^{(t+1)}, \beta^{(t)}) = 0$, and because $\rho > \rho_+(q, q)$,

$$\lim_{t \rightarrow \infty} (\beta^{(t+1)} - \beta^{(t)}) = 0.$$

Let $\hat{\beta}$ be any limit point of $\beta^{(t)}$ satisfying $\hat{\beta} = \lim_{k \rightarrow \infty} \beta^{(j_k)}$ for some sequence j_k . Then

$$\begin{aligned} 0 &= \lim_{k \rightarrow \infty} (\beta^{(j_k+1)} - \beta^{(j_k)}) \\ &= \lim_{k \rightarrow \infty} \Theta^\# \{ \beta^{(j_k)} - \nabla l(\beta^{(j_k)})/\rho; q, \eta_0/\rho \} - \hat{\beta} \\ &= \Theta^\# \{ \hat{\beta} - \nabla l(\hat{\beta})/\rho; q, \eta_0/\rho \} - \hat{\beta}, \end{aligned}$$

where the second equality is due to the continuity of $\nabla l(\beta)$ and the $\Theta^\#$ -uniqueness assumption.

Define $\hat{\mathcal{J}} = \{j : \hat{\beta}_j \neq 0\}$. Then we get

$$\hat{\beta}_{\hat{\mathcal{J}}} = \hat{\beta}_{\hat{\mathcal{J}}}/(1 + \eta_0/\rho) - X_{\hat{\mathcal{J}}}^T \nabla l_0(X_{\hat{\mathcal{J}}} \hat{\beta}_{\hat{\mathcal{J}}}; y)/(\rho + \eta_0),$$

or equivalently,

$$\eta_0 \hat{\beta}_{\hat{\mathcal{J}}} + X_{\hat{\mathcal{J}}}^T \nabla l_0(X_{\hat{\mathcal{J}}} \hat{\beta}_{\hat{\mathcal{J}}}; y) = 0.$$

Therefore, given $\hat{\mathcal{J}}$, $\hat{\beta}_{\hat{\mathcal{J}}}$ is a stationary point of

$$\min_{\gamma} l_0(X_{\hat{\mathcal{J}}} \gamma; y) + \eta_0 \|\gamma\|_2^2/2. \quad (\text{A.2})$$

When $l_0(\cdot; y)$ is convex and $\eta_0 > 0$, (A.2) is strongly convex and thus $\hat{\beta}_{\hat{\mathcal{J}}}$ is the unique minimizer.

By Ostrowski's convergence theorem, the set of limit points of $\beta^{(t)}$ must be connected. On the other hand, the set of all restricted optimal solutions $\{\hat{\beta}_{\hat{\mathcal{J}}}\}$ is finite, and so

$$\lim_{t \rightarrow \infty} \beta^{(t)} = \hat{\beta}.$$

Under $\|\hat{\mathcal{J}}\|_0 = q$, it is easy to see that the neighborhood $\{\beta : \|\beta - \hat{\beta}\|_\infty < \epsilon, J(\beta) \leq q\}$ with $0 < \epsilon < \min_{j \in \hat{\mathcal{J}}} |\hat{\beta}_j|$ is just $\{\beta : J(\beta) = \hat{\mathcal{J}}, |\beta_j - \hat{\beta}_j| < \epsilon, \forall j \in \hat{\mathcal{J}}\}$. The local optimality of $\hat{\beta}$ and support stability of $\beta^{(t)}$ thus follow.

B. Proof of Theorem 1

We first introduce some lemmas that are helpful in proving the theorem. The first is a generalization of Lemma 9 in [44].

Lemma A.1. *Let $\mathcal{J}(B)$ denote the row support of matrix B and define $J(B) = \|B\|_{2,0} = |\mathcal{J}(B)|$. Consider the following problem with $0 \leq q \leq p, \eta \geq 0$:*

$$\min_{B \in \mathbb{R}^{p \times m}} \frac{1}{2} \|Y - B\|_F^2 + \frac{\eta}{2} \|B\|_F^2 = l(B) \quad \text{subject to } \|B\|_{2,0} \leq q.$$

Then $\hat{B} = \vec{\Theta}^\#(Y; q, \eta)$ (recall $\vec{\Theta}^\#$ defined in Section I) gives a globally optimal solution, and for any B satisfying $J(B) \leq s$, we have

$$l(B) - l(\hat{B}) \geq (1 - \mathcal{L}(\mathcal{J}, \hat{\mathcal{J}}))(1 + \eta) \frac{\|\hat{B} - B\|_F^2}{2} \quad (\text{A.3})$$

where $\mathcal{J} = \mathcal{J}(B)$, $\hat{\mathcal{J}} = \mathcal{J}(\hat{B})$, and $\mathcal{L}(\mathcal{J}, \hat{\mathcal{J}}) = \sqrt{|\mathcal{J} \setminus \hat{\mathcal{J}}|/|\hat{\mathcal{J}} \setminus \mathcal{J}|}$. When $J(\hat{B}) = q$ with $\vartheta (\equiv q/s) \geq 1$, $\mathcal{L}(\mathcal{J}, \hat{\mathcal{J}}) \leq \sqrt{|\mathcal{J}|/|\hat{\mathcal{J}}|} \leq 1/\sqrt{\vartheta}$. In the above statement, 0/0 is understood as 1.

Lemma A.2. *There exist universal constants $A, C, c > 0$ such that for any $a > 0$, the following event*

$$\begin{aligned} & \sup_{\beta_1, \beta_2} \langle \epsilon, X(\beta_1 - \beta_2) \rangle - \frac{1}{2a} \|X(\beta_1 - \beta_2)\|_2^2 \\ & - \frac{a}{2} A \sigma^2 \{ J(\beta_1) \vee J(\beta_2) \} \log \left\{ \frac{ep}{J(\beta_1) \vee J(\beta_2)} \right\} \geq \frac{a}{2} \sigma^2 t \end{aligned} \quad (\text{A.4})$$

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

First, by definition, it is easy to show that $\hat{\beta}$ satisfies

$$\hat{\beta} \in \operatorname{argmin}_{\beta} g(\beta, \hat{\beta}),$$

where $g(\beta, \beta^-) = \|y - X\beta^-\|_2^2/2 + \langle X^T(X\beta^- - y), \beta - \beta^- \rangle + \rho\|\beta - \beta^-\|_2^2/2 + \eta_0\|\beta\|_2^2/2$. By $g(\hat{\beta}, \hat{\beta}) \leq g(\beta^*, \hat{\beta})$ and Lemma A.1,

$$\begin{aligned} & \frac{1}{2}\|\beta^* - \hat{\beta} + \frac{1}{\rho}X^T(X\hat{\beta} - y)\|_2^2 - \frac{1}{2}\|\frac{1}{\rho}X^T(X\hat{\beta} - y)\|_2^2 \\ & + \frac{\eta_0}{2\rho}\|\beta^*\|_2^2 - \frac{\eta_0}{2\rho}\|\hat{\beta}\|_2^2 \geq (1 + \frac{\eta_0}{\rho})\frac{1 - \mathcal{L}(\mathcal{J}^*, \hat{\mathcal{J}})}{2}\|\hat{\beta} - \beta^*\|_2^2, \end{aligned}$$

where $\mathcal{J}^* = \mathcal{J}(\beta^*)$, $\hat{\mathcal{J}} = \mathcal{J}(\hat{\beta})$, and $\mathcal{L}(\mathcal{J}^*, \hat{\mathcal{J}}) \leq 1/\sqrt{\vartheta}$.

It follows from the model $y = X\beta^* + \epsilon$ that

$$\begin{aligned} & \|X\hat{\beta} - X\beta^*\|_2^2 + \frac{\eta_0}{2}\|\hat{\beta}\|_2^2 \\ & \leq \frac{\rho - (\sqrt{\vartheta} - 1)\eta_0}{2\sqrt{\vartheta}}\|\hat{\beta} - \beta^*\|_2^2 + \frac{\eta_0}{2}\|\beta^*\|_2^2 + \langle X\hat{\beta} - X\beta^*, \epsilon \rangle, \end{aligned}$$

which gives

$$\begin{aligned} & \|X\hat{\beta} - X\beta^*\|_2^2 + \frac{\eta_0}{2}\|\hat{\beta} - \beta^*\|_2^2 \\ & \leq \frac{\rho - (\sqrt{\vartheta} - 1)\eta_0}{2\sqrt{\vartheta}}\|\hat{\beta} - \beta^*\|_2^2 + \eta_0\langle \hat{\beta} - \beta^*, -\beta^* \rangle \\ & \quad + \langle X\hat{\beta} - X\beta^*, \epsilon \rangle \\ & \leq \frac{\rho - (\sqrt{\vartheta} - 1)\eta_0}{2\sqrt{\vartheta}}\|\hat{\beta} - \beta^*\|_2^2 + \frac{b\eta_0}{2}\|\hat{\beta} - \beta^*\|_2^2 \\ & \quad + \frac{\eta_0}{2b}\|\beta^*\|_2^2 + \langle X\hat{\beta} - X\beta^*, \epsilon \rangle \end{aligned} \quad (\text{A.5})$$

for any $b > 0$. Applying Lemma A.2 with $t = 0$, we can show that for any $a > 0$, the following event

$$\langle X\hat{\beta} - X\beta^*, \epsilon \rangle \leq \frac{1}{2a}\|X\hat{\beta} - X\beta^*\|_2^2 + \frac{a}{2}A\sigma^2\vartheta s \log \frac{ep}{\vartheta s} \quad (\text{A.6})$$

occurs with probability at least $1 - Cp^{-c}$, where $A, C, c > 0$ are some universal constants.

Combining (A.5), (A.6) and the regularity condition (7) yields

$$\begin{aligned} & \frac{\eta_0(\varepsilon - b)}{2}\|\hat{\beta} - \beta^*\|_2^2 + \left(\frac{\delta}{2} - \frac{1}{2a}\right)\|X\hat{\beta} - X\beta^*\|_2^2 \\ & \leq \frac{\eta_0}{2b}\|\beta^*\|_2^2 + \frac{a}{2}A\sigma^2\vartheta s \log \frac{ep}{\vartheta s} \end{aligned}$$

with probability at least $1 - Cp^{-c}$. By choosing $a = 2/\delta$ and $b = \varepsilon/2$, we have the bound for the prediction error as

$$\begin{aligned} & \|X\hat{\beta} - X\beta^*\|_2^2 + \frac{\eta_0\varepsilon}{\delta}\|\hat{\beta} - \beta^*\|_2^2 \\ & \leq \frac{4\eta_0}{\delta\varepsilon}\|\beta^*\|_2^2 + \frac{4}{\delta^2}A\sigma^2\vartheta s \log \frac{ep}{\vartheta s} \lesssim \frac{\eta_0}{\delta\varepsilon}\|\beta^*\|_2^2 + \frac{1}{\delta^2}\sigma^2\vartheta s \log \frac{ep}{\vartheta s} \end{aligned}$$

which holds with probability at least $1 - Cp^{-c}$.

Proof of Lemma A.1 In this proof, given a matrix $B \in \mathbb{R}^{p \times m}$ and an index set $\mathcal{I} \subset [p]$, we use $B_{\mathcal{I}}$ to denote the submatrix of B by extracting its rows indexed by \mathcal{I} . Let $\mathcal{J}_1 = \mathcal{J} \cap \hat{\mathcal{J}}$, $\mathcal{J}_2 = \hat{\mathcal{J}} \setminus \mathcal{J}$ and $\mathcal{J}_3 = \mathcal{J} \setminus \hat{\mathcal{J}}$. Then $\mathcal{J} = \mathcal{J}_1 \cup \mathcal{J}_3$ and $\hat{\mathcal{J}} = \mathcal{J}_1 \cup \mathcal{J}_2$.

It can be easily shown that $\hat{B}_{\mathcal{J}_1} = Y_{\mathcal{J}_1}/(1 + \eta)$ and $\hat{B}_{\mathcal{J}_2} = Y_{\mathcal{J}_2}/(1 + \eta)$. By writing $B_{\mathcal{J}_1} = Y_{\mathcal{J}_1}/(1 + \eta) + \Delta_{\mathcal{J}_1}$ and $B_{\mathcal{J}_3} = Y_{\mathcal{J}_3}/(1 + \eta) + \Delta_{\mathcal{J}_3}$, we have

$$\begin{aligned} l(B) - l(\hat{B}) &= \frac{1 + \eta}{2}\|\Delta_{\mathcal{J}_1}\|_F^2 + \frac{1}{2(1 + \eta)}\|Y_{\mathcal{J}_2}\|_F^2 \\ &+ \frac{1 + \eta}{2}\|\Delta_{\mathcal{J}_3}\|_F^2 - \frac{1}{2(1 + \eta)}\|Y_{\mathcal{J}_3}\|_F^2, \\ \frac{1 + \eta}{2}\|\hat{B} - B\|_F^2 &= \frac{1 + \eta}{2}\|\Delta_{\mathcal{J}_1}\|_F^2 + \frac{1}{2(1 + \eta)}\|Y_{\mathcal{J}_2}\|_F^2 \\ &+ \frac{1 + \eta}{2}\|\frac{1}{1 + \eta}Y_{\mathcal{J}_3} + \Delta_{\mathcal{J}_3}\|_F^2. \end{aligned}$$

Let $K \leq 1$ satisfy

$$l(B) - l(\hat{B}) \geq \frac{K}{2}(1 + \eta)\|\hat{B} - B\|_F^2,$$

which is implied by

$$\begin{aligned} & \frac{1}{2(1 + \eta)}\|Y_{\mathcal{J}_2}\|_F^2 + \frac{1 + \eta}{2}\|\Delta_{\mathcal{J}_3}\|_F^2 - \frac{1}{2(1 + \eta)}\|Y_{\mathcal{J}_3}\|_F^2 \\ & \geq \frac{K}{2(1 + \eta)}\|Y_{\mathcal{J}_2}\|_F^2 + \frac{K(1 + \eta)}{2}\|\frac{1}{1 + \eta}Y_{\mathcal{J}_3} + \Delta_{\mathcal{J}_3}\|_F^2. \end{aligned} \quad (\text{A.7})$$

(A.7) is equivalent to

$$\begin{aligned} & (1 - K)\|Y_{\mathcal{J}_2}\|_F^2 + (1 + \eta)^2\|\Delta_{\mathcal{J}_3}\|_F^2 \\ & \geq (1 + \eta)^2K\|\frac{1}{1 + \eta}Y_{\mathcal{J}_3} + \Delta_{\mathcal{J}_3}\|_F^2 + \|Y_{\mathcal{J}_3}\|_F^2. \end{aligned} \quad (\text{A.8})$$

By construction, $\|y_i\|_2 \geq \|y_j\|_2$ for any $i \in \mathcal{J}_2$ and $j \in \mathcal{J}_3$. Thus $\|Y_{\mathcal{J}_2}\|_F^2 \geq J_2\|Y_{\mathcal{J}_3}\|_F^2/J_3$, from which it follows that (A.8) is implied by

$$\begin{aligned} & \{(1 - K)\frac{J_2}{J_3} - (1 + K)\}\|Y_{\mathcal{J}_3}\|_F^2 + (1 - K)(1 + \eta)^2\|\Delta_{\mathcal{J}_3}\|_F^2 \\ & \geq 2K(1 + \eta)\langle Y_{\mathcal{J}_3}, \Delta_{\mathcal{J}_3} \rangle. \end{aligned}$$

Therefore, restricting K to $(1 + K)/(1 - K) \leq J_2/J_3$ or $K \leq (J_2 - J_3)/(J_2 + J_3) \leq 1$, the largest possible K should satisfy

$$\{(1 - K)(J_2/J_3) - (1 + K)\} \cdot (1 - K) = |K|^2$$

or $(1 - K)^2 = J_3/J_2$, or $K = 1 - \sqrt{J_3/J_2} (\leq (J_2 - J_3)/(J_2 + J_3))$. This gives

$$\mathcal{L} = 1 - K = (J_3/J_2)^{1/2}.$$

Note that when $\mathcal{J}_2 = \emptyset$, K can take $-\infty$ for $\mathcal{J}_3 \neq \emptyset$ and 0 for $\mathcal{J}_3 = \emptyset$ to ensure (A.8).

Now assume $J(\hat{B}) = q$ with $\vartheta \geq 1$. If $\mathcal{J}_2 \neq \emptyset$, $\mathcal{L} \leq \sqrt{(J_3 + J_1)/(J_2 + J_1)} = \sqrt{J/\hat{J}} \leq 1/\sqrt{\vartheta}$. Otherwise, we must have $\mathcal{J}_3 = \emptyset$, $\mathcal{J} = \hat{\mathcal{J}}$ and $\vartheta = 1$. The proof is complete.

The lemma can be used in the analysis of ℓ_0 -constrained (elementwise) sparsity pursuit, as well as group variable selection (cf. Section IV-D).

Proof of Lemma A.2 Given a matrix A , denote by \mathcal{P}_A the orthogonal projection onto its range, and \mathcal{P}_A^\perp its orthogonal complement. In the proof, $\mathcal{P}_{\mathcal{J}}$ is used as a short notation for $\mathcal{P}_{X_{\mathcal{J}}}$ in the proof for any $J \subset [p]$. Let $\mathcal{J}_1 = \mathcal{J}(\beta_1)$, $\mathcal{J}_2 = \mathcal{J}(\beta_2)$, $J_1 = |\mathcal{J}_1|$, $J_2 = |\mathcal{J}_2|$.

First, note that the term $\{J(\beta_1) \vee J(\beta_2)\} \log[ep/\{J(\beta_1) \vee J(\beta_2)\}]$ is used in (A.4), instead of $J(\beta_1 - \beta_2) \log\{ep/J(\beta_1 - \beta_2)\}$, and although $J(\beta_1 - \beta_2) \leq J(\beta_1) + J(\beta_2)$, $J(\beta_1) + J(\beta_2)$ can be larger than p . To tackle the issue, we employ a decomposition trick

$$\begin{aligned} X\beta_1 - X\beta_2 &= \mathcal{P}_{\mathcal{J}_1}X(\beta_1 - \beta_2) + \mathcal{P}_{\mathcal{J}_1}^\perp X(\beta_1 - \beta_2) \\ &= \mathcal{P}_{\mathcal{J}_1}X(\beta_1 - \beta_2) + \mathcal{P}_{\mathcal{J}_1}^\perp \mathcal{P}_{\mathcal{J}_2}X(\beta_1 - \beta_2). \end{aligned}$$

Let $\Delta = \beta_1 - \beta_2$. Then

$$\langle \epsilon, X\Delta \rangle = \langle \epsilon, \mathcal{P}_{\mathcal{J}_1}X\Delta \rangle + \langle \epsilon, \mathcal{P}_{\mathcal{J}_1}^\perp \mathcal{P}_{\mathcal{J}_2}X\Delta \rangle. \quad (\text{A.9})$$

Let us bound the first term on the right-hand side of (A.9). Define $P_o(J) = \sigma^2 J \log(ep/J)$ for $0 \leq J \leq p$, which is an increasing function, and $\Gamma_J = \{\alpha \in \mathbb{R}^p : \|\alpha\|_2 \leq 1, \alpha \in \mathcal{P}_{\mathcal{J}} \text{ for some } \mathcal{J} \subset [p], |\mathcal{J}| \leq J\}$. Then for any $a, b > 0$

$$\begin{aligned} &\langle \epsilon, \mathcal{P}_{\mathcal{J}_1}X\Delta \rangle - \frac{1}{a}\|\mathcal{P}_{\mathcal{J}_1}X\Delta\|_2^2 - bLP_o(J_1) \\ &\leq \|\mathcal{P}_{\mathcal{J}_1}X\Delta\|_2 \langle \epsilon, \frac{\mathcal{P}_{\mathcal{J}_1}X\Delta}{\|\mathcal{P}_{\mathcal{J}_1}X\Delta\|_2} \rangle - 2\|\mathcal{P}_{\mathcal{J}_1}X\Delta\|_2 \sqrt{\frac{b}{a}LP_o(J_1)} \\ &\leq \frac{1}{a}\|\mathcal{P}_{\mathcal{J}_1}X\Delta\|_2^2 + \frac{a}{4} \sup_{J_1 \leq p} \sup_{\Delta \in \Gamma_{J_1}} \{\langle \epsilon, \Delta \rangle - 2\sqrt{(b/a)LP_o(J_1)}\}_+^2 \\ &\equiv \frac{1}{a}\|\mathcal{P}_{\mathcal{J}_1}X\Delta\|_2^2 + \frac{a}{4} \sup_{J_1 \leq p} R_{J_1}^2, \end{aligned}$$

where $R_{J_1} := \sup_{\Delta \in \Gamma_{J_1}} \{\langle \epsilon, \Delta \rangle - 2\sqrt{(b/a)LP_o(J_1)}\}_+$ with L a sufficiently large constant. When $J_1 = 0$, $R_{J_1} = 0$. When $J_1 \geq 1$, for any $t \geq 0$, if $4b/a$ is a constant greater than 1, we have

$$\begin{aligned} &\mathbb{P}(\sup_{1 \leq J_1 \leq p} R_{J_1} \geq t\sigma) \\ &\leq \sum_{J_1=1}^p \mathbb{P}\left(\sup_{\Delta \in \Gamma_{J_1}} \langle \epsilon, \Delta \rangle - \sqrt{LP_o(J_1)} \geq t\sigma\right. \\ &\quad \left. + 2\sqrt{\frac{b}{a}LP_o(J_1)} - \sqrt{LP_o(J_1)}\right) \\ &\leq C \exp(-ct^2) \sum_{J_1=1}^p \exp[-c(2\sqrt{b/a} - 1)^2 LP_o(J_1)/\sigma^2] \\ &\leq C \exp(-ct^2) \exp(-cL \log p) \sum_{J_1=1}^p \exp(-cLJ_1) \\ &\leq C \exp(-ct^2) p^{-cL}. \quad (\text{A.10}) \end{aligned}$$

The second inequality is due to Lemma 6 of [21], and we used $J \log(ep/J) \geq J + \log p$ for any $J \in [p]$ in the third inequality. Therefore, for any $a, b > 0$, $4b > a$ and $t \geq 0$, we have

$$\begin{aligned} &\mathbb{P}\left\{\langle \epsilon, \mathcal{P}_{\mathcal{J}_1}X\Delta \rangle - \frac{2}{a}\|\mathcal{P}_{\mathcal{J}_1}X\Delta\|_2^2 - bLP_o(J_1) \geq \frac{a}{4}t\sigma^2\right\} \\ &\leq C \exp(-ct)p^{-Lc}. \quad (\text{A.11}) \end{aligned}$$

Similarly, for the second term in (A.9), we can use Lemma 7 of [13] to prove that for any $t \geq 0$,

$$\begin{aligned} &\mathbb{P}\left[\langle \epsilon, \mathcal{P}_{\mathcal{J}_1}^\perp \mathcal{P}_{\mathcal{J}_2}X\Delta \rangle - \frac{2}{a}\|\mathcal{P}_{\mathcal{J}_1}^\perp \mathcal{P}_{\mathcal{J}_2}X\Delta\|_2^2 - bL\{P_o(J_1) + P_o(J_2)\} \right. \\ &\quad \left. \geq \frac{a}{4}t\sigma^2\right] \leq C \exp(-ct)p^{-Lc}. \quad (\text{A.12}) \end{aligned}$$

Combining (A.11), (A.12) and using the fact that $\|\mathcal{P}_{\mathcal{J}_1}X\Delta\|_2^2 + \|\mathcal{P}_{\mathcal{J}_1}^\perp \mathcal{P}_{\mathcal{J}_2}X\Delta\|_2^2 = \|X\Delta\|_2^2$, we get for any $a, b > 0$, $4b > a$ and $t \geq 0$,

$$\begin{aligned} &\mathbb{P}\left[\langle \epsilon, X\Delta \rangle - \frac{4}{a}\|X\Delta\|_2^2 - 3bL\{P_o(J_1) \vee P_o(J_2)\} \geq \frac{a}{2}t\sigma^2\right] \\ &\leq C \exp(-ct)p^{-Lc}. \quad (\text{A.13}) \end{aligned}$$

Finally, using the increasing property of $P_o(J)$ for $J \in [0, p]$, we have $P_o(J_1) \vee P_o(J_2) \leq (J_1 \vee J_2) \log\{ep/(J_1 \vee J_2)\}$. A reparameterization of (A.13) gives the conclusion.

C. Proof of Theorem 2

From the proof of Theorem 1, we get with probability $1 - Cp^{-c}$,

$$\begin{aligned} &\|X\hat{\beta} - X\beta^*\|_2^2 + \frac{\eta_0(1-b)}{2}\|\hat{\beta} - \beta^*\|_2^2 \\ &\leq \frac{\rho - (\sqrt{\vartheta} - 1)\eta_0}{2\sqrt{\vartheta}}\|\hat{\beta} - \beta^*\|_2^2 + \frac{\eta_0}{2b}\|\beta^*\|_2^2 + \\ &\quad \frac{1}{2a}\|X\hat{\beta} - X\beta^*\|_2^2 + \frac{a}{2}A\sigma^2\vartheta s \log \frac{ep}{\vartheta s}, \end{aligned}$$

which gives

$$\begin{aligned} &\|X\hat{\beta} - X\beta^*\|_2^2 - \frac{\eta_0 b}{2}\|\hat{\beta} - \beta^*\|_2^2 \\ &\leq \frac{\rho - (2\sqrt{\vartheta} - 1)\eta_0}{2\sqrt{\vartheta}}\|\hat{\beta} - \beta^*\|_2^2 + \frac{\eta_0}{2b}\|\beta^*\|_2^2 + \\ &\quad \frac{\rho_+((1+\vartheta)s)}{2a}\|\hat{\beta} - \beta^*\|_2^2 + \frac{a}{2}A\sigma^2\vartheta s \log \frac{ep}{\vartheta s}. \end{aligned}$$

Under the regularity condition (9), choosing $a = 2/\delta$ and $b = \delta\rho_+((1+\vartheta)s)/(4\eta_0)$ give (10). (The result applies to $\eta_0 = 0$ as well.)

To show the second result, note that from Theorem 1, the fixed-point solution $\hat{\beta}$ must satisfy $\hat{\beta} = \Theta^\# \{\hat{\beta} - X^T \nabla l_0(X\hat{\beta}; y)/\rho; q, \eta_0/\rho\}$, which means

$$\begin{aligned} &\left\|\hat{\beta}(1 + \frac{\eta_0}{\rho}) - \hat{\beta} + \frac{1}{\rho}X^T \nabla l_0(X\hat{\beta})\right\|_\infty \leq (1 + \frac{\eta_0}{\rho}) \min_{j \in \hat{\mathcal{J}}} |\hat{\beta}_j| \\ &\implies \left\|\eta_0\hat{\beta} + X^T(\nabla l_0(X\hat{\beta}) - \nabla l_0(X\beta^*)) - X^T\epsilon\right\|_\infty \\ &\leq (\rho + \eta_0) \min_{j \in \hat{\mathcal{J}}} |\hat{\beta}_j| \\ &\implies \left\|X^T(\nabla l_0(X\hat{\beta}) - \nabla l_0(X\beta^*)) + \eta_0(\hat{\beta} - \beta^*)\right\|_\infty \\ &\leq \|X^T\epsilon\|_\infty + \eta_0\|\beta^*\|_\infty + (\rho + \eta_0) \min_{j \in \hat{\mathcal{J}}} |\hat{\beta}_j|. \end{aligned}$$

By definition, $\nabla l_0(X\hat{\beta}) - \nabla l_0(X\beta^*) = X(\hat{\beta} - \beta^*)$.

Next, we introduce a lemma.

Lemma A.3. Let $\tilde{\beta}, \beta \in \mathbb{R}^p$ satisfying $\|\tilde{\beta}\|_0 = q > s \geq \|\beta\|_0$, and for short, denote $\mathcal{J}(\tilde{\beta})$ and $\mathcal{J}(\tilde{\beta})$ by $\tilde{\mathcal{J}}$ and \mathcal{J} , respectively. Then

$$\begin{aligned} \min_{j \in \tilde{\mathcal{J}}} |\tilde{\beta}_j| &\leq \min_{j \in \tilde{\mathcal{J}} \setminus \mathcal{J}} |\tilde{\beta}_j| \leq \frac{\|(\tilde{\beta} - \beta)_{\tilde{\mathcal{J}} \setminus \mathcal{J}}\|_2}{\sqrt{|\tilde{\mathcal{J}} \setminus \mathcal{J}|}} \\ &\leq \frac{\|(\tilde{\beta} - \beta)_{\tilde{\mathcal{J}} \setminus \mathcal{J}}\|_2}{\sqrt{q-s}} \leq \frac{\|\tilde{\beta} - \beta\|_2}{\sqrt{q-s}} \quad (\text{A.14}) \end{aligned}$$

$$\min_{j \in \tilde{\mathcal{J}}} |\tilde{\beta}_j| \leq \max_{j \in \tilde{\mathcal{J}} \setminus \mathcal{J}} |\tilde{\beta}_j| = \|(\tilde{\beta} - \beta)_{\tilde{\mathcal{J}} \setminus \mathcal{J}}\|_\infty \leq \|\tilde{\beta} - \beta\|_\infty. \quad (\text{A.15})$$

The proof is simple and omitted. Now, combining the regularity condition (11) and (A.14) or (A.15) gives the desired result.

D. Proof of Theorem 3

By definition, we have

$$\rho_+(2q) = \sup_{I \in [p]: |I|=2q} \lambda_{\max}(X_I^T X_I),$$

and under $q+s \leq n$,

$$\rho_-(q+s) = \inf_{I \in [p]: |I|=q+s} \lambda_{\min}(X_I^T X_I).$$

By Theorem of 6.1 of [23], we have

$$\begin{aligned} & \mathbb{P} \left\{ \sqrt{\frac{\lambda_{\max}(X_I^T X_I)}{n}} \geq (1+c_0) \sqrt{\lambda_{\max}(\Sigma_I)} + \sqrt{\frac{\text{tr}(\Sigma_I)}{n}} \right\} \\ & \leq \exp(-nc_0^2/2), \quad \forall I : |I|=2q \end{aligned}$$

and

$$\begin{aligned} & \mathbb{P} \left\{ \sqrt{\frac{\lambda_{\min}(X_I^T X_I)}{n}} \leq (1-c_0) \sqrt{\lambda_{\min}(\Sigma_I)} - \sqrt{\frac{\text{tr}(\Sigma_I)}{n}} \right\} \\ & \leq \exp(-nc_0^2/2), \quad \forall I : |I|=q+s \end{aligned}$$

for all $c_0 > 0$. Applying the union bound gives

$$\begin{aligned} & \mathbb{P} \left\{ \sqrt{\frac{\rho_+(2q)}{n}} \geq (1+c_0) \sqrt{\lambda_{\max}^{(2q)}} + \sqrt{\frac{2q}{n}} \right\} \\ & \leq \binom{p}{2q} \exp(-nc_0^2/2). \end{aligned} \quad (\text{A.16})$$

Let $nc^2 = nc_0^2 - \log \binom{p}{2q}$. Then using $\log \binom{p}{2q} \leq 2q \log(ep/q)$, $c_0 \leq c + \sqrt{2q \log(ep/q)/n}$. Therefore for any $c > 0$,

$$\begin{aligned} & \mathbb{P} \left\{ \sqrt{\frac{\rho_+(2q)}{n}} \geq (1+c) \sqrt{\lambda_{\max}^{(2q)}} \right. \\ & \left. + \sqrt{\frac{2q \log(ep/q)}{n}} \sqrt{\lambda_{\max}^{(2q)}} + \sqrt{\frac{2q}{n}} \right\} \leq \exp(-nc^2/2). \end{aligned} \quad (\text{A.17})$$

Similarly,

$$\begin{aligned} & \mathbb{P} \left\{ \sqrt{\frac{\rho_-(q+s)}{n}} \leq (1-c) \sqrt{\lambda_{\min}^{(q+s)}} \right. \\ & \left. - \sqrt{\frac{(q+s) \log(ep/q)}{n}} \sqrt{\lambda_{\min}^{(q+s)}} - \sqrt{\frac{q+s}{n}} \right\} \leq \exp(-nc^2/2). \end{aligned}$$

Let $c \in (0, 1)$ and assume $n \geq \{2(q+s)/(1-c)^2\} \{1/\lambda_{\min}^{(q+s)} + \log(ep/q)\}$. Then

$$\begin{aligned} & \frac{\rho_+(2q)}{\rho_-(q+s)} \leq \\ & \left\{ \frac{(1+c) \sqrt{\lambda_{\max}^{(2q)}} + \sqrt{\{2\lambda_{\max}^{(2q)} q \log(ep/q)\}/n} + \sqrt{2q/n}}{(1-c) \sqrt{\lambda_{\min}^{(q+s)}} - \sqrt{\{\lambda_{\min}^{(q+s)} (q+s) \log(ep/q)\}/n} - \sqrt{(q+s)/n}} \right\}^2 \end{aligned}$$

holds with probability at least $1 - 2 \exp(-nc^2/2)$.

E. Proof of Theorem 5

Let $E := \sigma^2 P_o(q) + \sigma^2$. Similar to the proof of Theorem 1, from the construction of g and Lemma A.1, we have

$$\rho(1 - 1/\sqrt{\vartheta})(1 + \eta_0/\rho) \mathbf{D}_2(\beta^*, \hat{\beta}) + g(\hat{\beta}, \hat{\beta}) \leq g(\beta^*, \hat{\beta}),$$

and thus

$$\begin{aligned} & 2\bar{\Delta}_{l_0}(X\hat{\beta}, X\beta^*) + \frac{\eta_0}{2} \|\hat{\beta}\|_2^2 \\ & \leq \frac{\rho - (\sqrt{\vartheta} - 1)\eta_0}{\sqrt{\vartheta}} \mathbf{D}_2(\hat{\beta}, \beta^*) + \frac{\eta_0}{2} \|\beta^*\|_2^2 + \langle \epsilon, X\hat{\beta} - X\beta^* \rangle. \end{aligned} \quad (\text{A.18})$$

Applying Lemma A.2 gives

$$\langle \epsilon, X\hat{\beta} - X\beta^* \rangle \leq \delta \mathbf{D}_2(X\hat{\beta}, X\beta^*) + \frac{1}{\delta} A \sigma^2 P_o(q) + R \quad (\text{A.19})$$

for any $\delta > 0$, where $R := \sup_{\beta_1, \beta_2} \{ \langle \epsilon, X\beta_1 - X\beta_2 \rangle - \delta \mathbf{D}_2(X\beta_1, X\beta_2) - A \sigma^2 P_o(q)/\delta \}_+$ and $\mathbb{P}(\delta R > \sigma^2 t) \leq C \exp(-ct) p^{-cA}$,

where $A, C, c > 0$ are some constants. Therefore,

$$\mathbb{E} \langle \epsilon, X\hat{\beta} - X\beta^* \rangle \leq \mathbb{E} \{ \delta \mathbf{D}_2(X\hat{\beta}, X\beta^*) \} + \frac{C}{\delta} (\sigma^2 P_o(q) + \sigma^2). \quad (\text{A.20})$$

Combining (A.18) and (A.20) gives

$$\begin{aligned} & \mathbb{E} \{ (2\bar{\Delta}_{l_0} - \delta \mathbf{D}_2)(X\hat{\beta}, X\beta^*) + \eta_0 \mathbf{D}_2(\hat{\beta}, \beta^*) \} \\ & \leq \mathbb{E} \left\{ \frac{\rho - (\sqrt{\vartheta} - 1)\eta_0}{\sqrt{\vartheta}} \mathbf{D}_2(\hat{\beta}, \beta^*) + \eta_0 \langle -\beta^*, \hat{\beta} - \beta^* \rangle \right\} + \frac{C}{\delta} E, \end{aligned} \quad (\text{A.21})$$

and so

$$\begin{aligned} & \mathbb{E} \left[(2\bar{\Delta}_{l_0} - \delta \mathbf{D}_2)(X\hat{\beta}, X\beta^*) - \frac{\rho - ((2-\varepsilon)\sqrt{\vartheta} - 1)\eta_0}{\sqrt{\vartheta}} \mathbf{D}_2(\hat{\beta}, \beta^*) \right] \\ & \leq \frac{C}{\delta} E + \frac{\eta_0}{2\varepsilon} \|\beta^*\|_2^2 \end{aligned} \quad (\text{A.22})$$

for any $\varepsilon, \delta > 0$.

Next, from $l_0(X\hat{\beta}) + \eta_0 \|\hat{\beta}\|_2^2/2 \leq l_0(X\beta^{(0)}) + \eta_0 \|\beta^{(0)}\|_2^2/2$, we have

$$\begin{aligned} & \Delta_{l_0}(X\hat{\beta}, X\beta^*) + \eta_0 \mathbf{D}_2(\hat{\beta}, \beta^*) \\ & \leq \Delta_{l_0}(X\beta^{(0)}, X\beta^*) + \eta_0 \mathbf{D}_2(\beta^{(0)}, \beta^*) + \eta_0 \langle -\beta^*, \hat{\beta} - \beta^* \rangle \\ & \quad - \eta_0 \langle -\beta^*, \beta^{(0)} - \beta^* \rangle \\ & \quad + \langle \epsilon, X\hat{\beta} - X\beta^* \rangle - \langle \epsilon, X\beta^{(0)} - X\beta^* \rangle. \end{aligned} \quad (\text{A.23})$$

Therefore, for any $\delta', \delta'', \varepsilon' > 0$

$$\begin{aligned} & \mathbb{E} \{ (\Delta_{l_0} - \delta' \mathbf{D}_2)(X\hat{\beta}, X\beta^*) + \eta_0 \mathbf{D}_2(\hat{\beta}, \beta^*) \} \\ & \leq \mathbb{E} \left\{ (\Delta_{l_0} + \delta'' \mathbf{D}_2)(X\beta^{(0)}, X\beta^*) + \eta_0 \mathbf{D}_2(\beta^{(0)}, \beta^*) \right. \\ & \quad \left. + \frac{\eta_0}{2\varepsilon} \|\beta^*\|_2^2 + \eta_0 \varepsilon \mathbf{D}_2(\hat{\beta}, \beta^*) + \frac{\eta_0}{2\varepsilon'} \|\beta^*\|_2^2 \right. \\ & \quad \left. + \eta_0 \varepsilon' \mathbf{D}_2(\beta^{(0)}, \beta^*) \right\} + CE \left(\frac{1}{\delta'} + \frac{1}{\delta''} \right). \end{aligned}$$

By the assumption of the starting point $\mathbb{E} \{ \mathbf{D}_2(\beta^{(0)}, \beta^*) \} \leq CME/n$, we have

$$\begin{aligned} & \mathbb{E} \{ \mathbf{D}_2(X\beta^{(0)}, X\beta^*) \} \leq C \rho_+(q+s) ME/n, \\ & \mathbb{E} \{ \Delta_{l_0}(X\beta^{(0)}, X\beta^*) \} \leq C \rho_+^l(q, s) ME/n. \end{aligned}$$

Taking $1/\delta'' = \sqrt{\rho_+(q+s)M/n}$, we obtain

$$\begin{aligned} & \mathbb{E}\{(\Delta_{l_0} - \delta' \mathbf{D}_2)(X\hat{\beta}, X\beta^*) + \eta_0(1-\varepsilon)\mathbf{D}_2(\hat{\beta}, \beta^*)\} \\ & \leq CE\left(\frac{1}{\delta'} + \sqrt{\frac{\rho_+(q+s)M}{n}} + \frac{\rho_+^l(q, s)}{n}M + \frac{\eta_0(1+\varepsilon')}{n}M\right) \\ & \quad + \eta_0\left(\frac{1}{\varepsilon} + \frac{1}{\varepsilon'}\right)\frac{\|\beta^*\|_2^2}{2}. \end{aligned}$$

Let $Q_0 := \sqrt{\rho_+(q+s)M/n} + \rho_+^l(q, s)M/n + \eta_0(1+\varepsilon')M/n$. Then

$$CE\left(\frac{1}{\delta'} + Q_0\right) \leq \frac{C}{c_1 \wedge c_2} E\left(\frac{c_1}{\delta'} + c_2 Q_0\right)$$

for any $c_1, c_2 > 0$. Taking $\delta' := \delta'^2/(c_1 + c_2 Q_0 \delta')$ and $\varepsilon' := 1/\varepsilon + 1/\varepsilon' = (1/\delta' + Q_0)c_3 \delta/\varepsilon$ for some large constant $c_3 > 0$, we get

$$\begin{aligned} & \mathbb{E}\left\{(\frac{\delta}{\delta'}\Delta_{l_0} - \delta \mathbf{D}_2)(X\hat{\beta}, X\beta^*) + \frac{\delta}{\delta'}\eta_0(1-\varepsilon)\mathbf{D}_2(\hat{\beta}, \beta^*)\right\} \\ & \leq \frac{CE}{c_1 \wedge c_2} \frac{1}{\delta} + c_3 \frac{\eta_0}{2\varepsilon} \|\beta^*\|_2^2. \end{aligned} \quad (\text{A.24})$$

Multiplying (A.22) by $(1-1/M)$ and (A.24) by $1/M$ and adding the two inequalities yield

$$\begin{aligned} & \mathbb{E}\left[(1-\frac{1}{M})\{2\bar{\Delta}_{l_0}(X\hat{\beta}, X\beta^*)\right. \\ & \quad \left.- \frac{\rho - \{(2-\varepsilon)\sqrt{\vartheta} - 1\}\eta_0}{\sqrt{\vartheta}}\mathbf{D}_2(\hat{\beta}, \beta^*)\}\right. \\ & \quad \left.+ (\frac{\delta}{M\delta'}\Delta_{l_0} - \delta \mathbf{D}_2)(X\hat{\beta}, X\beta^*)\right. \\ & \quad \left.+ \frac{\delta}{M\delta'}\eta_0(1-\varepsilon)\mathbf{D}_2(\hat{\beta}, \beta^*)\right] \\ & \leq C\left(\frac{E}{\delta} + \frac{\eta_0}{\varepsilon} \|\beta^*\|_2^2\right). \end{aligned} \quad (\text{A.25})$$

Simple calculation shows

$$\begin{aligned} \frac{\delta'}{\delta} &= \frac{c_2 Q_0 \delta + \sqrt{c_2^2 Q_0^2 \delta^2 + 4c_1}}{2} \\ &\leq \frac{\sqrt{2} + 1}{2} \{c_2 Q_0 \delta \vee \sqrt{4c_1}\} \leq C(Q_0 \delta \vee 1). \end{aligned}$$

It follows that

$$\varepsilon' \leq \frac{\varepsilon}{C(Q_0 \delta \vee 1) + \delta Q_0 - 1} \leq C \frac{\varepsilon}{Q_0 \delta \vee 1} \leq C\varepsilon$$

for some large constant C , and so $Q_0 \lesssim Q$. Under the condition that

$$\begin{aligned} & K\sigma^2 P_o(\vartheta s) + \left\{2\left(1-\frac{1}{M}\right)\bar{\Delta}_{l_0}\right. \\ & \quad \left.+\frac{C}{M(Q\delta \vee 1)}\Delta_{l_0} - 2\delta \mathbf{D}_2\right\}(X\hat{\beta}, X\beta^*) \\ & \geq \frac{1-1/M}{\sqrt{\vartheta}} [\rho - \{(2-\varepsilon)\sqrt{\vartheta} - 1\}\eta_0] \mathbf{D}_2(\hat{\beta}, \beta^*) \\ & \quad - \frac{C}{M(Q\delta \vee 1)} \eta_0(1-\varepsilon) \mathbf{D}_2(\hat{\beta}, \beta^*), \end{aligned} \quad (\text{A.26})$$

(A.25) yields

$$\begin{aligned} \mathbb{E}[\mathbf{D}_2(X\hat{\beta}, X\beta^*)] &\leq \frac{K}{\delta}\sigma^2 P_o(\vartheta s) + \frac{CE}{\delta^2} + C\frac{\eta_0}{\varepsilon} \|\beta^*\|_2^2 \\ &\lesssim \frac{K\delta \vee 1}{\delta^2} E + \frac{\eta_0}{\delta\varepsilon} \|\beta^*\|_2^2. \end{aligned} \quad (\text{A.27})$$

With a reparameterization, the regularity condition (30) implies (A.26).

F. Proof of Theorem 6

For convenience, denote $\mathbf{D}_2(X\beta, X\beta')$ by $\mathbf{D}_{2,X}(\beta, \beta')$. From Lemma A.1, we have

$$\begin{aligned} & g(\beta^*, \beta^{(t)}) - g(\beta^{(t+1)}; \beta^{(t)}) \\ & \geq \rho_{t+1}(1 - \mathcal{L}_{t+1})(1 + \bar{\eta}_{t+1})\mathbf{D}_2(\beta^{(t+1)}, \beta^*), \end{aligned} \quad (\text{A.28})$$

where $\mathcal{L}_{t+1} = \mathcal{L}(\mathcal{J}(\beta^*), \mathcal{J}(\beta^{(t+1)})) \leq 1/\sqrt{\vartheta_{t+1}}$. (Recall $\vartheta_{t+1} = q_{t+1}/s > 1$, and $s \geq \|\beta^*\|_0$.)

Substituting $g(\beta, \beta^{(t)}) = l(\beta) + \eta_{t+1}\mathbf{D}_2(\beta, 0) + (\rho_{t+1}\mathbf{D}_2 - \Delta_l)(\beta, \beta^{(t)})$ and $l(\beta^*) - l(\beta^{(t+1)}) = \langle \epsilon, X\beta^{(t+1)} - X\beta^* \rangle - \Delta_l(\beta^*, \beta^{(t+1)})$ into (A.28) gives

$$\begin{aligned} & \{\rho_{t+1}(1 - \mathcal{L}_{t+1})(1 + \bar{\eta}_{t+1})\mathbf{D}_2 + \Delta_l\}(\beta^*, \beta^{(t+1)}) \\ & \quad + \eta_{t+1}\mathbf{D}_2(\beta^*, \beta^{(t+1)}) + (\rho_{t+1}\mathbf{D}_2 - \Delta_l)(\beta^{(t+1)}, \beta^{(t)}) \\ & \leq (\rho_{t+1}\mathbf{D}_2 - \Delta_l)(\beta^*, \beta^{(t)}) + \langle \epsilon, X\beta^{(t+1)} - X\beta^* \rangle \\ & \quad + \eta_{t+1}\langle -\beta^*, \beta^{(t+1)} - \beta^* \rangle. \end{aligned} \quad (\text{A.29})$$

From Lemma A.2, with probability at least $1 - Cp^{-cA}$

$$\begin{aligned} & \langle \epsilon, X\beta^{(t+1)} - X\beta^* \rangle \\ & \leq \delta_{t+1}\mathbf{D}_{2,X}(\beta^*, \beta^{(t+1)}) + \delta_{t+1}^{-1}A\sigma^2 P_o(q_{t+1}), \text{ for all } t \geq 0 \end{aligned} \quad (\text{A.30})$$

given any $\delta_{t+1} > 0$, where A is a constant. Moreover, for any $\varepsilon_{t+1} > 0$,

$$\langle -\beta^*, \beta^{(t+1)} - \beta^* \rangle \leq \varepsilon_{t+1}\mathbf{D}_2(\beta^*, \beta^{(t+1)}) + \varepsilon_{t+1}^{-1}\mathbf{D}_2(\beta^*, 0). \quad (\text{A.31})$$

Plugging these bounds into (A.29) gives

$$\begin{aligned} & \{\rho_{t+1}(1 - \mathcal{L}_{t+1})(1 + \bar{\eta}_{t+1})\mathbf{D}_2 + \Delta_l + (1 - \varepsilon_{t+1})\eta_{t+1}\mathbf{D}_2 \\ & \quad - \delta_{t+1}\mathbf{D}_{2,X}\}(\beta^*, \beta^{(t+1)}) + (\rho_{t+1}\mathbf{D}_2 - \Delta_l)(\beta^{(t+1)}, \beta^{(t)}) \\ & \leq (\rho_{t+1}\mathbf{D}_2 - \Delta_l)(\beta^*, \beta^{(t)}) + \delta_{t+1}^{-1}A\sigma^2 P_o(q_{t+1}) \\ & \quad + \varepsilon_{t+1}^{-1}\eta_{t+1}\mathbf{D}_2(\beta^*, 0). \end{aligned} \quad (\text{A.32})$$

By the definition of (generalized) isometry numbers and using $\mathcal{L}_{t+1} \leq 1/\sqrt{\vartheta_{t+1}}$, we have

$$\begin{aligned} & \left\{\rho_{t+1}\left(1 - \frac{1}{\sqrt{\vartheta_{t+1}}}\right)(1 + \bar{\eta}_{t+1}) + \rho_-^l(q_{t+1}, s)\right. \\ & \quad \left.+ (1 - \varepsilon_{t+1})\eta_{t+1} - \delta_{t+1}\rho_+(q_{t+1} + s)\right\}\mathbf{D}_2(\beta^*, \beta^{(t+1)}) \\ & \quad + (\rho_{t+1}\mathbf{D}_2 - \Delta_l)(\beta^{(t+1)}, \beta^{(t)}) \\ & \leq \{\rho_{t+1} - \rho_-^l(s, q_{t+1})\}\mathbf{D}_2(\beta^*, \beta^{(t)}) + \delta_{t+1}^{-1}A\sigma^2 P_o(q_{t+1}) \\ & \quad + \varepsilon_{t+1}^{-1}\eta_{t+1}\mathbf{D}_2(\beta^*, 0). \end{aligned} \quad (\text{A.33})$$

Let ε_0 be any number $\in (0, 1]$. Taking $\varepsilon_{t+1} = \varepsilon_0/2, \delta_{t+1} = (\varepsilon_0\rho_-^l(q_{t+1}, s) + \varepsilon_0\eta_{t+1}/2)/\rho_+(q_{t+1} + s)$, we have

$$\begin{aligned} & (1 - 1/\sqrt{\vartheta_{t+1}})(1 + \bar{\eta}_{t+1})\rho_{t+1} + \rho_-^l(q_{t+1}, s) \\ & \quad + (1 - \varepsilon_{t+1})\eta_{t+1} - \delta_{t+1}\rho_+(q_{t+1} + s) \\ & = (1 - 1/\sqrt{\vartheta_{t+1}})(1 + \bar{\eta}_{t+1})\rho_{t+1} + (1 - \varepsilon_0)\rho_-^l(q_{t+1}, s) \\ & \quad + (1 - \varepsilon_0)\eta_{t+1}. \end{aligned}$$

Let

$$\begin{aligned}
E_{t+1} &= \frac{1}{\rho_{t+1} - \rho_-^l(s, q_{t+1})} \times \\
&\left\{ \frac{A\sigma^2}{\varepsilon_0} \frac{\rho_+(q_{t+1} + s)}{\rho_-^l(q_{t+1}, s) + \eta_{t+1}/2} P_o(q_{t+1}) + \frac{\eta_{t+1}}{\varepsilon_0} \|\beta^*\|_2^2 \right\} \\
&\leq \frac{A\sigma^2}{\varepsilon_0} \times \\
&\frac{\rho_+(q_{t+1} + s)}{(\rho_-^l(q_{t+1}, s)/\rho_{t+1} \vee \bar{\eta}_{t+1})(1 - \rho_-^l(s, q_{t+1})/\rho_{t+1})\rho_{t+1}^2} P_o(q_{t+1}) \\
&+ \frac{\bar{\eta}_{t+1}}{\varepsilon_0(1 - \rho_-^l(s, q_{t+1})/\rho_{t+1})} \|\beta^*\|_2^2
\end{aligned}$$

for any $t \geq 0$. By the definitions of κ_t, h_t , we can obtain

$$\begin{aligned}
&\mathbf{D}_2(\beta^*, \beta^{(t+1)}) + h_{t+1}(\rho_{t+1}\mathbf{D}_2 - \Delta_l)(\beta^{(t+1)}, \beta^{(t)}) \\
&\leq \kappa_{t+1}\mathbf{D}_2(\beta^*, \beta^{(t)}) + \kappa_{t+1}E_{t+1}.
\end{aligned} \tag{A.34}$$

Applying a recursive argument with $t = T, \dots, 0$ gives

$$\begin{aligned}
&\mathbf{D}_2(\beta^*, \beta^{(T+1)}) + \sum_{t=0}^T (\Pi_{\tau=t}^T h_{\tau+1})(\rho_{t+1}\mathbf{D}_2 - \Delta_l)(\beta^{(t+1)}, \beta^{(t)}) \\
&\leq \left(\Pi_{t=0}^T \kappa_{t+1} \right) \mathbf{D}_2(\beta^*, \beta^{(0)}) + \sum_{t=0}^T \left(\Pi_{\tau=t}^T \kappa_{\tau+1} \right) E_{t+1},
\end{aligned}$$

and thus the bound (35) follows.

To ensure

$$\frac{\rho_t - \rho_-^l(s, q_t)}{(1 - 1/\sqrt{\vartheta_t})(1 + \bar{\eta}_t)\rho_t + (1 - \varepsilon)(\rho_-^l(q_t, s) + \eta_t)} \leq \frac{1}{1 + \alpha} \tag{A.35}$$

for some $\alpha > 0$, we need

$$\bar{\eta}_t \geq \frac{(\alpha + 1/\sqrt{\vartheta_t}) - (2 + \alpha - \varepsilon)\{\rho_-^l(s, q_t) \wedge \rho_-^l(q_t, s)\}/\rho_t}{2 - 1/\sqrt{\vartheta_t} - \varepsilon}. \tag{A.36}$$

The result in the corollary follows by taking $\alpha = \varepsilon$ and noticing that $\rho_{t+1} \geq \rho_+^l(q_{t+1}, q_t)$ implies $(\rho_{t+1}\mathbf{D}_2 - \Delta_l)(\beta^{t+1}, \beta^t) \geq 0$.

G. A recursive coordinatewise error bound under restricted isometry

Recall the general procedure defined in (32),

$$\begin{aligned}
\beta^{(t+1)} &= \Theta^\# \left\{ \beta^{(t)} - \rho_{t+1}^{-1} X^T \nabla l_0(X\beta^{(t)}; y); q_{t+1}, \bar{\eta}_{t+1} \right\}, \\
&\text{with } \bar{\eta}_{t+1} = \eta_{t+1}/\rho_{t+1}.
\end{aligned} \tag{A.37}$$

Following a similar approach to Theorem 2 for the set of fixed points, an error bound for $\beta^{(t+1)}$ in the ∞ -norm can be established under appropriate regularity conditions.

To facilitate the proof, we first recall the definition of $\rho_-^l(s_1, s_2)$ as given in (21). In particular, in the regression setup, $\rho_-(s_1, s_2)$ satisfies $\|X(\beta_1 - \beta_2)\|_2^2 \geq \rho_-(s_1, s_2)\|\beta_1 - \beta_2\|_2^2, \forall \beta_i : \|\beta_i\|_0 \leq s_i$, or equivalently,

$$\begin{aligned}
&(\beta_1 - \beta_2)^T (\rho I - X^T X) (\beta_1 - \beta_2) \\
&\leq (\rho - \rho_-(s_1, s_2)) \|\beta_1 - \beta_2\|_2^2, \forall \beta_i : \|\beta_i\|_0 \leq s_i.
\end{aligned}$$

The presence of positive restricted eigenvalues in the Gram matrix $X^T X$ implies the existence of proper upper bounds on the restricted eigenvalues of the matrix $\rho I - X^T X$. So

when considering the ∞ -norm error for $\beta^{(t+1)}$, it appears more manageable to work with the matrix $\rho I - X^T X$ than with $X^T X$.

Motivated by this, given l , X , and s_i , we introduce a generalized restricted isometry number $v(s_1, s_2)$ that satisfies

$$\begin{aligned}
&\|\rho(\beta_1 - \beta_2) - X^T \{\nabla l_0(X\beta_1) - \nabla l_0(X\beta_2)\}\|_\infty \\
&\leq (\rho - v) \|\beta_1 - \beta_2\|_\infty \forall \beta_i : \|\beta_i\|_0 \leq s_i, \rho \geq v.
\end{aligned} \tag{A.38}$$

In the case where $l_0(X\beta) = \|X\beta - y\|_2^2/2$, we have $\nabla l_0(X\beta_1) - \nabla l_0(X\beta_2) = X(\beta_1 - \beta_2)$ and $\rho(\beta_1 - \beta_2) - X^T(\nabla l_0(X\beta_1) - \nabla l_0(X\beta_2)) = (\rho I - X^T X)(\beta_1 - \beta_2)$. Therefore, (A.38) can be understood as a variant of low coherence for the design matrix in the context of the ∞ -norm.

Theorem A.2. *For the sequence of iterates generated by procedure (A.37) and v_t denoting $v(q_t, s)$ as defined by (A.38), the following recursive coordinatewise error bound on $\beta^{(t+1)}$ holds for any $t \geq 0$:*

$$\begin{aligned}
&\|\beta^{(t+1)} - \beta^*\|_\infty \\
&\leq (1 - \frac{v_t + \eta_{t+1}}{\rho_{t+1} + \eta_{t+1}}) \|\beta^{(t)} - \beta^*\|_\infty + \frac{\|X^T \epsilon\|_\infty}{\rho_{t+1} + \eta_{t+1}} \\
&\quad + \frac{\eta_{t+1} \|\beta^*\|_\infty}{\rho_{t+1} + \eta_{t+1}} + \frac{1}{\sqrt{\vartheta_{t+1} - 1}} \frac{\|\beta^{(t+1)} - \beta^*\|_2}{\sqrt{s}}.
\end{aligned}$$

Proof. The proof follows similar lines of the proof of Theorem 2. First, by the definition of $\Theta^\#$,

$$\begin{aligned}
&\left\| (1 + \bar{\eta}_{t+1})\beta^{(t+1)} - \beta^{(t)} + \frac{1}{\rho_{t+1}} X^T \nabla l_0(X\beta^{(t)}) \right\|_\infty \\
&\leq (1 + \bar{\eta}_{t+1}) \min_{j \in \mathcal{J}(\beta_j^{(t+1)})} |\beta_j^{(t+1)}|
\end{aligned}$$

and so

$$\begin{aligned}
&\|(\rho_{t+1} + \eta_{t+1})\beta^{(t+1)} - \rho_{t+1}\beta^{(t)} \\
&\quad + X^T(\nabla l_0(X\beta^{(t)}) - \nabla l_0(X\beta^*)) - X^T \epsilon\|_\infty \\
&\leq (\rho_{t+1} + \eta_{t+1}) \min_{j \in \mathcal{J}(\beta_j^{(t+1)})} |\beta_j^{(t+1)}|.
\end{aligned}$$

Writing

$$\begin{aligned}
&(\rho_{t+1} + \eta_{t+1})\beta^{(t+1)} - \rho_{t+1}\beta^{(t)} \\
&= (\rho_{t+1} + \eta_{t+1})(\beta^{(t+1)} - \beta^*) - \rho_{t+1}(\beta^{(t)} - \beta^*) + \eta_{t+1}\beta^*
\end{aligned}$$

and using the sub-additivity of the ∞ -norm, we get

$$\begin{aligned}
&(\rho_{t+1} + \eta_{t+1})\|\beta^{(t+1)} - \beta^*\|_\infty \\
&\leq \|\rho_{t+1}(\beta^{(t)} - \beta^*) - X^T(\nabla l_0(X\beta^{(t)}) - \nabla l_0(X\beta^*))\|_\infty \\
&\quad + \|X^T \epsilon\|_\infty + \eta_{t+1} \|\beta^*\|_\infty + (\rho_{t+1} + \eta_{t+1}) \min_{j \in \mathcal{J}(\beta_j^{(t+1)})} |\beta_j^{(t+1)}|.
\end{aligned}$$

By (A.14) of Lemma A.3 and the definition of v_t , we get

$$\begin{aligned}
&(\rho_{t+1} + \eta_{t+1})\|\beta^{(t+1)} - \beta^*\|_\infty \\
&\leq (\rho_{t+1} - v_t)\|\beta^{(t)} - \beta^*\|_\infty + \|X^T \epsilon\|_\infty \\
&\quad + \eta_{t+1} \|\beta^*\|_\infty + (\rho_{t+1} + \eta_{t+1}) \frac{\|\beta^{(t+1)} - \beta^*\|_2}{\sqrt{\vartheta_{t+1} - 1}}.
\end{aligned}$$

Additionally, we can obtain $(\rho_{t+1} + \eta_{t+1})\|(\beta^{(t+1)} - \beta^*)_{\mathcal{J}^*}\|_{\infty} \leq (\rho_{t+1} - v_t)\|\beta^{(t)} - \beta^*\|_{\infty} + \|X^T\epsilon\|_{\infty} + \eta_{t+1}\|\beta^*\|_{\infty}$ or

$$\begin{aligned} & \|(\beta^{(t+1)} - \beta^*)_{\mathcal{J}^*}\|_{\infty} \\ & \leq (1 - \frac{v_t + \eta_{t+1}}{\rho_{t+1} + \eta_{t+1}})\|\beta^{(t)} - \beta^*\|_{\infty} \\ & \quad + \frac{\|X^T\epsilon\|_{\infty}}{\rho_{t+1} + \eta_{t+1}} + \frac{\eta_{t+1}\|\beta^*\|_{\infty}}{\rho_{t+1} + \eta_{t+1}}, \end{aligned}$$

by applying (A.15). \square

H. Model selection by predictive information criterion

Although parameter q as an upper bound of the true model support size can often be directly specified based on domain knowledge, this section develops a new information criterion for the tuning of q to achieve the best prediction performance in finite samples. We assume *multiple responses* to cover the application in Section IV-D. Let $Y \in \mathbb{R}^{n \times m}$, $X \in \mathbb{R}^{n \times p}$ be the response matrix and predictor matrix, respectively, and $l_0(XB; Y)$ be the given loss. We use $\mathcal{J}(B)$ to denote the row support of B and define $J(B) = |\mathcal{J}(B)|$. Assume the true $B^* \in \mathbb{R}^{p \times m}$ is row-sparse and let $s^* = J(B^*)$. The problem considered in the main sections corresponds to the special case $m = 1$. To choose the best (row) support size, we advocate the following complexity penalty to be added to the loss in the predictive information criterion:

$$P(B) = J(B)m + J(B)\log\{ep/J(B)\}. \quad (\text{A.39})$$

Recall $\mathbf{D}_2(A_1, A_2) = \|A_1 - A_2\|_F^2/2$ in the matrix context.

Theorem A.3. *Let the effective noise $E = -\nabla l_0(XB^*)$ be sub-Gaussian with mean zero and scale bounded by a constant and $B^* \in \mathcal{M}$ and $B^* \neq 0$. Assume that there exist constants $\delta > 0$ and $A_0 \geq 0$ such that $(\Delta_{l_0} - \delta\mathbf{D}_2)(XB, XB') + A_0(P(B) + P(B')) \geq 0$, for all $B, B' \in \mathcal{M}$. Then for a sufficiently large constant A , any \hat{B} that minimizes*

$$l_0(XB; Y) + AP(B) \quad (\text{A.40})$$

subject to $B \in \mathcal{M}$ must satisfy

$$\mathbb{E}\{\|X\hat{B} - XB^*\|_F^2 \vee P(\hat{B})\} \lesssim ms^* + s^*\log(ep/s^*). \quad (\text{A.41})$$

Theorem A.3 does not involve any regularization parameters (like q, λ), but it achieves the minimax optimal error rate (A.41). Moreover, the justification of (A.40) does not require an infinite-sample-size, design coherence or signal-to-noise ratio conditions.

When the noise distribution has a dispersion parameter σ^2 , Theorem A.3 still applies, but the penalty in (A.40) becomes $A\sigma^2P(B)$ with an unknown factor. A preliminary scale estimate can be possibly used. But an appealing result for regression is that the estimation of σ can be bypassed. We give a scale-free form of predictive information criterion by

$$mn\log\{\|Y - XB\|_F^2\} + AP(B), \quad (\text{A.42})$$

where A is an absolute constant.

Theorem A.4. *Let $Y = XB^* + \mathcal{E}$, where $E = [\epsilon_{i,k}]$ has independent centered sub-Gaussian(σ^2) entries and $\mathbb{E}\epsilon_{i,k}^2 \gtrsim \sigma^2$*

with σ^2 unknown. Define $l_0(XB; Y) = \|XB - Y\|_F^2$. Assume the true model is not over-complex in the sense that $P(B^*) \leq mn/A_0$ for some constant $A_0 > 0$. Let $\delta(B) = AP(B)/(mn)$, where A is a positive constant satisfying $A < A_0$, and so $\delta(B^*) < 1$. Then, for sufficiently large values of A_0 and A , any \hat{B} that minimizes $\log l_0(XB; Y) + \delta(B)$ subject to $\delta(B) < 1$ must satisfy $\mathbf{D}_2(X\hat{B}, XB^*) \lesssim \sigma^2\{s^*m + s^*\log(ep/s^*)\}$ with probability at least $1 - Cp^{-c}\exp\{-cm\} - C\exp(-cmn)$ for some constants $C, c > 0$.

A more general form of $AP(B)$ can be expressed as “ $\alpha_1 \times$ degrees-of-freedom + $\alpha_2 \times$ inflation” with α_1, α_2 as absolute constants. The two theorems can be proved based on modifying the proofs of Theorems 2 and 3 in [28]. For completeness, we present some details below. Note that although the logarithmic form of the scale-free predictive information criterion is widely used, other non-asymptotic forms exist [28]. In fact, a key trick in the proof is to convert these forms into a fractional scale-free predictive information criterion, which is essential for establishing the desired properties.

Proof. We first prove Theorem A.3 under the assumption that $\text{vec}(\mathcal{E})$ is subGaussian with mean 0 and scale σ . From the definition of \hat{B} , $\Delta_{l_0}(X\hat{B}, XB^*) + A\sigma^2P(\hat{B}) \leq A\sigma^2P(B^*) + \langle \mathcal{E}, X\hat{B} - XB^* \rangle$. Similar to the proof of Lemma A.2, we can show that for any $a, b, a' > 0$, $4b > a$, and $t > 0$,

$$\begin{aligned} \langle \mathcal{E}, XB - XB^* \rangle & \leq (\frac{2}{a} + \frac{2}{a'})\mathbf{D}_2(XB, XB^*) + a'\sigma^2t \\ & \quad + 4bL\sigma^2\{P(B^*) + P(B)\}, \forall B \in \mathbb{R}^{p \times m} \end{aligned} \quad (\text{A.43})$$

occurs with probability at least $1 - Cp^{-c}\exp(-cm)\exp(-ct)$, where L, c, C are positive constants. (The probability bound can be derived by setting L to a sufficiently large constant and observing that $Jm + J\log(ep/J) \geq m + \log(ep)$ holds for $J \geq 1$, and the union bound calculation, as in (A.10), does not need to cover the case $J = 0$.)

Now, substituting \hat{B} for B in (A.43) and taking the expectation, we have for any $a, b, a' > 0$, $4b > a$,

$$\begin{aligned} & \mathbb{E}\{\Delta_{l_0}(X\hat{B}, XB^*) + A\sigma^2P(\hat{B})\} \\ & \leq \mathbb{E}\left\{A\sigma^2P(B^*) + (\frac{2}{a} + \frac{2}{a'})\mathbf{D}_2(X\hat{B}, XB^*)\right. \\ & \quad \left.+ ca'\sigma^2 + 4bL\sigma^2[P(B^*) + P(\hat{B})]\right\}. \end{aligned}$$

Combining it with the regularity condition gives

$$\begin{aligned} & \mathbb{E}\left\{(\delta - \frac{2}{a} - \frac{2}{a'})\mathbf{D}_2(X\hat{B}, XB^*) + (A - 4bL - C)P(\hat{B})\right\} \\ & \leq (A + 4bL + C)\sigma^2P(B^*) + ca'\sigma^2. \end{aligned}$$

Since $P(B^*) \geq c > 0$, choosing the constants satisfying $(1/a + 1/a')(1 + 1/b') < \delta/2$, $4b > a$, and $A > 4bL + C$ yields the conclusion.

Next, we prove Theorem A.4. We begin with a proof for \hat{B} selected by a fractional form of scale-free form of predictive information criterion: $l_0(XB; Y)/(1 - \delta(B))$ subject to $\delta(B) \leq 1$. Let $h(B; A) = 1/\{mn - AP(B)\}$. From the optimality of \hat{B} , $l_0(X\hat{B}; Y)h(\hat{B}; A) \leq l_0(XB^*; Y)h(B^*; A)$ or

$$l_0(X\hat{B}; Y) - l_0(XB^*; Y) \leq l_0(XB^*; Y)\left(\frac{h(B^*; A)}{h(\hat{B}; A)} - 1\right),$$

where we used $h(\hat{B}; A) > 0$. Using the Bregman divergence for the quadratic function, we get

$$\begin{aligned} & \mathbf{D}_2(X\hat{B}, XB^*) \\ & \leq l_0(XB^*; Y) \left(\frac{h(B^*; A)}{h(\hat{B}; A)} - 1 \right) + \langle \mathcal{E}, X\hat{B} - XB^* \rangle. \end{aligned} \quad (\text{A.44})$$

From the definition of h and the model parsimony assumption, (A.44) becomes

$$\begin{aligned} & \mathbf{D}_2(X\hat{B}, XB^*) \\ & \leq l_0(XB^*; Y) \frac{AP(B^*) - AP(\hat{B})}{mn - AP(B^*)} + \langle \mathcal{E}, X\hat{B} - XB^* \rangle \\ & = \frac{1}{2} \frac{A\|\mathcal{E}\|_F^2}{mn\sigma^2 - A\sigma^2 P(B^*)} \sigma^2 P(B^*) - \frac{1}{2} \frac{A\|\mathcal{E}\|_F^2}{mn - AP(B^*)} \sigma^2 P(\hat{B}) \\ & \quad + \langle \mathcal{E}, X\hat{B} - XB^* \rangle \\ & \leq \frac{1}{2} \frac{A\|\mathcal{E}\|_F^2}{(1 - A/A_0)mn\sigma^2} \sigma^2 P(B^*) - \frac{1}{2} \frac{A\|\mathcal{E}\|_F^2}{mn\sigma^2} \sigma^2 P(\hat{B}) \\ & \quad + \langle \mathcal{E}, X\hat{B} - XB^* \rangle. \end{aligned} \quad (\text{A.45})$$

The stochastic term $\langle \mathcal{E}, X\hat{B} - XB^* \rangle$ can be bounded similarly by (A.43): for any $a_1, b_1, a_2 > 0$ satisfying $4b_1 > a_1$,

$$\begin{aligned} \langle \mathcal{E}, X\hat{B} - XB^* \rangle & \leq 2(1/a_1 + 1/a_2) \mathbf{D}_2(X\hat{B}, XB^*) \\ & \quad + (b_1)L_1\sigma^2\{P(\hat{B}) + P(B^*)\}, \end{aligned}$$

with probability at least $1 - Cp^{-c} \exp\{-cm\}$ for some $c, C, L_1 > 0$. Plugging it into (A.45) gives

$$\begin{aligned} & \left(1 - \frac{2}{a_1} - \frac{2}{a_2}\right) \mathbf{D}_2(X\hat{B}, XB^*) \\ & \leq \frac{1}{2} \left\{ \frac{A\|\mathcal{E}\|_F^2}{(1 - A/A_0)mn\sigma^2} + 2b_1L_1 \right\} \sigma^2 P(B^*) \\ & \quad - \frac{1}{2} \left\{ \frac{A\|\mathcal{E}\|_F^2}{mn\sigma^2} - 2b_1L_1 \right\} \sigma^2 P(\hat{B}). \end{aligned}$$

Since $\epsilon_{i,k}$ are independent and non-degenerate, $c_1mn\sigma^2 \leq \mathbb{E}\|\mathcal{E}\|_F^2 \leq c_2mn\sigma^2$ for some constants $c_1, c_2 > 0$. Let γ be some constant satisfying $0 < \gamma < 1$. On $\mathcal{E} = \{c_1(1 - \gamma)mn\sigma^2 \leq \|\mathcal{E}\|_F^2 \leq c_2(1 + \gamma)mn\sigma^2\}$, we have

$$\frac{A\|\mathcal{E}\|_F^2}{(1 - A/A_0)mn\sigma^2} \leq \frac{c_2(1 + \gamma)A_0A}{A_0 - A}, \quad \frac{A\|\mathcal{E}\|_F^2}{mn\sigma^2} \geq c_1(1 - \gamma)A.$$

Regarding the probability of the event, we write $\|\mathcal{E}\|_F^2 = \text{vec}(\mathcal{E})A\text{vec}(\mathcal{E})^T$ with $A = I \in \mathbb{R}^{nm \times nm}$ and bound it with the Hanson-Wright inequality. In fact, from $\text{Tr}(A) = mn$, $\|A\|_2 = 1$, $\|A\|_F = \sqrt{mn}$, the complement of \mathcal{E} occurs with probability at most $C' \exp\{-c'mn\}$.

Now, with A_0, A, a_1, a_2, b_1 large enough such that $(1/a_1 + 1/a_2) < 1/2$, $4b_1 > a_1$, $A > 2b_1L_1/\{c_1(1 - \gamma)\}$ and $A_0 > A$, we can obtain the desired prediction error rate for the fractional form. Finally, based on the fact that $1/(1 - \delta) \geq \exp(\delta) \geq 1/(1 - \delta/2)$ for any $0 \leq \delta < 1$, the same error rate holds for the logarithmic form (see [28] for more details). \square

I. More implementation details

Slow kill is extremely simple to implement and a summary is given below. For ease of presentation, we define an $\bar{\eta}$ function based on Theorem 6 and its discussions,

$$\bar{\eta}(q_+, \rho_+) = \begin{cases} \frac{1}{2\sqrt{q_+/\bar{s}-1}}, & \text{if } q_+ > 2q \text{ and } q \geq n/2 \\ \frac{\eta_0}{\rho_+}, & \text{if } q_+ \leq 2q, \\ \frac{\eta_0}{\rho_+} \wedge \frac{1}{2\sqrt{q_+/\bar{s}-1}}, & \text{otherwise,} \end{cases} \quad (\text{A.46})$$

where $\bar{s} = q \wedge nL^2/\log(ep) \geq s$ with L the Lipschitz parameter of ∇l_0 and η_0 is a user defined parameter. (Like q , η_0 is a regularization parameter customizable by the user.) We also define a β function

$$\beta(q_+, \rho_+, \beta^-) = \Theta^\# \{ \beta^- - \rho_+^{-1} X^T \nabla l_0(X\beta^-; y); q_+, \bar{\eta}(q_+, \rho_+) \}, \quad (\text{A.47})$$

based on (32). (Often, an intercept should be included (say β_1) that is subject to no regularization. We can add a column of ones in the design matrix and redefine the $\Theta^\#$ in (A.47) to keep the first entry and perform quantile-thresholding on the remaining subvector.)

Recall the line search criterion for a trial ρ :

$$(\rho \mathbf{D}_2 - \Delta_l)(\beta(q_{t+1}, \rho, \beta^{(t)}), \beta^{(t)}) \geq 0 \quad (\text{A.48})$$

or

$$\begin{aligned} & \frac{1}{2} \|\beta(q_{t+1}, \rho, \beta^{(t)}) - \beta^{(t)}\|_2^2 \\ & \geq l_0(X\beta(q_{t+1}, \rho, \beta^{(t)})) - l_0(X\beta^{(t)}) \\ & \quad - \langle \nabla l_0(X\beta^{(t)}), X\beta(q_{t+1}, \rho, \beta^{(t)}) - X\beta^{(t)} \rangle. \end{aligned}$$

Then the algorithm can be summarized as follows.

Input: X, y , a quantile parameter sequence $q_t \rightarrow q \in [p]$, a target ℓ_2 -shrinkage $\eta_0 \geq 0$.
Initialization: $\beta^{(0)}, \rho_0$ (say 0 and $L\|X\|_2^2$, respectively).
For each q_{t+1} ($t \geq 0$), perform the following
a) Find ρ_{t+1} by line search with the criterion (A.48).
b) Perform $\beta^{(t+1)} \leftarrow \beta(q_{t+1}, \rho_{t+1}, \beta^{(t)})$ according to (A.47).

We can also add a squeezing operation as step c): $X \leftarrow X_{\mathcal{J}(\beta^{(t+1)})}$ from time to time (say when q_{t+1} reaches $p/2^k$ for k greater than some k_0). In addition, after q_{t+1} reaches q and when the sparsity pattern of $\beta^{(t+1)}$ stabilizes, one can use a classical optimization method to solve a smooth problem to get the nonzero entries of the final estimate. As for step a), many standard line search methods can be used, e.g., backtracking [45]. We use an adaptive search with warm starts. Concretely, given $\alpha \in (0, 1)$, we begin with $\rho \leftarrow \rho_t$, and set $\rho \leftarrow \alpha\rho$ if (A.48) is satisfied for $\beta(q_{t+1}, \rho, \beta^{(t)})$ and $\rho \leftarrow \rho/\alpha$ otherwise, until a small enough ρ_{t+1} makes (A.48) hold while $\alpha\rho_{t+1}$ does not. In practice, it is wise to limit the number (M) of searches. We use $\alpha = 0.5, M = 5$ for implementation.

REFERENCES

- [1] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, “Simultaneous analysis of Lasso and Dantzig selector,” *The Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 08 2009.

[2] P. C. Bellec, G. Lecué, and A. B. Tsybakov, “Slope meets lasso: improved oracle bounds and optimality,” *The Annals of Statistics*, vol. 46, no. 6B, pp. 3603–3642, 2018.

[3] J. Fan and J. Lv, “Sure independence screening for ultrahigh dimensional feature space,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 5, pp. 849–911, 2008.

[4] C.-H. Zhang, “Nearly unbiased variable selection under minimax concave penalty,” *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 04 2010.

[5] P. Bühlmann and B. Yu, “Boosting with the L_2 loss: regression and classification,” *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 324–339, 2003.

[6] D. Needell and J. A. Tropp, “CoSaMP: Iterative signal recovery from incomplete and inaccurate samples,” *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[7] T. Zhang, “Sparse recovery with orthogonal matching pursuit under RIP,” *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 6215–6221, 2011.

[8] ———, “Multi-stage convex relaxation for feature selection,” *Bernoulli*, vol. 19, no. 5B, pp. 2277–2293, 2013.

[9] T. Zhao, H. Liu, and T. Zhang, “Pathwise coordinate optimization for sparse learning: Algorithm and theory,” *The Annals of Statistics*, vol. 46, no. 1, pp. 180–218, 2018.

[10] E. J. Candes and T. Tao, “Decoding by linear programming,” *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[11] Y. She, J. Wang, H. Li, and D. Wu, “Group iterative spectrum thresholding for super-resolution sparse spectral selection,” *IEEE Transactions on Signal Processing*, vol. 61, no. 24, pp. 6371–6386, 2013.

[12] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[13] Y. She, Z. Wang, and J. Shen, “Gaining Outlier Resistance with Progressive Quantiles: Fast Algorithms and Theoretical Studies,” *Journal of the American Statistical Association*, vol. 117, pp. 1282–1295, 2022.

[14] T. Blumensath and M. E. Davies, “Iterative thresholding for sparse approximations,” *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 629–654, 2008.

[15] ———, “Iterative hard thresholding for compressed sensing,” *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[16] G. Raskutti, M. J. Wainwright, and B. Yu, “Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls,” *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6976–6994, 2011.

[17] H. Liu and R. F. Barber, “Between hard and soft thresholding: optimal iterative thresholding algorithms,” *Information and Inference: A Journal of the IMA*, vol. 9, no. 4, pp. 899–933, 2020.

[18] P. Jain, A. Tewari, and P. Kar, “On iterative hard thresholding methods for high-dimensional m-estimation,” *In Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.

[19] Z. Wang, H. Liu, and T. Zhang, “Optimal computational and statistical rates of convergence for sparse nonconvex learning problems,” *Annals of statistics*, vol. 42, no. 6, p. 2164, 2014.

[20] P.-L. Loh and M. J. Wainwright, “Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima,” *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 559–616, 2015.

[21] Y. She, “On the finite-sample analysis of Θ -estimators,” *Electronic Journal of Statistics*, vol. 10, no. 2, pp. 1874–1895, 2016.

[22] Y. She, Z. Wang, and J. Jin, “Analysis of generalized Bregman surrogate algorithms for nonsmooth nonconvex statistical learning,” *The Annals of Statistics*, vol. 49, no. 6, pp. 3434–3459, 2021.

[23] M. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[24] L. M. Bregman, “The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming,” *USSR Computational Mathematics and Mathematical Physics*, vol. 7, pp. 200–217, 1967.

[25] C. Zhang, Y. Jiang, and Y. Chai, “Penalized Bregman divergence for large-dimensional regression and classification,” *Biometrika*, vol. 97, no. 3, pp. 551–566, 2010.

[26] Y. She, “Reduced rank vector generalized linear models for feature extraction,” *Statistics and Its Interface*, vol. 6, pp. 197–209, 2013.

[27] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[28] Y. She and H. Tran, “On cross-validation for sparse reduced rank regression,” *Journal of the Royal Statistical Society: Series B*, vol. 81, pp. 145–161, 2019.

[29] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pp. 267–288, 1996.

[30] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[31] T. Blumensath and M. E. Davies, “Normalized iterative hard thresholding: Guaranteed stability and performance,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 298–309, 2010.

[32] D. Kim, S. Lee, and S. Kwon, “A unified algorithm for the non-convex penalized estimation: The ncpen package,” *arXiv preprint arXiv:1811.05061*, 2018.

[33] Guyon, Isabelle, G. Steve, B.-H. Asa, and G. Dror, “Result analysis of the NIPS 2003 feature selection challenge,” *In Advances in Neural Information Processing Systems*, pp. 545–552, 2005.

[34] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for generalized linear models via coordinate

descent,” *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[35] ———, “Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors),” *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[37] J. Ge, X. Li, H. Jiang, H. Liu, T. Zhang, M. Wang, and T. Zhao, “Picasso: A sparse learning library for high dimensional data analysis in R and Python.” *Journal of Machine Learning Research*, vol. 20, pp. 44–1, 2019.

[38] B. C. Feltes, E. B. Chandelier, B. I. Grisci, and M. Dorn, “CuMiDa: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research,” *Journal of Computational Biology*, vol. 26, no. 4, pp. 376–386, 2019.

[39] M. Mishali and Y. C. Eldar, “Blind multiband signal reconstruction: Compressed sensing for analog signals,” *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 993–1009, 2009.

[40] Z. Song, H. Qi, and Y. Gao, “Real-time multi-gigahertz sub-Nyquist spectrum sensing system for mmwave,” in *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, 2019, pp. 33–38.

[41] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, “Simultaneous sparse approximation via greedy pursuit,” in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 5. IEEE, 2005, pp. v–721.

[42] H. Qi, X. Zhang, and Y. Gao, “Low-complexity subspace-aided compressive spectrum sensing over wideband whitespace,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 11 762–11 777, 2019.

[43] Y. Gao, Z. Song, H. Zhang, S. Fuller, A. Lambert, Z. Ying, P. Mähönen, Y. Eldar, S. Cui, M. D. Plumbley *et al.*, “Sub-Nyquist spectrum sensing and learning challenge,” *Frontiers of Computer Science*, vol. 15, no. 4, pp. 1–5, 2021.

[44] Y. She and K. Chen, “Robust reduced-rank regression,” *Biometrika*, vol. 104, no. 3, pp. 633–647, 2017.

[45] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: Cambridge University Press, 2004.

PLACE
PHOTO
HERE

Yiyuan She received Ph.D. in Statistics from Stanford University in 2008. Since 2008, he has been on the faculty of Department of Statistics at Florida State University. His research interests are in the areas of high dimensional statistics, statistical machine learning, optimization, signal processing, robust statistics, and network science. Currently, he serves as an Associate Editor for Statistics Sinica and Journal of the American Statistical Association. He received Florida State University Developing Scholar Award and NSF CAREER award. He is an Elected Member of the International Statistical Institute, Fellow of the Institute of Mathematical Statistics, and Fellow of the American Statistical Association.

PLACE
PHOTO
HERE

Jiahui Shen received the B.S. degree in Mathematics in 2016 from City University of Hong Kong, and Ph.D. in Statistics in 2022 from Florida State University. He is currently leading on phase II and phase III studies as a full-time senior biostatistician at SONOFI. His research interests are in the areas of deep learning, signal processing, and nonconvex optimization.

PLACE
PHOTO
HERE

Adrian Barbu received a PhD in Mathematics in 2000 from Ohio State University and a PhD in Computer Science in 2005 from University of California, Los Angeles. From 2005 to 2007, he worked in Siemens Corporate Research, first as a research scientist and later as a project manager, with the focus on medical imaging. He received the 2011 Thomas A. Edison Patent Award with his Siemens coauthors for their work on Marginal Space Learning. In 2007, he joined the Statistics Department at Florida State University as an assistant professor, and since 2019 as a professor. He has published more than 70 papers in computer vision, machine learning and medical imaging, and has more than 25 patents related to medical imaging and image denoising. In 2020 he published a Springer book on ‘Monte Carlo Methods’ with his Ph.D. advisor Dr. Song-Chun Zhu.