FROS: Fast Regularized Optimization by Sketching

Yingzhen Yang[1], **and Ping Li**[2]

[1] Arizona State University

[2] Baidu Research, 10900 NE 8th ST, Bellevue, WA 98004, USA

Abstract

Randomized algorithms are important for solving large-scale optimization problems. In this paper, we propose Fast Regularized Optimization by Sketching (FROS) as an efficient solver for a general class of regularized optimization problems. FROS first generates a sketch of the original data matrix, then solves the sketched problem. Different from existing randomized algorithms, FROS handles general Frechet subdifferentiable regularization functions in an unified framework. It is proved that FROS achieves relative-error bounds for the approximation error between the optimization results of the sketched problem and that of the original problem for all convex and certain non-convex regularization. We further propose Iterative FROS which reduces the approximation error exponentially by iteratively invoking FROS. To the best of our knowledge, our results are among the very few results in approximation error of sketching algorithms for a broad class of optimization problems with general regularization. Experimental results demonstrate the effectiveness of the proposed FROS and Iterative FROS algorithms.

Yingzhen Yang's work was conducted while he was a consulting researcher at Baidu Research - Bellevue.

I. INTRODUCTION

Efficient optimization by randomized algorithms is an important topic in machine learning and optimization, and it has broad applications in numerical linear algebra, data analysis and scientific computing. Randomized algorithms based on matrix sketching or random projection have received a lot of attention [1], [2], [3], [4], [5], which solve sketched problems of much smaller scale. Sketching algorithms has been used to approximately solve various large-scale problems including least square regression, robust regression, low-rank approximation and singular value decomposition [6], [7], [8], [9], [10], [11], [12], [13]. On the other hand, regularized problems with convex or non-convex regularization, such as the well-known ℓ^1 or ℓ^2-norm regularized least square estimation, also known as Lasso or ridge regression, play essential roles in machine learning and statistics. While most existing research works demonstrate the potential of random projection and sketching on problems with common convex regularization [14] or convex constraints [15], few efforts are made in the analysis of regularized problems with general convex or non-convex regularization.

In this paper, we study efficient sketching algorithm for a general class of optimization problems with convex or non-convex regularization, which is presented as follows:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) = \frac{1}{2}g(\|\mathbf{A}\mathbf{x}\|_2^2) + \mathbf{b}^\top\mathbf{x} + \lambda h(\mathbf{x}). \tag{1}$$

$\mathbf{A} \in \mathbb{R}^{n\times d}$ is the data matrix or design matrix for regression problems, $\mathbf{b} \in \mathbb{R}^d$, $h\colon \mathbb{R}^d \to \mathbb{R}$ is a regularization function, λ is a positive regularization weight. We study the regime that $n \gg \mathrm{rank}(\mathbf{A})$ where r is the rank of $\mathbf{A}$, and it is a popular setting for large-scale problems such as fast least square estimation by sketching [3]. For sparse linear regression, matrix $\mathbf{A}$ is usually low-rank in practice and it is assumed that $n \gg \mathrm{rank}(\mathbf{A})$. The functions g and h in (1) satisfy the following assumptions:

- A1: $g\colon \mathbb{R} \to \mathbb{R}$ has L_g-Lipschitz continuous gradient, i.e. $|g'(x) - g'(y)| \le L_g|x - y|$. In addition, g' is bounded from both sides, e.g. $g_0 \le g'(x) \le G$ holds for any $x \ge 0$ with positive constants $0 < g_0 \le G$.
- A2: h is bounded from below, i.e. $h(\mathbf{x}) \ge h_0$ for all $\mathbf{x} \in \mathbb{R}^d$.

Such assumptions are mild, and (1) encompasses a broad class of optimization problems in machine learning and optimization. For example, when $g(x) = x + \|\mathbf{y}\|_2^2$, $\mathbf{b} = \mathbf{A}^\top\mathbf{y}$, (1) is equivalent to regularized least square estimation, i.e. $\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda h(\mathbf{x})$ for a response vector $\mathbf{y} \in \mathbb{R}^n$. Furthermore, when $h(\cdot) = \|\cdot\|_2^2$ or $h(\mathbf{x}) = \|\mathbf{x}\|_1$, (1) is the optimization problem for ridge regression or ℓ^1 regularized least square estimation (Lasso).

Optimization for (1) is time consuming when n is large, and such large-scale regularized optimization problems are important due to increasing interest in massive data. To this end, we propose Fast Regularized

Optimization by Sketching (FROS) in this paper as an efficient randomized algorithm for problem (1). With $\tilde{n} < n$ where $\tilde{n}$ is the target row number of a sketch of the data matrix $\mathbf{A}$, FROS first generates a sketched version of $\mathbf{A}$ by $\tilde{\mathbf{A}} = \mathbf{PA}$, then solves the following sketched problem,

$$\min_{\mathbf{x}\in\mathbb{R}^d} \tilde{f}(\mathbf{x}) = \frac{1}{2} g(\|\tilde{\mathbf{A}}\mathbf{x}\|_2^2) + \mathbf{b}^\top \mathbf{x} + \lambda h(\mathbf{x}). \tag{2}$$

One hopes that the optimization result of the sketched problem (2), denoted by $\tilde{\mathbf{x}}^*$, is a good approximation to that of the original problem (1), denoted by $\mathbf{x}^*$. The optimization and theoretical computer science literature are particulary interested in the solution approximation measure defined as the semi-norm induced by the data matrix $\mathbf{A}$, i.e. $\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}} \triangleq \|\mathbf{A}(\tilde{\mathbf{x}}^* - \mathbf{x}^*)\|_2$ where $\|\mathbf{u}\|_{\mathbf{A}} \triangleq \|\mathbf{Au}\|_2$ for any vector $\mathbf{u}$. Existing research, such as Iterative Hessian Sketch (IHS) [15], prefers relative-error approximation to the solution of the original problem in the following form:

$$\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}} \leq \rho \|\mathbf{x}^*\|_A, \tag{3}$$

where $0 < \rho < 1$ is a positive constant. With the relative-error approximation (3), IHS proposes an interesting iterative sketching method to reduce the approximation error $\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}$ geometrically in the iteration number.

A. Contributions and Main Results

Firstly, we prove that FROS for a general class of optimization problems in the form of (1) enjoys an universal approximation error bound for arbitrary regularization function h which is Frechet subdifferentiable, i.e.

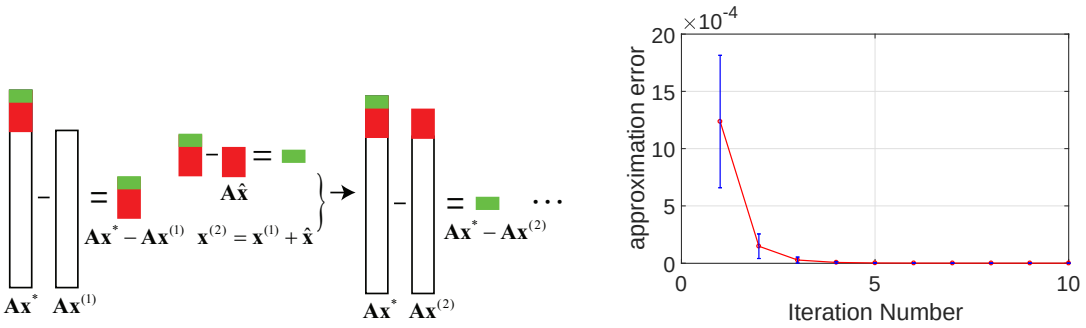
$$C_1(1-\varepsilon)\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}^2 - C_2\varepsilon\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}\|\mathbf{x}^*\|_{\mathbf{A}} \leq Q_h, \tag{4}$$

where C_1 and C_2 are constants determined by parameters of problem (1), $\varepsilon > 0$ is a small positive number which appears in the approximation error bound to be presented. Q_h is a quantity depending on the degree of nonconvexity of h. In particular, $Q_h = 0$ for arbitrary convex regularization function h, leading to a preferred relative-error approximation essential to Iterative FROS to be explained soon, i.e.

$$\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}} \leq \frac{C\varepsilon}{1-\varepsilon}\|\mathbf{x}^*\|_{\mathbf{A}}, \tag{5}$$

where C is a positive constant and $C = 1$ for regularized least square problems. With a small positive ε, the relative-error approximation (3) is achieved. Furthermore, if h is strongly convex, we prove in Theorem 7 that the relative-error approximation also applies to $\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2^2$, the gap between the actual optimization results of the original problem and the sketched problem. If h is non-convex but not "very" non-convex with limited Q_h, we prove that sketching still admits a form of relative-error approximation

bound in Corollary 2. Our theoretical results convey a clear message that relative-error approximation "favors" convex regularization, and show how convexity and nonconvexity of regularization affect the accuracy of sketching by the quantity Q_h on the RHS of (4).



(a) (b)

Fig. 1. (a) Illustration of Iterative FROS. (b) The approximation error of Iterative FROS with respect to iteration number for ridge regression.

Secondly, based on FROS, we present an iterative sketching algorithm termed Iterative FROS which provably reduces the approximation error $\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}$ geometrically in an iterative manner. Albeit being bounded, the approximation error of one-time FROS is still not small enough for various applications. To this end, Iterative FROS iteratively calls FROS to approximate the residual or the approximation error of the last iteration so as to further reduce the approximation error, given the relative-error approximation (3). Figure 1(a) illustrates the idea of Iterative FROS. $\mathbf{x}^{(1)}$ is the result of calling FROS one time at the first iteration. In order to further reduce the approximation error $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_{\mathbf{A}}$, FROS is invoked again to produce $\hat{\mathbf{x}}$ so as to approximate $\mathbf{x}^* - \mathbf{x}^{(1)}$, where an adjusted optimization problem is formed whose solution is $\mathbf{x}^* - \mathbf{x}^{(1)}$. Thanks to the relative-error approximation (3), the approximation error of $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \hat{\mathbf{x}}$ is reduced by a factor of ρ compared to $\mathbf{x}^{(1)}$. In fact, the approximation error provably drops exponentially in the iteration number. Figure 1(b) shows the approximation error of Iterative FROS in red curve for ridge regression, where the blue bar represents standard deviation caused by random data sampling. More details are introduced in Section IV and Section B.

There are two key differences between Iterative FROS and IHS [15]. First of all, utilizing the subspace embedding as the projection matrix $\mathbf{P}$, Iterative FROS does not need to sample $\mathbf{P}$ and compute the sketched matrix $\tilde{\mathbf{A}} = \mathbf{PA}$ at each iteration, in contrast with IHS where a separate $\mathbf{P}$ is sampled and $\tilde{\mathbf{A}}$ is recomputed at each iteration. This advantage saves considerable computation and storage for large-scale problems. Secondly, while IHS is restricted to constrained least-square problems with convex constraints, FROS (Corollary 2) and Iterative FROS (Theorem 8) is capable of handling all convex regularization

and certain non-convex regularization in a unified framework for a more general class of optimization problems, including least square problems as special cases. For example, while Generalized Lasso [16] has not received a lot of attention in randomized optimization, we show that it can be efficiently and effectively solved by Iterative FROS in Section B.

B. Notations

Throughout this paper, we use bold letters for matrices and vectors, regular lower letters for scalars. The bold letter with subscript indicates the corresponding element of a matrix or vector, and the bold letter with superscript indicates the corresponding column of a matrix, i.e. $\mathbf{A}^i$ indicates the i-th column of matrix $\mathbf{A}$. $\|\cdot\|_p$ denotes the ℓ^p-norm of a vector, or the p-norm of a matrix. $\sigma_t(\cdot)$ is the t-th largest singular value of a matrix, and $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ indicate the smallest and largest singular value of a matrix respectively. $\mathrm{tr}(\cdot)$ is the trace of a matrix. $f_1(n) = \Theta(f_2(n))$ if there exist constants $k_1, k_2 > 0$ and n_0 such that $k_1 f_(n) \leq f_1(n) \leq k_2 f_2(n)$. We use $\mathbf{A} \succcurlyeq \mathbf{B}$ to indicate that $\mathbf{A} - \mathbf{B}$ is a positive semi-definite matrix, and $\mathbf{I}_d$ indicates the $d \times d$ identity matrix. $\mathrm{rank}(\mathbf{A})$ means the rank of a matrix $\mathbf{A}$. We use $[m \dots n]$ to indicate numbers between m and n inclusively, and $[n]$ denotes the natural numbers between 1 and n inclusively. $\mathsf{nnz}(\mathbf{A})$ indicates the number of nonzero elements of a matrix $\mathbf{A}$.

II. The FROS Algorithm

In order to improve the efficiency of optimization for (1), we propose Fast Reguarlzied Optimization by Sketching (FROS) in this section. The key idea is to sketch matrix $\mathbf{A}$ in the quadratic term of (1) by random projection. It is comprised of two steps:

Step 1. Project the matrix $\mathbf{A}$ onto a lower dimensional space by a linear transformation $\mathbf{P} \in \mathbb{R}^{\tilde{n} \times n}$ with $\tilde{n} < n$, i.e. $\tilde{\mathbf{A}} = \mathbf{P}\mathbf{A}$.

Step 2. Solve the sketched problem (2).

Using sketching matrix $\mathbf{A}$ only in the quadratic term $\|\mathbf{A}\mathbf{x}\|_2^2$ is proposed in [15] for constrained least square problems with convex constraints. FROS adopts this idea for general regularized problems.

The linear transformation $\mathbf{P}$ is required to be a subspace embedding [17] defined in Definition 1. The literature [18], [19], [20] extensively studies such random transformation which is also closely related to the proof of the Johnson-Lindenstrauss lemma [21].

Definition 1. *Suppose $\mathcal{P}$ is a distribution over $\tilde{n} \times n$ matrices, where $\tilde{n}$ is a function of n, d, ε, and δ. Suppose that with probability at least $1 - \delta$, for any fixed $n \times d$ matrix $\mathbf{A}$, a matrix $\mathbf{P}$ drawn from distribution $\mathcal{P}$ has the property that $\mathbf{P}$ is a $(1 \pm \varepsilon)$ ℓ^2-subspace embedding for $\mathbf{A}$, i.e.*

$$(1 - \varepsilon)\|\mathbf{A}\mathbf{x}\|_2^2 \leq \|\mathbf{P}\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \varepsilon)\|\mathbf{A}\mathbf{x}\|_2^2 \tag{6}$$

holds for all $\mathbf{x} \in \mathbb{R}^d$. *Then we call* $\mathcal{P}$ *an* (ε, δ) *oblivious* ℓ^2*-subspace embedding.*

Remark 1. (Gaussian Subspace Embedding, Theorem 6 in Chapter 2.1, [17]) Let $0 < \varepsilon, \delta < 1$, $\mathbf{P} = \frac{\mathbf{P}'}{\sqrt{\tilde{n}}}$ where $\mathbf{P}^{'} \in \mathbb{R}^{\tilde{n} \times n}$ is a matrix whose elements are i.i.d. samples from the standard Gaussian distribution $\mathcal{N}(0, 1)$. Then if $\tilde{n} = \Theta((r + \log \frac{1}{\delta})\varepsilon^{-2})$, for any matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with $r = \mathrm{rank}(\mathbf{A})$, with probability $1 - \delta$, $\mathbf{P} = \frac{\mathbf{P}'}{\sqrt{\tilde{n}}}$ is a $(1 \pm \varepsilon)$ ℓ^2-subspace embedding for $\mathbf{A}$. $\mathbf{P}$ is named a Gaussian subspace embedding.

Definition 2. (Sparse Subspace Embedding) *Let* $\mathbf{P} \in \mathbb{R}^{\tilde{n} \times n}$. *For each* $i \in [n]$, $h(i) \in [\tilde{n}]$ *is uniformly chosen from* $[\tilde{n}]$, *and* $\sigma(i)$ *is a uniformly random element of* $\{1, -1\}$. *We then set* $\mathbf{P}_{h(i)i} = \sigma(i)$ *and set* $\mathbf{P}_{ji} = 0$ *for all* $j \neq i$. *As a result,* $\mathbf{P}$ *has only a single nonzero element per column, and it is called a sparse subspace embedding.*

Lemma 1 below, also presented in [22], shows that the sparse subspace embedding defined above is indeed a subspace embedding with a high probability.

Lemma 1. ([22]) Let $\mathbf{P} \in \mathbb{R}^{\tilde{n} \times n}$ be a sparse embedding matrix with $\tilde{n} = \mathcal{O}(r^2/\varepsilon^2 \mathrm{poly}(\log(r/\varepsilon)))$ rows, for any fixed $n \times d$ matrix $\mathbf{A}$ with $r = \mathrm{rank}(\mathbf{A})$, with probability .99, $\mathbf{P}$ is a $(1 \pm \varepsilon)$ ℓ^2-subspace embedding for $\mathbf{A}$. Furthermore, $\mathbf{PA}$ can be computed in $\mathcal{O}(\mathrm{nnz}(\mathbf{A}))$ time, where $\mathrm{nnz}(\mathbf{A})$ is the number of nonzero elements of $\mathbf{A}$.

A. Error Bounds

The solutions to the original problem (1) and the sketched problem (2) by typical iterative optimization algorithms, such as gradient descent for smooth h or proximal gradient method for non-smooth h, are always critical points of the corresponding objective functions under mild conditions [23]. Therefore, the analysis in the gap between $\tilde{\mathbf{x}}^*$ and $\mathbf{x}^*$ amounts to the analysis in the distance between critical points of the objective functions of (2) and that of (1), which is presented in Section III. In the sequel, $\tilde{\mathbf{x}}^*$ is a critical point of the objective function (2) and $\mathbf{x}^*$ is a critical point of the objective function (1), if no confusion arises. More details about optimization algorithms are deferred to supplementary.

III. Approximation Error Bounds

We present an universal approximation error bound for FROS on the general problem (1) in Section III-A. We then apply this universal result to problems with convex, strongly convex and non-convex regularization and derive the corresponding relative-error approximation bounds. Before stating our results, the definition of Frechet subdifferential, critical point, strong convexity and degree of nonconvexity are introduced below, which are essential to our analysis.

Definition 3. (Subdifferential and critical points) Given a non-convex function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ which is a proper and lower semi-continuous function,

- for a given $\mathbf{x} \in \mathrm{dom} f$, its Frechet subdifferential of f at $\mathbf{x}$, denoted by $\tilde{\partial} f(\mathbf{x})$, is the set of all vectors $\mathbf{u} \in \mathbb{R}^d$ which satisfy

$$\liminf_{\mathbf{y}\neq\mathbf{x},\mathbf{y}\to\mathbf{x}} \frac{f(\mathbf{y}) - f(\mathbf{x}) - \langle \mathbf{u}, \mathbf{y}-\mathbf{x}\rangle}{\|\mathbf{y}-\mathbf{x}\|} \geq 0.$$

- The limiting-subdifferential of f at $\mathbf{x} \in \mathbb{R}^d$, denoted by $\partial f(\mathbf{x})$, is defined by

$$\partial f(\mathbf{x}) = \{\mathbf{u} \in \mathbb{R}^d \colon \exists \mathbf{x}^k \to \mathbf{x}, f(\mathbf{x}^k) \to f(\mathbf{x}), \tilde{\mathbf{u}}^k \in \tilde{\partial} f(\mathbf{x}^k) \to \mathbf{u}\}.$$

The point $\mathbf{x}$ is a critical point of f if $\mathbf{0} \in \partial f(\mathbf{x})$.

Note that Frechet subdifferential generalizes the notions of Frechet derivative and subdifferential of convex functions. If f is a convex function, then $\tilde{\partial} f(\mathbf{x})$ is also the subdifferential of f at $\mathbf{x}$. In addition, if f is a real-valued differentiable function with gradient $\nabla f(\mathbf{x})$, then $\tilde{\partial} f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

Definition 4. (Strongly convex function) A differential function $h\colon \mathbb{R}^d \to \mathbb{R}$ is σ-strongly convex for $\sigma > 0$ if for any $\mathbf{y}, \mathbf{x} \in \mathbb{R}^d$,

$$h(\mathbf{y}) \geq h(\mathbf{x}) + \langle \nabla h(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{\sigma}{2}\|\mathbf{y}-\mathbf{x}\|_2^2. \tag{7}$$

In order to analyze the relative-error approximation bound, we need the following definition of degree of nonconvexity in terms of the Frechet subdifferential in Definition 3. It is an extension of the univariate degree of nonconvexity presented in [24] used to analyze the consistency of non-convex sparse estimation models with concave regularization.

Definition 5. The degree of nonconvexity of a function $h\colon \mathbb{R}^d \to \mathbb{R}$ at a point $\mathbf{t} \in \mathbb{R}^d$ is defined as

$$\theta_h(\mathbf{t}, \kappa) \triangleq \sup_{\mathbf{s}\in\mathbb{R}^d, \mathbf{s}\neq\mathbf{t}, \mathbf{u}\in\tilde{\partial}h(\mathbf{s}), \mathbf{v}\in\tilde{\partial}h(\mathbf{t})} \frac{-(\mathbf{s}-\mathbf{t})^\top(\mathbf{u}-\mathbf{v}) - \kappa\|\mathbf{s}-\mathbf{t}\|_2^2}{\|\mathbf{s}-\mathbf{t}\|_2}, \tag{8}$$

where $\kappa \in \mathbb{R}$. We abbreviate (8) as $\theta_h(\mathbf{t}, \kappa) \triangleq \sup_{\mathbf{s}\in\mathbb{R}^d, \mathbf{s}\neq\mathbf{t}}\{-\frac{1}{\|\mathbf{s}-\mathbf{t}\|_2}(\mathbf{s}-\mathbf{t})^\top(\tilde{\partial}h(\mathbf{s}) - \tilde{\partial}h(\mathbf{t})) - \kappa\|\mathbf{s}-\mathbf{t}\|_2\}$ in the following text.

Remark 2. It can be verified that the degree of nonconvexity of any convex function h is zero with $\kappa = 0$, i.e. $\theta_h(\mathbf{t}, 0) = \mathbf{0}$ when h is convex. If h is σ-strongly convex and h is twice continuously differentiable, then $\nabla^2 h(\mathbf{x}) \succcurlyeq \sigma \mathbf{I}_d$ for any $\mathbf{x} \in \mathbb{R}^d$, and its degree of nonconvexity is zero with $\kappa = -\sigma$.

A. General Bound

Let $f_0 \triangleq f(\mathbf{x}^0)$, $\tilde{f}_0 \triangleq \tilde{f}(\mathbf{x}^0)$, $h_0 \triangleq \lambda h(\mathbf{x}^0)$, $M_{\varepsilon_0} \triangleq g^{-1}(2(\tilde{f}_0 + h_0)) + \big(g^{-1}(2(\tilde{f}_0 + h_0))\big)^{\frac{1}{2}} \cdot (1 + \varepsilon_0)\big(g^{-1}(2(f_0 + h_0))\big)^{\frac{1}{2}}$

Theorem 1. Suppose $\tilde{\mathbf{x}}^*$ is any critical point of the objective function in (2), and $\mathbf{x}^*$ is any critical point of the objective function in (1). Suppose $0 < \varepsilon < \varepsilon_0 < 1$ where ε_0 is a small positive constant, $0 < \delta < 1$, $\mathbf{P}$ is drawn from an (ε, δ) oblivious ℓ^2-subspace embedding over $\tilde{n} \times n$ matrices, $L_g < \frac{g_0}{M_{\varepsilon_0}}$. Then with probability $1 - \delta$,

$$C_1(1-\varepsilon)\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}^2 - C_2\varepsilon\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}\|\mathbf{x}^*\|_{\mathbf{A}} \leq \lambda\theta_h(\mathbf{x}^*, \kappa)\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2 + \lambda\kappa\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2^2. \quad (9)$$

$C_1 \triangleq g_0 - L_g M_{\varepsilon_0}$, M_{ε_0} is a constant determined by $\mathbf{x}^0$ which is the initialization point for optimization and ε_0, $C_2 \triangleq G + L_g g^{-1}(2(f_0 + h_0))$. In particular, if $\mathbf{P}$ is a Gaussian subspace embedding, then $\tilde{n} = \Theta((r + \log\frac{1}{\delta})\varepsilon^{-2})$. If $\mathbf{P}$ is a sparse subspace embedding, then $\tilde{n} = \mathcal{O}(r^2/\varepsilon^2 \mathrm{poly}(\log(r/\varepsilon)))$ and $\delta = 0.99$.

κ in the definition of degree of nonconvexity and the general approximation error bound (25) reflects how "nonconvexity" of h affects the accuracy of sketching. The following corollary shows that if the Frechet subdifferential of h is Lipschitz continuous with limited Lipschitz constant, or κ equivalently, the degree of nonconvexity can be set to 0, rendering a desirable relative-error approximation bound.

Corollary 1. Under the conditions of Theorem 5, suppose $\mathbf{A}$ has full column rank and h is non-convex. If the Frechet subdifferential of h is L_h-smooth, i.e. $\sup_{\mathbf{u}\in\tilde{\partial}h(\mathbf{x}),\mathbf{v}\in\tilde{\partial}h(\mathbf{y})} \|\mathbf{u} - \mathbf{v}\|_2 \leq L_h\|\mathbf{x} - \mathbf{y}\|_2$. If $\frac{\lambda L_h}{\sigma_{\min}^2(\mathbf{A})} < C_1(1 - \varepsilon_0)$, then

$$\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}} \leq \frac{C_2}{C_1(1-\varepsilon) - \frac{\lambda L_h}{\sigma_{\min}^2(\mathbf{A})}} \cdot \varepsilon\|\mathbf{x}^*\|_{\mathbf{A}}. \quad (10)$$

Remark 3. To the best of our knowledge, (60) is among the very few results in theoretical guarantee of sketching for non-convex regularization. If h is twice continuously differentiable, the condition about L_h-smoothness is reduced to $\nabla^2 h(\mathbf{x}) \succcurlyeq -L_h\mathbf{I}_d$ for all $\mathbf{x} \in \mathbb{R}^d$. The corollary above states that if h is non-convex but not "very" non-convex with a limited L_h or the regularization weight λ is small enough, i.e. $\lambda L_h < \sigma_{\min}^2(\mathbf{A})C_1(1 - \varepsilon_0)$, then we have the relative-error approximation bound (60). Consider the case that $g(x) = x$, then $L_g = 0$ and $C_1 = 1$, and the condition $\frac{\lambda L_h}{\sigma_{\min}^2(\mathbf{A})} < C_1(1 - \varepsilon_0)$ indicates that f is strongly convex although h is non-convex. This confirms the intuition that relative-error approximation is prone to occur for convex optimization.

B. Relative-Error Approximation Bound with Convex Regularization

It can be verified that the degree of nonconvexity vanishes with $\kappa = 0$ when h is convex. As a result, we have relative-error approximation bound for $\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}$ shown in Theorem 6 below.

Theorem 2. If h is convex, then under the conditions of Theorem 5, with probability $1 - \delta$,

$$\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}} \leq \frac{C\varepsilon}{1-\varepsilon}\|\mathbf{x}^*\|_{\mathbf{A}}, \tag{11}$$

where $C \triangleq \frac{C_2}{C_1}$, C_1 and C_2 are defined in Theorem 5.

When h is strongly convex, more fine-grained approximation error bound in terms of the actual gap between $\tilde{\mathbf{x}}^*$ and $\mathbf{x}^*$ is presented in the following theorem.

Theorem 3. Suppose h is σ-strongly convex, then under the conditions of Theorem 5, with probability $1 - \delta$,

$$\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2^2 \leq \frac{C_2^2}{4\lambda\sigma C_1(1-\varepsilon)} \cdot \varepsilon^2 \|\mathbf{x}^*\|_{\mathbf{A}}^2, \tag{12}$$

where C_1 and C_2 are defined in Theorem 5.

IV. Iterative FROS

Algorithm 1 Iterative FROS

Input: Initialize $\mathbf{x}^{(0)} = \mathbf{0}$, iteration number $N > 0$, $t = 0$.

Generate a sketch matrix $\mathbf{P}$ which can either be a Gaussian subspace embedding or sparse subspace embedding, then compute $\tilde{\mathbf{A}} = \mathbf{PA}$.

for $t \leftarrow 1$ to N **do**

Set

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2} g(\|\tilde{\mathbf{A}}(\mathbf{x} + \mathbf{x}^{(t-1)})\|_2^2) + \mathbf{b}^\top \mathbf{x} + \lambda h(\mathbf{x} + \mathbf{x}^{(t-1)}), \tag{13}$$

$$\mathbf{x}^{(t)} = \hat{\mathbf{x}} + \mathbf{x}^{(t-1)} \tag{14}$$

end for

Return $\mathbf{x}^{(N)}$

Inspired by Iterative Hessian Sketch [15], we introduce an iterative sketching method for FROS so that the gap between solutions to the original problem and the sketched problem can be further reduced. The key idea is to iteratively apply FROS to generate a sequence $\{\mathbf{x}^{(t)}\}_{t=1}^N$ such that $\mathbf{x}^{(t)}$ is a more accuracy approximation to $\mathbf{x}^*$, the solution to the original problem (1), than $\mathbf{x}^{(t-1)}$. Consider the optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2} g(\|\mathbf{A}(\mathbf{x} + \mathbf{x}^{(t-1)})\|_2^2) + \mathbf{b}^\top \mathbf{x} + \lambda h(\mathbf{x} + \mathbf{x}^{(t-1)}), \tag{15}$$

then $\mathbf{x}^* - \mathbf{x}^{(t-1)}$ is a solution to (15). We apply FROS to problem (15) and suppose $\hat{\mathbf{x}}$ is an solution to the sketched problem, i.e.

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2} g(\|\tilde{\mathbf{A}}(\mathbf{x} + \mathbf{x}^{(t-1)})\|_2^2) + \mathbf{b}^\top \mathbf{x} + \lambda h(\mathbf{x} + \mathbf{x}^{(t-1)}). \tag{16}$$

$\hat{\mathbf{x}}$ is supposed to be an approximation to $\mathbf{x}^* - \mathbf{x}^{(t-1)}$. If $\hat{\mathbf{x}}$ admits the relative-error approximation bound (3), then $\mathbf{x}^{(t)} = \hat{\mathbf{x}} + \mathbf{x}^{(t-1)}$ becomes a more accurate approximation to $\mathbf{x}^*$ than $\mathbf{x}^{(t-1)}$ by a factor of ρ. This can be verified by noting that $\|\mathbf{x}^{(t)} - \mathbf{x}^*\|_{\mathbf{A}} = \|\hat{\mathbf{x}} - (\mathbf{x}^* - \mathbf{x}^{(t-1)})\|_{\mathbf{A}} \leq \rho \|\mathbf{x}^* - \mathbf{x}^{(t-1)}\|_{\mathbf{A}}$. By mathematical induction, we have Theorem 8 below showing that the approximation error of Iterative FROS, which is formally described by Algorithm 1, drops geometrically in the iteration number. It should be emphasized that Theorem 8 also handles certain non-convex regularization.

Theorem 4. Under the conditions of Theorem 5, with probability $1 - \delta$, the output of Iterative FROS described by Algorithm 1 satisfies

$$\|\mathbf{x}^{(N)} - \mathbf{x}^*\|_{\mathbf{A}} \leq \rho^N \|\mathbf{x}^*\|_{\mathbf{A}} \tag{17}$$

for a constant $0 < \rho < 1$ if h is convex, or the Frechet subdifferential of h is L_h-smooth and $\mathbf{A}$ has full column rank with $\frac{\lambda L_h}{\sigma_{\min}^2(\mathbf{A})} < C_1(1 - \varepsilon_0)$. In particular, if $\mathbf{P}$ is a Gaussian subspace embedding, then $\tilde{n} = \Theta(r + \log \frac{1}{\delta})$. If $\mathbf{P}$ is a sparse subspace embedding, then $\tilde{n} = \mathcal{O}(r^2 \mathrm{poly}(\log r))$ and $\delta = 0.99$, where $r = \mathrm{rank}(\mathbf{A})$.

V. Time Complexity

We compare the time complexity of solving the original problem (1) to that of solving the sketched problem (2) with Iterative FROS. We employ Proximal Gradient Descent (PGD) or Gradient Descent (GD) in our analysis, which are widely used in the machine learning and optimization literature. If $\mathbf{P}$ is a Gaussian subspace embedding in Remark 1, it takes $\mathcal{O}(\tilde{n}nd)$ operations to compute the sketched matrix $\tilde{\mathbf{A}} = \mathbf{PA}$ and then form the sketched problem (2). Let $C(\tilde{n}, d)$ be the time complexity of solving the sketched problem (2), and suppose iterative sketching is performed for N iterations, then the overall time complexity of Iterative FROS in Algorithm 1 is $\mathcal{O}\left(\tilde{n}nd + NC(\tilde{n}, d)\right)$. If $\mathbf{P}$ is a sparse subspace embedding in Definition 2, then it only takes $\mathcal{O}\left(\mathsf{nnz}(\mathbf{A})\right)$ operations to compute the sketched matrix $\tilde{A}$. In this case, the overall time complexity of Iterative FROS is $\mathcal{O}\left(\mathsf{nnz}(\mathbf{A}) + NC(\tilde{n}, d)\right)$. If the data matrix is sparse, then the efficiency of Iterative FROS can be significantly boosted by a sparse subspace embedding. We present a more concrete complexity analysis if PGD, such as that analyzed in [23], or GD, is used to solve problem (1) and (2). It can be verified that each iteration of PGD or GD takes $\mathcal{O}(nd)$ operations or $\mathcal{O}(\tilde{n}d)$ operations for (1) and (2) respectively. Suppose the maximum number of iterations

for PGD or GD is M, then $C(\tilde{n}, d) = \mathcal{O}(M\tilde{n}d)$. If a sparse subspace embedding is used for sketching, then the overall time complexity of Iterative FROS is $\mathcal{O}(\mathsf{nnz}(\mathbf{A}) + NM\tilde{n}d)$. Noting that $N \leq \log n$ [15] and in many practical cases N is a constant, and $\tilde{n} \ll n$, such complexity is much lower than that of solving the original problem with complexity of $\mathcal{O}(Mnd)$.

VI. Experimental Results

A. Ridge Regression

We employ Iterative FROS to approximate solution to Ridge Regression in this subsection, whose optimization problem is $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2n}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2+\lambda\|\mathbf{x}\|_2$. We assume a linear model $\mathbf{y} = \mathbf{A}\mathbf{x}^o+\mathbf{w}$ where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ is the Gaussian noise with unit variance. The unknown regression vector $\mathbf{x}^o$ is also sampled according to $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. We randomly sample $\mathbf{A} \in \mathbb{R}^{n\times d}$ of rank $\frac{n}{10}$ with $n = 5000$ and $d = 10000$. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be the Singular Value Decomposition of $\mathbf{A}$ where $\mathbf{\Sigma}$ is a diagonal matrix whose diagonal elements are the singular values of $\mathbf{A}$, then the elements of each column of $\mathbf{U}$ and $\mathbf{V}$ and the diagonal elements of $\mathbf{\Sigma}$ are i.i.d. standard Gaussian samples. We set $\lambda = \sqrt{\frac{\log(d)}{n}}$. Figure 6 illustrates the logarithm of approximation error $\frac{\|\mathbf{x}^{(i)}-\mathbf{x}^*\|_\mathbf{A}^2}{n}$ with respect to the iteration number i of Iterative FROS for different choices of projection dimension $\tilde{n}$, and the maximum iteration number of Iterative FROS is set to $N = 10$. We let $\tilde{n} = \gamma\mathrm{rank}(\mathbf{A})$ where γ ranges over $\{12, 14, 16, 18, 20\}$, and sample $\mathbf{A}, \mathbf{y}$ and $\mathbf{w}$ 100 times for each γ. The average approximation errors are illustrated in Figure 6. It can be observed from Figure 6 that the convergence rate of approximation error drops geometrically, or its logarithm drops linearly, evidencing our Theorem 8 for Iterative FROS. Moreover, as suggested by Theorem 5, larger $\tilde{n}$ leads to smaller approximation error. We also report approximation error to the true unknown regression vector $\mathbf{x}^o$ by Iterative FROS in the supplementary. For example, over all the trials of data sampling for $\gamma = 16$, $\frac{\|\mathbf{x}^*-\mathbf{x}^o\|_\mathbf{A}}{n}$ has an average value of 0.021, and $\frac{\|\mathbf{x}^{(0)}-\mathbf{x}^o\|_\mathbf{A}}{n} \in 0.0297 \pm 0.013$, $\frac{\|\mathbf{x}^{(N)}-\mathbf{x}^o\|_\mathbf{A}}{n} \in 0.022 \pm 0.008$.

B. Generalized Lasso

We study the performance of Iterative FROS for Generalized Lasso (GLasso) [16] in this subsection. The optimization problem of an instance of GLasso studied here is $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2n}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\sum_{i=1}^{d-1}|\mathbf{x}_i - \mathbf{x}_{i+1}|$, which is solved by Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [25], an accelerated version of PGD. $\mathbf{A} \in \mathbb{R}^{n\times d}$ have i.i.d. standard Gaussian entries with $n = 80000$ and $d = 600$, and all the elements of $\mathbf{y} \in \mathbb{R}^n$ are also i.i.d. Gaussian samples. Figure 6 illustrates the approximation error of FROS and Iterative FROS, i.e. $\frac{\|\mathbf{x}^{(1)}-\mathbf{x}^*\|_\mathbf{A}^2}{n}$ and $\frac{\|\mathbf{x}^{(N)}-\mathbf{x}^*\|_\mathbf{A}^2}{n}$ respectively, for different choices of projection dimension $\tilde{n}$ with $\tilde{n} = \gamma d$. We set $N = 10$ and employ either sparse subspace embedding or

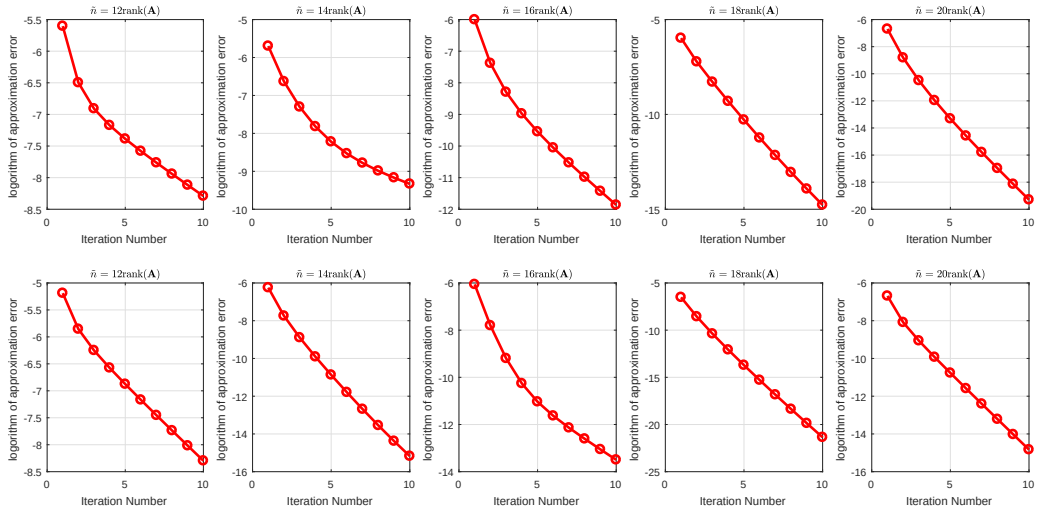


Fig. 2. Approximation Error of Iterative FROS with respect to different projection dimension $\tilde{n}$ for ridge regression. The first row corresponds to sparse subspace embedding defined in Definition 2 and the second row is produced by Gaussian subspace embedding defined in Remark 1.

Gaussian subspace embedding. The average approximation errors are reported over 100 trials of data sampling for each γ. It can be observed that Iterative FROS significantly reduces the approximation error and its approximation error is roughly $\frac{1}{3}$ of that of FROS, demonstrating the effectiveness of Iterative FROS. We defer the actual running times of Iterative FROS, more experimental results such as signal recovery by Lasso and Lasso subspace clustering, and more analysis about Iterative FROS for convex and non-convex regularization to the supplementary. In our experiment, due to the significant reduction in the sample size, e.g. $\frac{\gamma d}{n} = 0.03$ when $\gamma = 4$, the running time of Iterative FROS is always less than half of that required to solve the original problem with small γ.

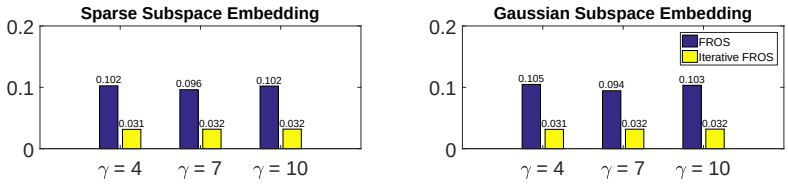


Fig. 3. Approximation error of Iterative FROS vs. FROS for GLasso with respect to different projection dimension $\tilde{n}$. Iterative FROS and FROS are equipped with either sparse subspace embedding (left) or Gaussian Subspace Embedding (right).

VII. Conclusion

We present Fast Regularized Optimization by Sketching (FROS) as a sketching method for efficiently solving general regularized optimization problems with convex or non-convex regularization. FROS generates a sketch of the data matrix by randomized linear transformation, then solves the sketched problem with provable approximation error. We further propose Iterative FROS to reduce the approximation error of FROS geometrically. Experimental results evidence that Iterative FROS can effectively and efficiently approximate the optimization result of the original problem.

References

[1] S. S. Vempala, *The Random Projection Method*, ser. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. DIMACS/AMS, 2004, vol. 65.

[2] C. Boutsidis and P. Drineas, "Random projections for the nonnegative least-squares problem," *Linear Algebra and its Applications*, vol. 431, no. 5, pp. 760 – 771, 2009.

[3] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, "Faster least squares approximation," *Numerische Mathematik*, vol. 117, no. 2, pp. 219–249, 2011.

[4] M. W. Mahoney, "Randomized algorithms for matrices and data," *Foundations and Trends® in Machine Learning*, vol. 3, no. 2, pp. 123–224, 2011.

[5] D. M. Kane and J. Nelson, "Sparser johnson-lindenstrauss transforms," *J. ACM*, vol. 61, no. 1, pp. 4:1–4:23, Jan. 2014.

[6] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions," *SIAM REV*, vol. 53, no. 2, pp. 217–288, 2011.

[7] Y. Lu, P. S. Dhillon, D. P. Foster, and L. H. Ungar, "Faster ridge regression via the subsampled randomized hadamard transform," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, 2013, pp. 369–377.

[8] A. Alaoui and M. W. Mahoney, "Fast randomized kernel ridge regression with statistical guarantees," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 775–783.

[9] G. Raskutti and M. W. Mahoney, "A statistical perspective on randomized sketching for ordinary least-squares," *J. Mach. Learn. Res.*, vol. 17, pp. 214:1–214:31, 2016.

[10] T. Yang, L. Zhang, R. Jin, and S. Zhu, "Theory of dual-sparse regularized randomized reduction," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, vol. 37. JMLR.org, 2015, pp. 305–314.

[11] P. Drineas and M. W. Mahoney, "Randnla: Randomized numerical linear algebra," *Commun. ACM*, vol. 59, no. 6, p. 80–90, May 2016.

[12] S. Oymak, B. Recht, and M. Soltanolkotabi, "Isometric sketching of any set via the restricted isometry property," *Information and Inference: A Journal of the IMA*, vol. 7, no. 4, pp. 707–726, 03 2018.

[13] S. Oymak and J. A. Tropp, "Universality laws for randomized dimension reduction, with applications," *Information and Inference: A Journal of the IMA*, vol. 7, no. 3, pp. 337–446, 11 2017.

[14] W. Zhang, L. Zhang, R. Jin, D. Cai, and X. He, "Accelerated sparse linear regression via random projection," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, 2016, pp. 2337–2343.

[15] M. Pilanci and M. J. Wainwright, "Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1842–1879, Jan. 2016.

[16] R. J. Tibshirani and J. Taylor, "The solution path of the generalized lasso," *Ann. Statist.*, vol. 39, no. 3, pp. 1335–1371, 06 2011. [Online]. Available: https://doi.org/10.1214/11-AOS878

[17] D. P. Woodruff, "Sketching as a tool for numerical linear algebra," *Foundations and Trends® in Theoretical Computer Science*, vol. 10, no. 1–2, pp. 1–157, 2014.

[18] P. Frankl and H. Maehara, "The johnson-lindenstrauss lemma and the sphericity of some graphs," *J. Comb. Theory Ser. A*, vol. 44, no. 3, pp. 355–362, Jun. 1987.

[19] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, ser. STOC '98. New York, NY, USA: ACM, 1998, pp. 604–613.

[20] L. Zhang, T. Yang, R. Jin, and Z. Zhou, "Sparse learning for large-scale and high-dimensional data: A randomized convex-concave optimization approach," in *Algorithmic Learning Theory - 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings*, 2016, pp. 83–97.

[21] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of johnson and lindenstrauss," *Random Struct. Algorithms*, vol. 22, no. 1, pp. 60–65, Jan. 2003.

[22] K. L. Clarkson and D. P. Woodruff, "Low rank approximation and regression in input sparsity time," in *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, ser. STOC '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 81–90. [Online]. Available: https://doi.org/10.1145/2488608.2488620

[23] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, no. 1-2, pp. 459–494, Aug. 2014.

[24] C.-H. Zhang and T. Zhang, "A general theory of concave regularization for high-dimensional sparse estimation problems," *Statist. Sci.*, vol. 27, no. 4, pp. 576–593, 11 2012.

[25] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Img. Sci.*, vol. 2, no. 1, pp. 183–202, Mar. 2009.

[26] P. Bhlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, 1st ed. Springer Publishing Company, Incorporated, 2011.

[27] Y. Wang and H. Xu, "Noisy sparse subspace clustering," in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, 2013, pp. 89–97.

[28] M. Soltanolkotabi and E. J. Candés, "A geometric analysis of subspace clustering with outliers," *Ann. Statist.*, vol. 40, no. 4, pp. 2195–2238, 08 2012.

[29] X. Zheng, D. Cai, X. He, W.-Y. Ma, and X. Lin, "Locality preserving clustering for image database," in *Proceedings of the 12th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '04. New York, NY, USA: ACM, 2004, pp. 885–891.

[30] E. Elhamifar and R. Vidal, "Sparse manifold clustering and embedding," in *NIPS*, 2011, pp. 55–63.

[31] E. L. Dyer, A. C. Sankaranarayanan, and R. G. Baraniuk, "Greedy feature selection for subspace clustering," *Journal of Machine Learning Research*, vol. 14, pp. 2487–2517, 2013.

[32] K. Davidson and S. Szarek, "Local operator theory, random matrices and Banach spaces," in *Handbook on the Geometry of Banach spaces*, Lindenstrauss, Ed. Elsevier Science, 2001, vol. 1, pp. 317–366.

APPENDIX A
MORE DISCUSSIONS

1) Why do we need Iterative FROS to further reduce the approximation error of FROS?

In many case, the solution to the sketched problem (2) should be close to the solution to the original problem (1) enough to match the prediction error of statistical models. Section B-C of this supplementary provides a concrete analysis for the case of Lasso explaining why Iterative FROS is required to further reduce the approximation error of FROS.

2) Can FROS and Iterative FROS perform sketch for non-convex optimization problems?

We have the following claim indicating that $\frac{1}{2}g(\|\mathbf{A}\mathbf{x}\|_2^2)$ can be non-convex under the conditions and assumptions required by all the theoretical results of this paper.

Claim A. If $L_g > 0$ and $\mathrm{rank}(\mathbf{A}) > 1$, then $g_1(\mathbf{x}) \triangleq \frac{1}{2}g(\|\mathbf{A}\mathbf{x}\|_2^2)$ in the original problem (1) is non-convex if g satisfies

$$-L_g \leq g''(x) < a_0 \tag{18}$$

for some negative constant a_0 and any nonnegative $x \in \mathbb{R}$. Moreover, there exists a subspace $\mathcal{R}$ of $\mathbb{R}^n$ and a positive constant $c_{\mathbf{x}}$ such that when $\mathbf{x} \in \mathcal{R}$ and $\|\mathbf{x}\|_2 > c_{\mathbf{x}}$, $\nabla^2 g_1(\mathbf{x})$ is indefinite.

The proof of the above claim is in Section C of this supplementary. The condition that $L_g > 0$ is compatible with Assumptions A1-A2 and the conditions required by all the theoretical results of this paper. **As a result, all the theoretical results of this paper hold for non-convex instances of the original problem (1)** if $\mathrm{rank}(\mathbf{A}) > 1$, which is a rather mild condition. According to Claim A, when $L_g > 0$, $\mathrm{rank}(\mathbf{A}) > 1$ and g satisfies (18), $\nabla^2 g_1(\mathbf{x})$ is indefinite when $\mathbf{x} \in \mathcal{R}$ and $\|\mathbf{x}\|_2 > c_{\mathbf{x}}$. Furthermore, if the regularization function is chosen as $h(\mathbf{x}) = \|\mathbf{x}\|_1$, it can be verified that $\nabla^2 f(\mathbf{x})$ and $\nabla^2 \tilde{f}(\mathbf{x})$ are indefinite when $\mathbf{x}_i \neq 0$ for all $i \in [d]$ and $\mathbf{x} \in \mathcal{R}$. Therefore, both $f(\mathbf{x})$ and $\tilde{f}(\mathbf{x})$ are non-convex when $L_g > 0$, $\mathrm{rank}(\mathbf{A}) > 1$ and g satisfies (18), and all the theoretical results of this paper still hold. An example of g satisfying (18) is $g(x) = -ax^2 + bx + c$ for some constant $|a_0| < a \leq L_g$ and arbitrary $b, c \in \mathbb{R}$. **Therefore, FROS and Iterative FROS provably renders approximation solutions with proved approximation error in the paper for the general optimization problem (1) which can be either convex or non-convex. It is worthwhile to mention that $\mathbf{x}^{(t)}$ in Algorithm 1 for Iterative FROS is a critical point of the objective function (13) in the paper for all $t \in [N]$.**

3) Time Complexity of Iterative FROS and IHS [15]

The time complexity of Iterative FROS is analyzed in Section 5 of the paper. If the sketch matrix $\mathbf{P}$ is a Gaussian subspace embedding, the overall time complexity of Iterative FROS is $\mathcal{O}\left(\tilde{n}nd + NC(\tilde{n}, d)\right)$,

where $C(\tilde{n}, d)$ is the time complexity of solving the sketched problem. If $\mathbf{P}$ is a sparse subspace embedding, the overall time complexity of Iterative FROS is $\mathcal{O}\left(\mathsf{nnz}(\mathbf{A}) + NC(\tilde{n}, d)\right)$. Since IHS needs to sample an independent sketch matrix at each iteration, the time complexity of IHS using fast Johnson-Lindenstrauss sketches is $\mathcal{O}\left(Nnd\log\tilde{n} + NC(\tilde{n}, d)\right)$, which is higher than that of Iterative FROS with sparse subspace embedding. It should be emphasized that Iterative FROS is even more efficient for sparse data matrix $\mathbf{A}$.

APPENDIX B
COMPLETE EXPERIMENTAL RESULTS

We demonstrate complete experimental results of FROS and Iterative FROS in this section for three instances of the general optimization problem (1), i.e. ridge regression where $h(\mathbf{x}) = \|\mathbf{x}\|_2^2$, Generalized Lasso where $h(\mathbf{x}) = \sum_{i=1}^{d-1} |\mathbf{x}_i - \mathbf{x}_{i+1}|$, and Lasso where $h(\mathbf{x}) = \|\mathbf{x}\|_1$. We choose g and $\mathbf{b}$ such that $\frac{1}{2}g(\|\mathbf{A}\mathbf{x}\|_2^2) + \mathbf{b}^\top\mathbf{x} = \frac{1}{2n}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ for all the three instances studied where $\mathbf{y}$ is an given input signal. We further show the application of FROS in subspace clustering in the last part of this section.

A. Ridge Regression of Larger Scale

We present more experimental results for ridge regression under similar setting as that in Section 6.1 of this paper. We assume a linear model $\mathbf{y} = \mathbf{A}\mathbf{x}^o + \mathbf{w}$ where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ is the Gaussian noise with unit variance. The unknown regression vector $\mathbf{x}^o$ is also sampled according to $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. We randomly sample $\mathbf{A} \in \mathbb{R}^{n\times d}$ of rank $\frac{n}{100}$ with $n = 10000$ and $d = 100000$ (the paper used $n = 5000$ and $d = 10000$). Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be the Singular Value Decomposition of $\mathbf{A}$ where $\mathbf{\Sigma}$ is a diagonal matrix whose diagonal elements are the singular values of $\mathbf{A}$, then the elements of each column of $\mathbf{U}$ and $\mathbf{V}$ and the diagonal elements of $\mathbf{\Sigma}$ are i.i.d. standard Gaussian samples. We use Gram–Schmidt process to produce orthogonal columns of $\mathbf{U}$ and $\mathbf{V}$. We let $\tilde{n} = \gamma\mathrm{rank}(\mathbf{A})$ where γ ranges over $\{12, 14, 16, 18, 20\}$, and sample $\mathbf{A}, \mathbf{y}$ and $\mathbf{w}$ 100 times for each γ.

Figure 4 illustrates the logarithm of approximation error, i.e. $\log \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_{\mathbf{A}}^2}{n}$, with respect to the iteration number i of Iterative FROS for different choices of projection dimension $\tilde{n}$, and the maximum iteration number of Iterative FROS is set to $N = 10$. It can be observed that the logarithm of approximation error drops linearly with respect to the iteration number in most cases, evidencing our theory that the approximation error of Iterative FROS drops geometrically in the iteration number.

Figure 5 illustrates the approximation error of Iterative FROS in red curve for ridge regression, i.e. $\frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_{\mathbf{A}}^2}{n}$. The blue bar represents standard deviation caused by random data sampling and random subspace embedding. A single subspace embedding is sampled for each sampled data, and this subspace embedding is used throughout all the iterations of Iterative FROS, in contrast with Iterative Hessian

Sketch (IHS) [15] which samples a separate random projection matrix for each iteration of the iterative sketch procedure.

Table I demonstrates the approximation error to the true unknown regression vector $\mathbf{x}^o$ by FROS, Iterative FROS and the solution $\mathbf{x}^*$ to the original problem (1). It can be observed that Iterative FROS consistently reduces the approximation error to the true regression vector, and its approximation error is very close to that of $\mathbf{x}^*$, the solution to the original problem, demonstrating the effectiveness of Iterative FROS for ridge regression. Similar to Figure 5, standard deviation is showed in this table which is caused by random data sampling and random subspace embedding.

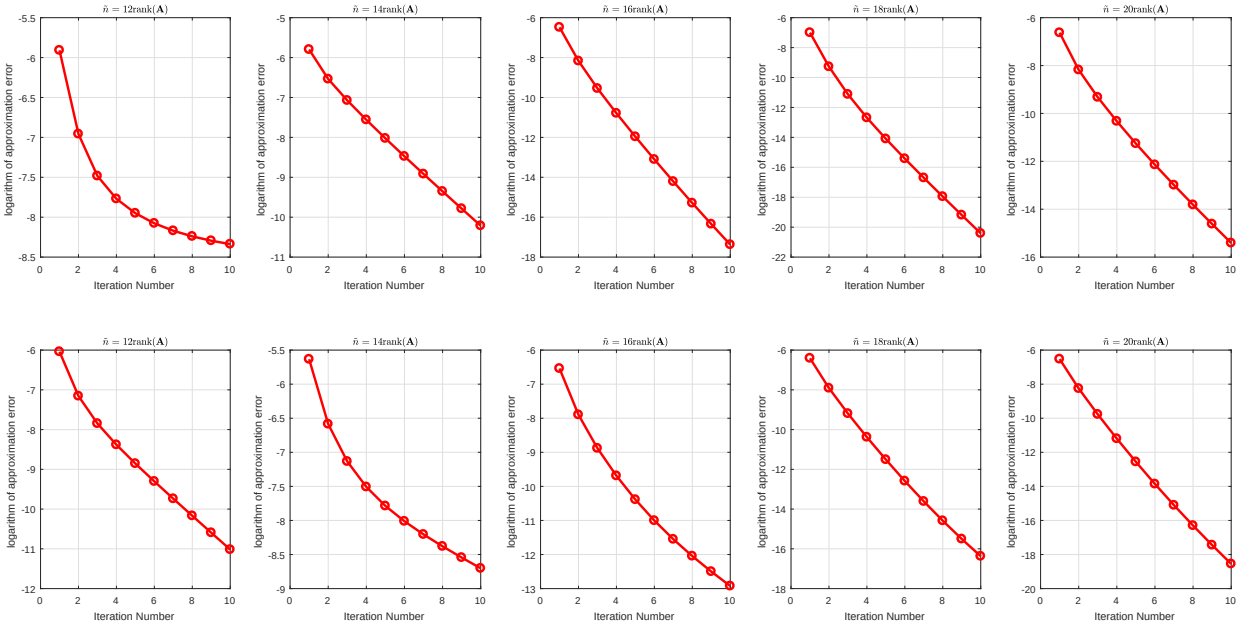


Fig. 4. Logarithm of approximation error of Iterative FROS with respect to different projection dimension $\tilde{n}$ for ridge regression. The first row corresponds to sparse subspace embedding and the second row is produced by Gaussian subspace embedding.

B. Generalized Lasso

In this subsection, we add more details to Section 6.2 of the paper for Generalized Lasso (GLasso) [16]. The optimization problem of GLasso studied here is $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2n}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \sum_{i=1}^{d-1} |\mathbf{x}_i - \mathbf{x}_{i+1}|$, which is solved by Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [25]. We construct $\mathbf{D} \in \mathbb{R}^{(d-1)\times d}$ by setting $\mathbf{D}_{i,i} = -1$, $\mathbf{D}_{i,i+1} = 1$ for all $i \in [d-1]$, then $\sum_{i=1}^{d-1} |\mathbf{x}_i - \mathbf{x}_{i+1}| = \|\mathbf{D}\mathbf{x}\|_1$. Let

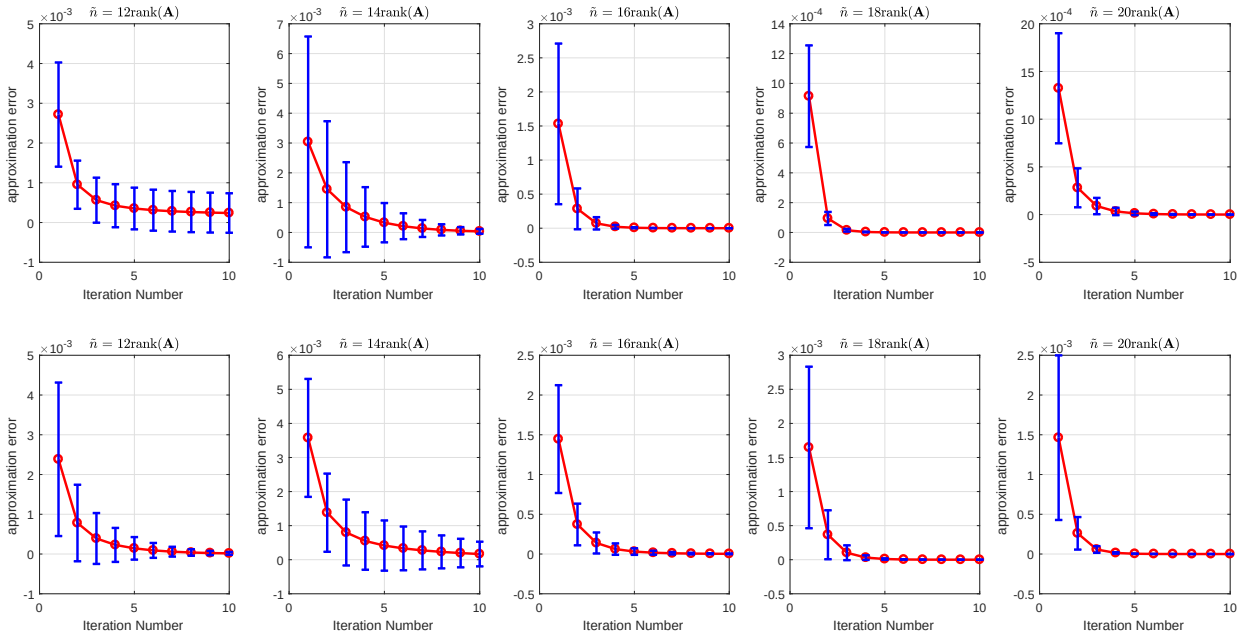


Fig. 5. Approximation error of Iterative FROS with respect to different projection dimension $\tilde{n}$ for ridge regression. The first row corresponds to sparse subspace embedding and the second row is produced by Gaussian subspace embedding.

TABLE I

APPROXIMATION ERROR TO THE TRUE UNKNOWN REGRESSION VECTOR $\mathbf{x}^o$ BY FROS, ITERATIVE FROS AND THE SOLUTION $\mathbf{x}^*$ TO THE ORIGINAL PROBLEM (1) FOR RIDGE REGRESSION.

	Error \ γ	12	14	16	18	20
Sparse Subspace Embedding	FROS	0.023 ± 0.014	0.027 ± 0.015	0.023 ± 0.009	0.026 ± 0.016	0.0184 ± 0.004
	Iterative FROS	0.017 ± 0.009	0.018 ± 0.005	0.021 ± 0.007	0.023 ± 0.013	0.0180 ± 0.004
	Error of $\mathbf{x}^*$	0.017 ± 0.009	0.019 ± 0.005	0.021 ± 0.007	0.023 ± 0.013	0.0180 ± 0.004
Gaussian Subspace Embedding	FROS	0.032 ± 0.019	0.031 ± 0.011	0.018 ± 0.004	0.022 ± 0.007	0.021 ± 0.010
	Iteratie FROS	0.021 ± 0.010	0.022 ± 0.009	0.015 ± 0.005	0.018 ± 0.003	0.015 ± 0.007
	Error of $\mathbf{x}^*$	0.021 ± 0.010	0.022 ± 0.008	0.015 ± 0.006	0.018 ± 0.003	0.015 ± 0.007

$\mathbf{D}^{\text{ext}} = \begin{bmatrix} \mathbf{D} \\ 0, \ldots, 1 \end{bmatrix}$ with $\mathbf{D}^{\text{ext}}_{dd} = 1$. Denote $\boldsymbol{\beta} = \mathbf{D}\mathbf{x}$, then $\mathbf{x} = \left(\mathbf{D}^{\text{ext}}\right)^{-1}\boldsymbol{\beta}$ because $\mathbf{D}^{\text{ext}}$ is nonsingular. The instance of GLasso considered above is then rewritten as $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{2n}\|\mathbf{y} - \mathbf{A}\left(\mathbf{D}^{\text{ext}}\right)^{-1}\boldsymbol{\beta}\|_2^2 + \lambda\sum_{i=1}^{d-1}|\boldsymbol{\beta}_i|$, which can be solved by FISTA.

We also report the running time of GLasso under the setting of Section 6.2 of the paper with $n = 80000$

and $d = 600$ in Table II. We use $M = 10000$ iterations for FISTA, and the running time is reported for $\gamma = 4$. Iterative FROS-IHS is the "IHS" version of Iterative FROS where a new sketch matrix $\mathbf{P}$ is sampled and the sketched data $\tilde{\mathbf{A}} = \mathbf{PA}$ is recomputed at each iteration. We can see that Iterative FROS-IHS incurs extra running time due to the extra operations at each iteration.

TABLE II

RUNNING TIME (IN SECONDS) OF FROS, ITERATIVE FROS WITH $\gamma = 4$ AND ITERATIVE FROS-IHS FOR GLASSO.

FROS	Iterative FROS	Iterative FROS-IHS
$11.57s$	$5.21s$	$5.32s$

C. Signal Recovery by Lasso

We present experimental results for signal recovery/approximation by Lasso in this subsection. We assume a linear model $\mathbf{y} = \mathbf{A}\mathbf{x}^o + \mathbf{w}$ where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ is the Gaussian noise with unit variance. The optimization problem of Lasso considered here is $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2n}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$. Following the setting in (IHS) [15], we set $\lambda = 0.1 \cdot \frac{\log d}{n}$, sparsity $s = \lfloor 3\log d \rfloor$, and choose the unknown regression vector $\mathbf{x}^o$ with its support uniformly random with entries $\pm\frac{1}{\sqrt{s}}$ with equal probability. We randomly sample $\mathbf{A} \in \mathbb{R}^{n\times d}$ of rank $\frac{n}{100}$ with $n = 5000$ and $d = 100000$ in the same manner as that used for ridge regression in Section B-A. We let $\tilde{n} = \gamma\mathrm{rank}(\mathbf{A})$ where γ ranges over $\{12, 14, 16, 18\}$, and sample $\mathbf{A}, \mathbf{y}$ and $\mathbf{w}$ 100 times for each γ. Figure 6 illustrates the approximation error of FROS and Iterative FROS, i.e. $\frac{\|\mathbf{x}^{(1)}-\mathbf{x}^*\|_{\mathbf{A}}^2}{n}$ and $\frac{\|\mathbf{x}^{(N)}-\mathbf{x}^*\|_{\mathbf{A}}^2}{n}$ respectively, for different choices of projection dimension $\tilde{n}$ with different γ. It can be observed that Iterative FROS significantly and constantly reduces approximation error of FROS.

This is a good example for which we explain why Iterative FROS is required to reduce the approximation error of FROS. As a well studied model, the prediction error of Lasso has received extensive study in statistics. For example, [26] proves that the prediction error of Lasso satisfies $\frac{1}{n}\|\mathbf{A}(\mathbf{x}^* - \mathbf{x}^o)\|_2^2 \lesssim \frac{s\log d}{n}$, where $a \lesssim b$ indicates that $a \leq c \cdot b$ for some positive constant c. We would like to approximate $\mathbf{x}^*$ by $\tilde{\mathbf{x}}^*$, the solution to the sketched problem, and the approximation error of $\tilde{\mathbf{x}}^*$ should be on a par with $\frac{s\log d}{n}$ so that $\tilde{\mathbf{x}}^*$ can be a reasonable approximation to $\mathbf{x}^o$, as suggested by [15]. Suppose FROS is used as the sketching method, then by Theorem 2 of the paper, ε on the RHS of equation (11) should be $\Theta\left(\sqrt{\frac{s\log d}{n}}\right)$. It follows that projection dimension for Gaussian subspace embedding should be $\Theta((r + \log\frac{1}{\delta})\frac{n}{s\log d})$, where $r = \mathrm{rank}(\mathbf{A})$ and $1 - \delta$ with $0 < \delta < 1$ is the probability of success for subspace embedding, which could be even larger than n. Similar conclusion is deduced for sparse subspace embedding. To

address this problem, we employ Iterative FROS to further reduce the approximation error of FROS so as to match the prediction error of Lasso, while mainitaining a small projection dimension $\tilde{n}$ for efficiency. In this experiment, we find that a maximum iteration number of $N = 4$ for Iterative FROS suffices for this purpose. Table III shows the approximation error to the true unknown regression vector $\mathbf{x}^o$ by FROS, Iterative FROS and the solution $\mathbf{x}^*$ to the original problem (1). Standard deviation is caused by random data sampling and random subspace embedding. It can be seen that Iterative FROS approximates the true regression vector much better than FROS with the projection dimension $\tilde{n}$ being a fraction of n, especially with Gaussian subspace embedding, justifying our theoretical analysis.

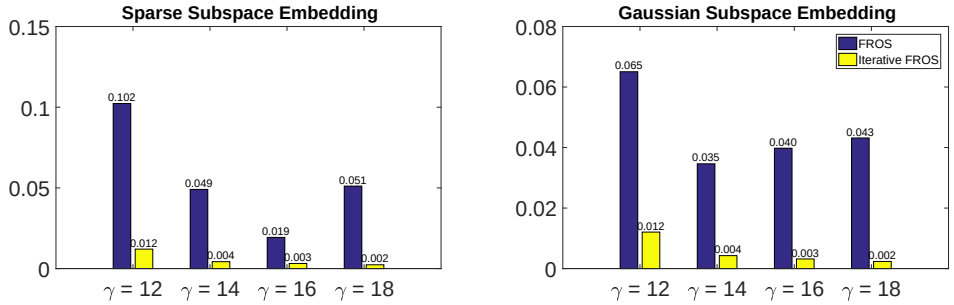


Fig. 6. Approximation error of Iterative FROS and FROS for Lasso with respect to different projection dimension $\tilde{n} = \gamma \mathrm{rank}(\mathbf{A})$. Iterative FROS and FROS are equipped with either sparse subspace embedding (left) or Gaussian Subspace Embedding (right).

TABLE III

APPROXIMATION ERROR TO THE TRUE UNKNOWN REGRESSION VECTOR $\mathbf{x}^o$ BY FROS, ITERATIVE FROS AND THE SOLUTION $\mathbf{x}^*$ TO THE ORIGINAL PROBLEM (1) FOR LASSO.

	Error \ γ	12	14	16	18
Sparse Subspace Embedding	FROS	0.115 ± 0.149	0.063 ± 0.023	0.025 ± 0.010	0.059 ± 0.054
	Iteratie FROS	0.021 ± 0.017	0.016 ± 0.009	0.008 ± 0.002	0.011 ± 0.005
	Error of $\mathbf{x}^*$	0.006 ± 0.002	0.007 ± 0.006	0.003 ± 0.001	0.006 ± 0.004
Gaussian Subspace Embedding	FROS	0.067 ± 0.092	0.043 ± 0.066	0.044 ± 0.038	0.046 ± 0.030
	Iteratie FROS	0.008 ± 0.003	0.010 ± 0.007	0.007 ± 0.002	0.007 ± 0.004
	Error of $\mathbf{x}^*$	0.003 ± 0.001	0.004 ± 0.003	0.003 ± 0.001	0.002 ± 0.002

D. Application in Lasso Subspace Clustering

We demonstrate application of FROS in subspace clustering in this subsection. Given a data matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ comprised of d data points in $\mathbb{R}^n$ which lie in a union of subspaces in $\mathbb{R}^n$, classical subspace

clustering methods using sparse codes, such as Noisy Sparse Subspace Clustering (Noisy SSC) [27], recovers the subspace structure by solving the Lasso problem

$$\mathbf{x}^{i*} = \underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min} \|\mathbf{A}^i - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1, \quad \mathbf{x}_i = 0, \tag{19}$$

for each $i \in [d]$. $\mathbf{A}^i$ is the i-th column of $\mathbf{A}$, which is also the i-th data point, $\lambda > 0$ is the weight for the ℓ^1 regularization. Under certain conditions on $\mathbf{A}$ and the underlying subspaces, it is proved by [28], [27] that nonzero elements of $\mathbf{x}^*$ correspond to data points lying in the same subspace as $\mathbf{A}^i$, and in this case $\mathbf{x}^*$ is said to satisfy the Subspace Detection Property (SDP). It has been proved that SDP is crucial for subspace recovery in the subspace clustering literature. By solving (19) for all $i \in [d]$, we have a sparse code matrix $\mathbf{X} = [\mathbf{x}^{1*}, \mathbf{x}^{2*}, \ldots, \mathbf{x}^{d*}] \in \mathbb{R}^{d\times d}$, and a sparse similarity matrix $\mathbf{W}$ is constructed by $\mathbf{W} = \frac{|\mathbf{X}|+|\mathbf{X}^\top|}{2}$. Spectral clustering is performed on $\mathbf{W}$ to produce the final clustering result of Noisy SSC. Two measures are used to evaluate the performance of different clustering methods, i.e. the Accuracy (AC) and the Normalized Mutual Information (NMI) [29]. In this experiment, we employ FROS to solve the sketched version of problem (19) with $\tilde{n} = \frac{n}{15}$. Note that FROS is equivalent to Iterative FROS with $N = 1$, and we do not incur more iterations by Iterative FROS since FROS produces satisfactory results. Figure 7 and Figure 8 illustrate the accuracy (left) and NMI (right) of sketched Noisy SSC by FROS with respect to various choices of the regularization weight λ on the Extended Yale-B Dataset. The Extended Yale-B Dataset contains face images for 38 subjects with about 64 frontal face images of size 32×32 taken under different illuminations for each subject, and $\mathbf{A}$ is of size 1024×2414. FROS-GSE stands for FROS with Gaussian subspace embedding, and FROS-SSE stands for FROS with sparse subspace embedding. We compare FROS-GSE and FROS-SSE to K-means (KM), Spectral Clustering (SC), Sparse Manifold Clustering and Embedding (SMCE) [30] and SSC by Orthogonal Matching Pursuit (SSC-OMP) [31]. It can be observed that FROS-GSE and FROS-SSE outperform other competing clustering methods by a notable margin, and they perform even better than the original Noisy SSC for most values of λ, due to the fact that sketching potentially reduces the adverse effect of noise in the original data.

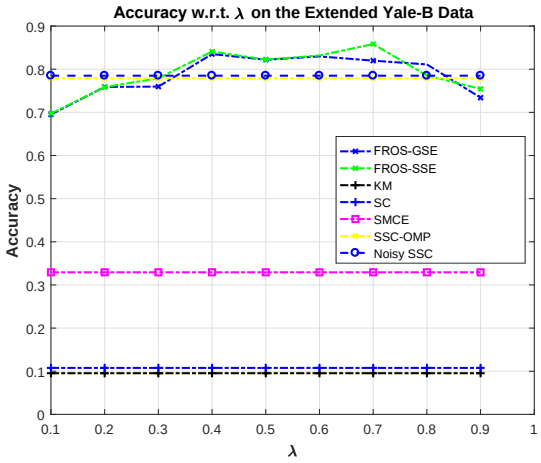


Fig. 7. Accuracy with respect to different values of λ on the Extended Yale-B Dataset

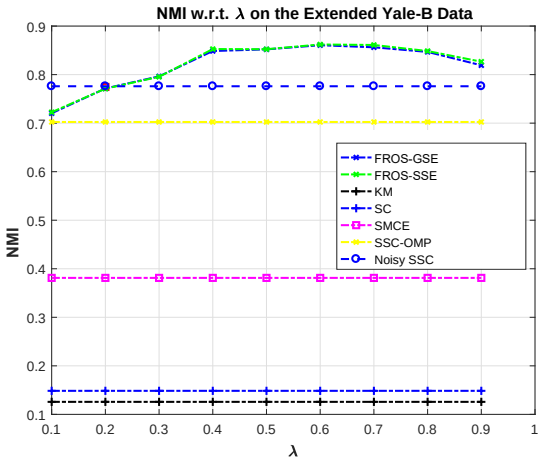


Fig. 8. NMI with respect to different values of λ on the Extended Yale-B Dataset

APPENDIX C
PROOFS

We present proofs of theoretical results of the original paper in this section.

Proof of Claim A. Since $g_1(\mathbf{x}) \triangleq \frac{1}{2}g(\|\mathbf{A}\mathbf{x}\|_2^2)$, by computing the Hessian of g_1, we have

$$\begin{aligned}\nabla^2 g_1(\mathbf{x}) &= g''(\|\mathbf{A}\mathbf{x}\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top\mathbf{A}^\top\mathbf{A} + g'(\|\mathbf{A}\mathbf{x}\|_2^2)\mathbf{A}^\top\mathbf{A} \\ &\overset{(i)}{=} g''(\|\mathbf{A}\mathbf{x}\|_2^2)\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top\mathbf{x}\mathbf{x}^\top\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top + g'(\|\mathbf{A}\mathbf{x}\|_2^2)\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top \end{aligned} \tag{20}$$

where (i) is due to the Singular Value Decomposition of $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n\times r}$. $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ is a diagonal matrix with diagonal elements being singular values of $\mathbf{A}$, $\mathbf{V} \in \mathbb{R}^{d\times r}$, $r = \text{rank}(\mathbf{A}) \leq \min\{n, d\}$. Suppose the diagonal elements of $\mathbf{\Sigma}$ are in descending order, i.e. $\Sigma_{11} \geq \Sigma_{22} \geq \ldots \Sigma_{rr}$. Let the r columns of $\mathbf{V}$ be $\mathbf{V}^1, \mathbf{V}^2, \ldots, \mathbf{V}^r$, and $\mathbf{x} = c\mathbf{V}^1$ for $c > 0$. Then it can be verified that when $c > c_{\mathbf{x}} \triangleq \frac{1}{\Sigma_{11}}\sqrt{\frac{G}{|a_0|}}$, $\nabla^2 g_1(\mathbf{x})$ is indefinite since it has at least one negative eigenvalue and one positive eigenvalue value. □

Before presenting the proof of Theorem 5, the following lemma is introduced which shows that subspace embedding approximately preserves inner product with high probability.

Lemma A. Suppose $\mathbf{P}$ is a $(1 \pm \varepsilon)$ ℓ^2-subspace embedding for $\mathbf{A}$, and let $\mathcal{C}(\mathbf{A})$ denote the column space of $\mathbf{A}$. Then with probability $1 - \delta$, for any two vectors $\mathbf{u} \in \mathcal{C}(\mathbf{A})$, $\mathbf{v} \in \mathcal{C}(\mathbf{A})$,

$$|\mathbf{u}^\top\mathbf{P}^\top\mathbf{P}\mathbf{v} - \mathbf{u}^\top\mathbf{v}| \leq \|\mathbf{u}\|_2\|\mathbf{v}\|_2\varepsilon. \tag{21}$$

Proof. If $\mathbf{u}=\mathbf{0}$ or $\mathbf{v}=\mathbf{0}$, then (21) holds trivially. Otherwise, let $\mathbf{u}'=\frac{\mathbf{u}}{\|\mathbf{u}\|_2}$, $\mathbf{v}'=\frac{\mathbf{v}}{\|\mathbf{v}\|_2}$, and $\mathbf{u}',\mathbf{v}'\in\mathcal{C}(\mathbf{A})$. According to the definition of $(1\pm\varepsilon)$ ℓ^2-subspace embedding for $\mathbf{A}$, with probability $1-\delta$,

$$(1-\varepsilon)\|\mathbf{u}'+\mathbf{v}'\|_2^2\le\|\mathbf{P}(\mathbf{u}'+\mathbf{v}')\|_2^2\le(1+\varepsilon)\|\mathbf{u}'+\mathbf{v}'\|_2^2, \tag{22}$$

$$(1-\varepsilon)\|\mathbf{u}'-\mathbf{v}'\|_2^2\le\|\mathbf{P}(\mathbf{u}'-\mathbf{v}')\|_2^2\le(1+\varepsilon)\|\mathbf{u}'-\mathbf{v}'\|_2^2. \tag{23}$$

Subtracting (23) from (22), we have

$$|\mathbf{u}'^\top\mathbf{P}^\top\mathbf{P}\mathbf{v}'-\mathbf{u}'^\top\mathbf{v}'|\le\varepsilon, \tag{24}$$

and (21) holds by scaling (24) by $\|\mathbf{u}\|_2\|\mathbf{v}\|_2$. □

Theorem 5. Suppose $\tilde{\mathbf{x}}^*$ is any critical point of the objective function in (2), and $\mathbf{x}^*$ is any critical point of the objective function in (1). Suppose $0<\varepsilon<\varepsilon_0<1$ where ε_0 is a small positive constant, $0<\delta<1$, $\mathbf{P}$ is drawn from an (ε,δ) oblivious ℓ^2-subspace embedding over $\tilde{n}\times n$ matrices, $L_g<\frac{g_0}{M_{\varepsilon_0}}$. Then with probability $1-\delta$, either $\tilde{\mathbf{x}}^*=\mathbf{x}^*$, or

$$C_1(1-\varepsilon)\|\tilde{\mathbf{x}}^*-\mathbf{x}^*\|_{\mathbf{A}}^2-C_2\varepsilon\|\tilde{\mathbf{x}}^*-\mathbf{x}^*\|_{\mathbf{A}}\|\mathbf{x}^*\|_{\mathbf{A}}\le\lambda\theta_h(\mathbf{x}^*,\kappa)\|\tilde{\mathbf{x}}^*-\mathbf{x}^*\|_2+\lambda\kappa\|\tilde{\mathbf{x}}^*-\mathbf{x}^*\|_2^2, \tag{25}$$

where $C_1\triangleq g_0-L_gM_{\varepsilon_0}$, M_{ε_0} is a constant determined by $\mathbf{x}^{(0)}$ and ε_0, $C_2\triangleq G+L_gg^{-1}(2(f_0+h_0))$. In particular, if $\mathbf{P}$ is a Gaussian subspace embedding, then $\tilde{n}=\Theta((r+\log\frac{1}{\delta})\varepsilon^{-2})$. If $\mathbf{P}$ is a sparse subspace embedding, then $\tilde{n}=\mathcal{O}(r^2/\varepsilon^2\mathrm{poly}(\log(r/\varepsilon)))$ and $\delta=0.99$.

Proof of Theorem 5. By the optimality of $\tilde{\mathbf{x}}^*$, we have

$$\|g'(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*+\mathbf{b}+\lambda\mathbf{v}\|_2=0 \tag{26}$$

for some $\mathbf{v}\in\tilde{\partial}h(\tilde{\mathbf{x}}^*)$, where $\tilde{\mathbf{y}}\in\mathbb{R}^n$ satisfies $\tilde{\mathbf{A}}^\top\tilde{\mathbf{y}}=\mathbf{A}^\top\mathbf{y}$. We will also write (26) as $\|g'(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*+\lambda\tilde{\partial}h(\tilde{\mathbf{x}}^*)\|_2=0$ for simplicity without confusion. In the sequel, we will use $\tilde{\partial}h(\cdot)$ to indicate an element belonging to $\tilde{\partial}h(\cdot)$ if no special note is made.

By the optimality of $\mathbf{x}^*$, we have

$$\|g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^*+\mathbf{b}+\lambda\tilde{\partial}h(\mathbf{x}^*)\|_2=0. \tag{27}$$

Again, (27) indicates that $\|g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^*+\lambda\mathbf{u}\|_2=0$ for some $\mathbf{u}\in\tilde{\partial}h(\mathbf{x}^*)$.

Let $\Delta=\tilde{\mathbf{x}}^*-\mathbf{x}^*$, $\tilde{\Delta}=\lambda\big(\tilde{\partial}h(\tilde{\mathbf{x}}^*)-\tilde{\partial}h(\mathbf{x}^*)\big)$. By (26) and (27), we have

$$\begin{aligned}&\|g'(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*+\mathbf{b}+\lambda\tilde{\partial}h(\tilde{\mathbf{x}}^*)-g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^*-\mathbf{b}-\lambda\tilde{\partial}h(\mathbf{x}^*)\|_2\\&=\|g'(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*-g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^*+\lambda\tilde{\Delta}\|_2=0\end{aligned} \tag{28}$$

By Cauchy–Schwarz inequality,

$$\begin{aligned}&\Delta^\top\big(g'(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^* - g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^* + \lambda\tilde{\Delta}\big)\\ &\le \|\Delta\|_2\|g'(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^* - g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^* + \lambda\tilde{\Delta}\|_2 = 0. \end{aligned} \tag{29}$$

On the other hand, the LHS of (29) can be written as

$$\begin{aligned}&\Delta^\top\big(g'(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^* - g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^* + \lambda\tilde{\Delta}\big)\\ &= \Delta^\top\big(g'(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^* - g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\mathbf{x}^*\big) + \Delta^\top\big(g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\mathbf{x}^*\big) - g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^*\big)\\ &\quad +\lambda\Delta^\top\tilde{\Delta}. \end{aligned} \tag{30}$$

By (29) and (30), we have

$$\begin{aligned}&\Delta^\top\big(g'(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^* - g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\mathbf{x}^*\big) + \Delta^\top\big(g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\mathbf{x}^*\big) - g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^*\big)\\ &\le -\lambda\Delta^\top\tilde{\Delta} \end{aligned} \tag{31}$$

Now we derive lower bounds for the two terms on the RHS of (31). First, we have

$$\begin{aligned}&\Delta^\top\big(g'(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^* - g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\mathbf{x}^*\big)\\ &= \Delta^\top\big(g'(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^* - g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^* + g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^* - g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\mathbf{x}^*\big)\\ &= \Delta^\top g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\Delta + \Delta^\top\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\big(g'(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2^2) - g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\big)\\ &\ge g_0\|\tilde{\mathbf{A}}\Delta\|_2^2 - \Big|\Delta^\top\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\Big|\, L_g\,\Big|\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2^2 - \|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2\Big|\\ &= g_0\|\tilde{\mathbf{A}}\Delta\|_2^2 - \Big|\Delta^\top\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\Big|\, L_g(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2 + \|\tilde{\mathbf{A}}\mathbf{x}^*\|_2)\,\Big|\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2 - \|\tilde{\mathbf{A}}\mathbf{x}^*\|_2\Big|\\ &\ge g_0\|\tilde{\mathbf{A}}\Delta\|_2^2 - \Big|\Delta^\top\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\Big|\, L_g(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2 + \|\tilde{\mathbf{A}}\mathbf{x}^*\|_2)\|\tilde{\mathbf{A}}\Delta\|_2\\ &\ge g_0\|\tilde{\mathbf{A}}\Delta\|_2^2 - L_g\|\tilde{\mathbf{A}}\Delta\|_2^2\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2 + \|\tilde{\mathbf{A}}\mathbf{x}^*\|_2)\\ &\overset{(i)}{\ge} g_0\|\tilde{\mathbf{A}}\Delta\|_2^2 - L_g\|\tilde{\mathbf{A}}\Delta\|_2^2\Big(g^{-1}(2(\tilde{f}_0+h_0)) + \big(g^{-1}(2(\tilde{f}_0+h_0))\big)^{\frac{1}{2}}\cdot(1+\varepsilon_0)\big(g^{-1}(2(f_0+h_0))\big)^{\frac{1}{2}}\Big)\\ &= \underbrace{(g_0 - L_g M_{\varepsilon_0})}_{C_1}\|\tilde{\mathbf{A}}\Delta\|_2^2\\ &= C_1\|\tilde{\mathbf{A}}\Delta\|_2^2 \overset{(ii)}{\ge} C_1(1-\varepsilon)\|\mathbf{A}\Delta\|_2^2. \end{aligned} \tag{32}$$

Since optimization algorithms are supposed to decrease the objective value evaluated at the initialization $\mathbf{x}^0$, we have $\tilde{f}(\tilde{\mathbf{x}}^*) \le \tilde{f}(\mathbf{x}^0)$ and $f(\mathbf{x}^*) \le \tilde{f}(\mathbf{x}^0)$, which lead to $g(\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}\|_2^2) \le 2(\tilde{f}_0+h_0)$ and $g(\|\mathbf{A}\mathbf{x}^*\|_2^2) \le 2(f_0+h_0)$ respectively. As a result, $\|\tilde{\mathbf{A}}\tilde{\mathbf{x}}^*\|_2 \le g^{-1}(2(\tilde{f}_0+h_0))$ and $\|\mathbf{A}\mathbf{x}^*\|_2 \le g^{-1}(2(f_0+h_0))$. By

the property of subspace embedding, $\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2 \le (1+\varepsilon_0)\|\mathbf{A}\mathbf{x}^*\|_2 \le (1+\varepsilon_0)\big(g^{-1}(2(f_0+h_0))\big)^{\frac{1}{2}}$, and (i) follows.

Moreover,

$$\begin{aligned}&\Delta^\top\big(g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\mathbf{x}^* - g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^*\big)\\ &= \Delta^\top\big(g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\mathbf{x}^* - g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^*\big) + \Delta^\top\big(g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^* - g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^*\big).\end{aligned} \tag{33}$$

Now we derive analyze the two terms on the RHS of (33). First, we have

$$\begin{aligned}\Delta^\top\big(g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\mathbf{x}^* - g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^*\big) &= g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\Delta^\top(\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}} - \mathbf{A}^\top\mathbf{A})\mathbf{x}^*\\ &\ge -g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)|\Delta^\top(\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}} - \mathbf{A}^\top\mathbf{A})\mathbf{x}^*|\\ &\overset{(i)}{\ge} -G\varepsilon\|\mathbf{A}\Delta\|_2\|\mathbf{A}\mathbf{x}^*\|_2,\end{aligned} \tag{34}$$

where (i) follows by applying Lemma A and Cauchy–Schwarz inequality.

$$\begin{aligned}\Delta^\top\big(g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^* - g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^*\big) &= \big(g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2) - g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\big)\Delta^\top\mathbf{A}^\top\mathbf{A}\mathbf{x}^*\\ &\overset{(i)}{\ge} -|g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2) - g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)|\cdot\|\mathbf{A}\Delta\|_2\|\mathbf{A}\mathbf{x}^*\|_2\\ &\overset{(ii)}{\ge} -L_g|\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2 - \|\mathbf{A}\mathbf{x}^*\|_2^2|\cdot\|\mathbf{A}\Delta\|_2\|\mathbf{A}\mathbf{x}^*\|_2\\ &\overset{(iii)}{\ge} -L_g\varepsilon\|\mathbf{A}\mathbf{x}^*\|_2^2\cdot\|\mathbf{A}\Delta\|_2\|\mathbf{A}\mathbf{x}^*\|_2\\ &\overset{(iv)}{\ge} -\varepsilon L_g g^{-1}(2(f_0+h_0))\|\mathbf{A}\Delta\|_2\|\mathbf{A}\mathbf{x}^*\|_2.\end{aligned} \tag{35}$$

(i) is due to Cauchy–Schwarz inequality, (i) follows from the L_g-smoothness of g', (iii) is due to the definition of subspace embedding, i.e. $(1-\varepsilon)\|\mathbf{A}\mathbf{x}^*\|_2^2 \le \|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2 \le (1+\varepsilon)\|\mathbf{A}\mathbf{x}^*\|_2^2$. By the optimality of $\mathbf{x}^*$, $f(\mathbf{x}^*) \le f(\mathbf{x}^{(0)}) = f_0$, and it follows that $g(\|\mathbf{A}\mathbf{x}^*\|_2^2) \le 2(f_0+h_0)$. Since g is increasing, we have $\|\mathbf{A}\mathbf{x}^*\|_2^2 \le g^{-1}(2(f_0+h_0))$. Therefore, (iv) follows.

Combining (34) and (35), we have

$$\begin{aligned}&\Delta^\top\big(g'(\|\tilde{\mathbf{A}}\mathbf{x}^*\|_2^2)\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}\mathbf{x}^* - g'(\|\mathbf{A}\mathbf{x}^*\|_2^2)\mathbf{A}^\top\mathbf{A}\mathbf{x}^*\big)\\ &\ge -G\varepsilon\|\mathbf{A}\Delta\|_2\|\mathbf{A}\mathbf{x}^*\|_2 - \varepsilon L_g g^{-1}(2(f_0+h_0))\|\mathbf{A}\Delta\|_2\|\mathbf{A}\mathbf{x}^*\|_2\\ &\ge -\underbrace{\big(G + L_g g^{-1}(2(f_0+h_0))\big)}_{C_2}\varepsilon\|\mathbf{A}\Delta\|_2\|\mathbf{A}\mathbf{x}^*\|_2\end{aligned} \tag{36}$$

Plugging (32) and (36) in (31), we have

$$C_1(1-\varepsilon)\|\mathbf{A}\Delta\|_2^2 - C_2\varepsilon\|\mathbf{A}\Delta\|_2\|\mathbf{A}\mathbf{x}^*\|_2 \leq -\lambda\Delta^\top\tilde{\Delta}. \tag{37}$$

By (37) and the definition of degree of nonconvexity, we have

$$C_1(1-\varepsilon)\|\mathbf{A}\Delta\|_2^2 - C_2\varepsilon\|\mathbf{A}\Delta\|_2\|\mathbf{A}\mathbf{x}^*\|_2 \leq -\lambda\Delta^\top\tilde{\Delta} \leq \lambda\theta_h(\mathbf{x}^*,\kappa)\|\Delta\|_2 + \lambda\kappa\|\Delta\|_2^2. \tag{38}$$

□

Corollary 2. Under the conditions of Theorem 5, suppose $\mathbf{A}$ is nonsingular and the Frechet subdifferential of h is L_h-smooth, i.e. $\sup_{\mathbf{u}\in\tilde{\partial}h(\mathbf{x}),\mathbf{v}\in\tilde{\partial}h(\mathbf{y})}\|\mathbf{u}-\mathbf{v}\|_2 \leq L_h\|\mathbf{x}-\mathbf{y}\|_2$. If $\frac{\lambda L_h}{\sigma_{\min}^2(\mathbf{A})} < C_1(1-\varepsilon_0)$, then

$$\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}} \leq \frac{C_2}{C_1(1-\varepsilon) - \frac{\lambda L_h}{\sigma_{\min}^2(\mathbf{A})}} \cdot \varepsilon\|\mathbf{x}^*\|_{\mathbf{A}}. \tag{39}$$

Proof of Corollary 2. Because the Frechet subdifferential of h is L_h-smooth, namely $\sup_{\mathbf{u}\in\tilde{\partial}h(\mathbf{x}),\mathbf{v}\in\tilde{\partial}h(\mathbf{y})}\|\mathbf{u}-\mathbf{v}\|_2 \leq L_h\|\mathbf{x}-\mathbf{y}\|_2$, it can be veried that $\theta_h(\mathbf{x}^*,\kappa) \leq 0$ with holds with $\kappa = L_h$. This is due to the fact that for any $\mathbf{u}\in\tilde{\partial}h(\mathbf{x})$ and any $\mathbf{v}\in\tilde{\partial}h(\mathbf{y})$,

$$\begin{aligned}
&-(\mathbf{s}-\mathbf{t})^\top(\mathbf{u}-\mathbf{v}) - L_h\|\mathbf{s}-\mathbf{t}\|_2^2 \leq \|\mathbf{s}-\mathbf{t}\|_2\|\mathbf{u}-\mathbf{v}\|_2 - L_h\|\mathbf{s}-\mathbf{t}\|_2^2 \\
&\leq L_h\|\mathbf{x}-\mathbf{y}\|_2^2 - L_h\|\mathbf{s}-\mathbf{t}\|_2^2 \leq 0.
\end{aligned} \tag{40}$$

Therefore,

$$\theta_h(\mathbf{t}, L_h) = \sup_{\mathbf{s}\in\mathbb{R}^d,\mathbf{s}\neq\mathbf{t},\mathbf{u}\in\tilde{\partial}h(\mathbf{s}),\mathbf{v}\in\tilde{\partial}h(\mathbf{t})} \frac{-(\mathbf{s}-\mathbf{t})^\top(\mathbf{u}-\mathbf{v}) - L_h\|\mathbf{s}-\mathbf{t}\|_2^2}{\|\mathbf{s}-\mathbf{t}\|_2} \leq 0 \tag{41}$$

holds for arbitrary $\mathbf{t}$, and $\theta_h(\mathbf{x}^*,\kappa) \leq 0$.

Moreover, since $\mathbf{A}$ is nonsingular, $\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2^2 \leq \frac{\|\tilde{\mathbf{x}}^*-\mathbf{x}^*\|_{\mathbf{A}}^2}{\sigma_{\min}^2(\mathbf{A})}$.

Plugging the above results in (25) and setting κ to L_h, we have

$$\begin{aligned}
&C_1(1-\varepsilon)\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}^2 - C_2\varepsilon\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}\|\mathbf{x}^*\|_{\mathbf{A}} \leq \lambda\theta_h(\mathbf{x}^*, L_h)\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2 + \lambda L_h\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2^2 \\
&\leq \frac{\lambda L_h}{\sigma_{\min}^2(\mathbf{A})} \cdot \|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}^2,
\end{aligned} \tag{42}$$

and (60) immediately follows from (42). □

Theorem 6. If h is convex, then under the conditions of Theorem 5, with probability $1-\delta$,

$$\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}} \leq \frac{C\varepsilon}{1-\varepsilon}\|\mathbf{x}^*\|_{\mathbf{A}}, \tag{43}$$

where $C \triangleq \frac{C_2}{C_1}$, C_1 and C_2 are defined in Theorem 5.

Proof of Theorem 6. When h is convex, then its Frechet differential coincides with its subdifferential. As a result, for any $\mathbf{u} \in \tilde{\partial} h(\mathbf{s})$ and $\mathbf{v} \in \tilde{\partial} h(\mathbf{t})$, we have $(\mathbf{s}-\mathbf{t})(\mathbf{u}-\mathbf{v}) \geq 0$, and it follows that

$$-(\mathbf{s}-\mathbf{t})^{\top}(\mathbf{u}-\mathbf{v}) \leq 0. \tag{44}$$

For $\kappa = 0$ and $\mathbf{s} \neq \mathbf{t}$, (44) suggests that

$$\theta_h(\mathbf{t}, 0) = \sup_{\mathbf{s}\in\mathbb{R}^d, \mathbf{s}\neq\mathbf{t}} \{-\frac{1}{\|\mathbf{s}-\mathbf{t}\|_2}(\mathbf{s}-\mathbf{t})^{\top}(\tilde{\partial} h(\mathbf{s}) - \tilde{\partial} h(\mathbf{t}))\} \leq 0, \tag{45}$$

By (25) in Theorem 5 and (45), setting $\kappa = 0$, we have

$$C_1(1-\varepsilon)\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}^2 - C_2\varepsilon\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}\|\mathbf{x}^*\|_{\mathbf{A}} \leq \lambda\theta_h(\mathbf{x}^*, \kappa)\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2 + \lambda\kappa\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2^2 \leq 0 \tag{46}$$

if $\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}} \neq 0$, it follows from (46) that

$$\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}} \leq \frac{C_2}{C_1}\frac{\varepsilon}{1-\varepsilon}\|\mathbf{x}^*\|_{\mathbf{A}} = \frac{C\varepsilon}{1-\varepsilon}\|\mathbf{x}^*\|_{\mathbf{A}}. \tag{47}$$

□

Theorem 7. Suppose h is σ-strongly convex, then under the conditions of Theorem 5, with probability $1-\delta$,

$$\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2^2 \leq \frac{C_2^2}{4\lambda\sigma C_1(1-\varepsilon)} \cdot \varepsilon^2\|\mathbf{x}^*\|_{\mathbf{A}}^2, \tag{48}$$

where C_1 and C_2 are defined in Theorem 5.

Proof of Theorem 7. For strongly convex function h, we have

$$-(\mathbf{s}-\mathbf{t})^{\top}\big(\nabla h(\mathbf{s}) - \nabla h(\mathbf{t})\big) \leq -\sigma\|\mathbf{s}-\mathbf{t}\|_2^2. \tag{49}$$

Therefore, the degree of nonconvexity $\theta_h(\mathbf{t}, \kappa) = 0$ with $\kappa = -\sigma$. By (25), setting $\kappa = -\sigma$, we have

$$\begin{aligned} &C_1(1-\varepsilon)\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}^2 - C_2\varepsilon\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}}\|\mathbf{x}^*\|_{\mathbf{A}} \leq \lambda\theta_h(\mathbf{x}^*, -\sigma)\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2 - \lambda\sigma\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2^2 \\ &\leq -\lambda\sigma\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2^2. \end{aligned} \tag{50}$$

(50) indicates that

$$\begin{aligned} &4\lambda\sigma\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2^2 \cdot C_1(1-\varepsilon) \leq C_2^2\varepsilon^2\|\mathbf{x}^*\|_{\mathbf{A}}^2 \\ &\Rightarrow \|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_2^2 \leq \frac{C_2^2}{4\lambda\sigma C_1(1-\varepsilon)} \cdot \varepsilon^2\|\mathbf{x}^*\|_{\mathbf{A}}^2 \end{aligned} \tag{51}$$

□

Theorem 8. Under the conditions of Theorem 5, with probability $1-\delta$, the output of Iterative FROS described by Algorithm 1 satisfies

$$\|\mathbf{x}^{(N)}-\mathbf{x}^*\|_{\mathbf{A}} \leq \rho^N \|\mathbf{x}^*\|_{\mathbf{A}} \tag{52}$$

for a constant $0<\rho<1$ if h is convex, or the Frechet subdifferential of h is L_h-smooth and $\mathbf{A}$ has full column rank with $\frac{\lambda L_h}{\sigma^2_{\min}(\mathbf{A})} < C_1(1-\varepsilon_0)$. In particular, if $\mathbf{P}$ is a Gaussian subspace embedding, then $\tilde{n}=\Theta(r+\log\frac{1}{\delta})$. If $\mathbf{P}$ is a sparse subspace embedding, then $\tilde{n}=\mathcal{O}(r^2\mathrm{poly}(\log r))$ and $\delta=0.99$, where $r=\mathrm{rank}(\mathbf{A})$.

Proof of Theorem 8. This proof mostly follows from the proof of our main Theorem 5, Theorem 6 and Corollary 2. We first consider that case that h is convex and $\mathbf{P}$ is a Gaussian subspace embedding. Note that $\mathbf{x}^*-\mathbf{x}^{(t-1)}$ is a solution to

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2} g(\|\mathbf{A}(\mathbf{x}+\mathbf{x}^{(t-1)})\|_2^2) + \mathbf{b}^\top \mathbf{x} + \lambda h(\mathbf{x}+\mathbf{x}^{(t-1)}). \tag{53}$$

Since $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2} g(\|\tilde{\mathbf{A}}(\mathbf{x}+\mathbf{x}^{(t-1)})\|_2^2) + \mathbf{b}^\top \mathbf{x} + \lambda h(\mathbf{x}+\mathbf{x}^{(t-1)})$, by repeating the proof of Theorem 5 and Theorem 6, with probability $1-\delta$,

$$\|\hat{\mathbf{x}}-(\mathbf{x}^*-\mathbf{x}^{(t-1)})\|_{\mathbf{A}} \leq \frac{C'\varepsilon}{1-\varepsilon}\|\mathbf{x}^*-\mathbf{x}^{(t-1)}\|_{\mathbf{A}} \tag{54}$$

for some positive constant C' and for all $t\in[N]$. For any $t\geq 1$, it follows from (54) that

$$\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_{\mathbf{A}} = \|\hat{\mathbf{x}}+\mathbf{x}^{(t-1)}-\mathbf{x}^*\|_{\mathbf{A}} = \|\hat{\mathbf{x}}-(\mathbf{x}^*-\mathbf{x}^{(t-1)})\|_{\mathbf{A}} \leq \frac{C'\varepsilon}{1-\varepsilon}\|\mathbf{x}^*-\mathbf{x}^{(t-1)}\|_{\mathbf{A}} \tag{55}$$

For a constant $0<\rho<1$, by choosing $\varepsilon < \frac{\rho}{\rho+C'}$ and (55), we have

$$\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_{\mathbf{A}} \leq \rho\|(\mathbf{x}^*-\mathbf{x}^{(t-1)}\|_{\mathbf{A}} \tag{56}$$

for any $t\geq 1$. It follows that

$$\|\mathbf{x}^{(N)}-\mathbf{x}^*\|_{\mathbf{A}} \leq \rho^N\|\mathbf{x}^*\|_{\mathbf{A}} \tag{57}$$

with $\mathbf{x}^{(0)}=\mathbf{0}$. The same proof is applied for the case that h is L_h-smooth and $\mathbf{A}$ has full column rank with $\frac{\lambda L_h}{\sigma^2_{\min}(\mathbf{A})} < C_1(1-\varepsilon_0)$, or $\mathbf{P}$ is a sparse subspace embedding. □

As the last part of this section, we present an additional theoretical result in Corollary A showing that when $\mathbf{A}$ has i.i.d. standard Gaussian entries, relative-error approximation bound for non-convex regularization function h under the conditions of Corollary 2 holds with a high probability. We first present Lemma B below about the bounds for maximum and minimum singular values of $\mathbf{A}$ with i.i.d. standard Gaussian entries.

Lemma B. (Spectrum bound for Gaussian random matrix, Theorem $II.13$ in [32]) Suppose $\mathbf{A} \in \mathbb{R}^{n\times d}$ ($n \geq d$) is a random matrix whose entries are i.i.d. samples generated from the standard Gaussian distribution $\mathcal{N}(0, \frac{1}{n})$. Then

$$1 - \sqrt{\frac{d}{n}} \leq \mathbb{E}[\sigma_d(\mathbf{A})] \leq \mathbb{E}[\sigma_1(\mathbf{A})] \leq 1 + \sqrt{\frac{d}{n}}. \tag{58}$$

Also, for any $t > 0$,

$$\begin{aligned} \Pr[\sigma_d(\mathbf{A}) \leq 1 - \sqrt{\frac{d}{n}} - t] < e^{-\frac{nt^2}{2}}, \\ \Pr[\sigma_1(\mathbf{A}) \geq 1 + \sqrt{\frac{d}{n}} + t] < e^{-\frac{nt^2}{2}}. \end{aligned} \tag{59}$$

Corollary A. (Relative-error approximation with Gaussian random matrix $\mathbf{A}$) Under the conditions of Corollary 2, suppose $\mathbf{A} \in \mathbb{R}^{n\times d}$ ($n \geq d$) is a random matrix whose entries are i.i.d. samples generated from the standard Gaussian distribution $\mathcal{N}(0, \frac{1}{n})$, and $\frac{d}{n} \to 0$ when $n \to \infty$. If $\lambda L_h < C_1(1-\varepsilon_0)$, then there exists $n_0 > 0$ such that when $n > n_0$

$$\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}} \leq \frac{C_2}{C_1(1-\varepsilon) - \frac{\lambda L_h}{\left(1-\sqrt{\frac{d}{n}}-n^{-\frac{\alpha}{2}}\right)^2}} \cdot \varepsilon \|\mathbf{x}^*\|_{\mathbf{A}}. \tag{60}$$

Proof of Corollary A. Let $0 < \alpha < 1$. According to Lemma B, with probability at least $1 - e^{-\frac{n^{1-\alpha}}{2}}$,

$$\sigma_{\min}(\mathbf{A}) = \sigma_d(\mathbf{A}) \geq 1 - \sqrt{\frac{d}{n}} - n^{-\frac{\alpha}{2}}. \tag{61}$$

Because $\frac{d}{n} \to 0$ and $n^{-\frac{\alpha}{2}} \to 0$ when $n \to \infty$, there exists $n_0 > 0$ such that when $n > n_0$, $\frac{\lambda L_h}{\sigma^2_{\min}(\mathbf{A})} < C_1(1-\varepsilon)$.

Plugging (61) in (60), we have

$$\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|_{\mathbf{A}} \leq \frac{C_2}{C_1(1-\varepsilon) - \frac{\lambda L_h}{\left(1-\sqrt{\frac{d}{n}}-n^{-\frac{\alpha}{2}}\right)^2}} \cdot \varepsilon \|\mathbf{x}^*\|_{\mathbf{A}}. \tag{62}$$

□