Journal of Machine Learning Research 27 (2026) 1-48 Submitted 6/24; Revised 9/25; Published 5/26

Online Generalized Sparse Regression: How Does Overparametrization Help?

Shuoguang Yang YANGSG@UST.HK
The Hong Kong University of Science and Technology

Qiang Sun* QSUNSTATS@GMAIL.COM
University of Toronto & MBZUAI

Editor: Ji Zhu

Abstract

Regularized sparse regression has been extensively studied in the offline setting, but online formulation remains relatively under-explored. This gap stems from four key challenges: (i) the infeasibility of dynamically updating the regularization parameter in every online round, (ii) managing storage and memory complexity, (iii) enabling real-time computation via *closed-form updates* rather than solving full optimization problems at each round, and (iv) achieving optimal statistical guarantees under realistic assumptions. In this paper, we propose an online generalized-sparsity-constrained regression framework, focusing on online cardinality-constrained linear regression and low-rank matrix sensing. Unlike online regularized regression, our constrained formulation eliminates the need for dynamic parameter tuning. We introduce an efficient online hard-thresholding algorithm that performs closed-form updates and requires storing only summary statistics, making it computationally, memory, and storage efficient. Despite the inherent nonconvexity and combinatorial nature of the formulation, our algorithm achieves global convergence at the optimal statistical rate under realistic assumptions, provided that the projection set is properly overparameterized. Numerical experiments demonstrate that our method consistently outperforms state-of-the-art alternatives.

Keywords: cardinality constraints, generalized sparsity, online hard thresholding, optimality, rank constraints, streaming data

1. Introduction

Modern data analysis often involves a massive number of features, which poses both statistical and computational challenges. A common assumption is that the underlying signals are sparse, a property believed to hold in many applications such as genome-wide association studies, image processing, signal processing, and medical imaging. To exploit this sparse structure, a popular approach is to adopt the penalized empirical risk minimization approach:

$$\widehat{\Theta}^{\text{pen}} = \underset{\Theta\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}} \left\{\mathcal{L}(\Theta) + \mathcal{R}(\Theta;\lambda)\right\}, \tag{1}$$

*. Corresponding author.

where $\mathcal{L}(\cdot)$ is a loss function, $\Theta \in \mathbb{R}^{d_1 \times d_2}$ is a matrix or a vector ($d_2 = 1$), $\mathcal{R}(\Theta; \lambda)$ is a sparsity-inducing regularizer such as the LASSO penalty (Tibshirani, 1996) or the nuclear norm penalty (Recht et al., 2010), and λ is a regularization parameter. Both statistical and computational aspects of this penalized approach have been studied extensively in the offline setting, where all data are readily accessible at once (Agarwal et al., 2012; Loh and Wainwright, 2015; Fan et al., 2018b). By contrast, sparse regression in the online setting, where data arrive sequentially, has received relatively little attention.

Consider the online setting, where in each round $t \geq 1$, we receive a label-feature pair $(Y_t, X_t) \in \mathbb{R} \times \mathbb{R}^{d_1 \times d_2}$. Our goal is to construct an estimator Θ_t in real time to estimate the ground-truth coefficient matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$. Although online regression and stochastic optimization have been studied extensively (Kushner and Yin, 1997; Rakhlin et al., 2012; Fan et al., 2018a), extending the offline sparse regression (1) to the online setting is challenging due to presence of the regularizer $\mathcal{R}(\Theta, \lambda)$. The main challenges are fourfold:

1. **Infeasibility of dynamic regularization.** Achieving optimal statistical performance at each round requires dynamically updating the regularization parameter λ, which depends on both the unknown error distribution and the growing sample size. Manually tuning λ in every round is often infeasible in practice.

2. **Memory and storage complexity.** Naively storing all past data up to round t leads to $\mathcal{O}(td_1d_2)$ storage and memory costs, which are impractical for large-scale settings.

3. **Requirement for real-time computation.** Parameter updates must be computationally simple to enable real-time operation, especially in large-scale applications.

4. **Realistic assumptions.** Achieving optimal statistical guarantees under realistic assumptions, such as allowing arbitrary regularized/sparse condition numbers, is essential for practical deployment.

Several existing methods have attempted to address these challenges, but none simultaneously resolve all four. For instance, Kale et al. (2017) proposed an online sparse linear regression algorithm that periodically solves a randomized Dantzig-selector linear program (Candès and Tao, 2007) and then applies a sparse projection; the linear program is solved when the round index is a power of two. The method also stores an estimated design matrix whose number of rows grows with t, so it does not provide bounded storage. Moreover, their results require a bound on the restricted isometry property constant $\leq 1/5$, implying a sparse condition number $\leq 3/2$, which is restrictive, since real-world high-dimensional datasets can have much larger condition numbers. Steinhardt et al. (2014) proposed a streaming sparse-regression method with stochastic-gradient-like computational and memory requirements, but its support-recovery theory requires an irrepresentability condition (Zhao and Yu, 2006). Yang et al. (2023) introduced a new loss function and a memory-efficient online linearized LASSO algorithm that allows arbitrary condition numbers, but each round is still defined by solving an ℓ_1-regularized optimization problem rather than by a fixed number of closed-form updates. This naturally raises the question:

Can we design online algorithms that are storage- and memory-efficient, admit simple updates at each round, and achieve optimal statistical convergence guarantees under reasonable assumptions, without requiring dynamic regularization?

This paper provides a positive answer to the question above. Instead of a regularized approach, we propose the following online constrained sparse regression framework:

$$\widehat{\Theta}_t \in \underset{\Theta \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \{\mathcal{L}_t(\Theta) : \Upsilon(\Theta) \leq s^*\} \quad \text{for} \quad t \geq 1, \tag{2}$$

where $\Upsilon(\Theta)$ represents the *generalized sparsity* of Θ, and s^* denotes the unknown true generalized sparsity of the ground-truth coefficient matrix Θ^*. We focus on problems: (i) online cardinality-constrained linear regression, where $\Theta \in \mathbb{R}^d$ is a vector and $\Upsilon(\Theta) = \|\Theta\|_0$, the number of non-zero entries; (ii) online rank-constrained matrix sensing, where $\Theta \in \mathbb{R}^{d_1 \times d_2}$ is a matrix and $\Upsilon(\Theta) = \operatorname{rank}(\Theta)$. A solution to problem (i) is called *sparse* if it has a small cardinality, and a solution to problem (ii) is called *low-rank* if it has a small rank.

One immediate advantage of formulating online generalized sparse regression in this constrained form is that it eliminates the need to dynamically tune a regularization parameter such as λ in (1). The drawback is also apparent: the resulting generalized-sparsity-constrained problem is combinatorial and typically NP-hard (Natarajan, 1995; Foster et al., 2015). Assuming natural sparse strong convexity and smoothness conditions, which often hold in practice, Jain et al. (2014) showed that projected gradient descent with an enlarged projection set $\{\Theta : \|\Theta\|_0 \leq s = 32\kappa^2 s^*\}$ and a strict step size $\eta = 2/(3L)$ converges to the underlying Θ^* in function value, where κ and L are the sparse condition number and smoothness parameter, respectively[1]. However, efficient online algorithms that admit simple updates and provable statistical guarantees remain unavailable, particularly when the loss function changes across rounds. Additionally, strictly requiring a step size of $2/(3L)$ may be overly restrictive.

We propose an online hard thresholding (OHT) algorithm to solve (2). Instead of solving each subproblem exactly, we perform K hard thresholding steps in each online round:

$$\Theta_{t,k+1} = \Pi_{\mathcal{C}_s}\left(\Theta_{t,k} - \eta_{t,k} \nabla \mathcal{L}_t(\Theta_{t,k})\right), \quad \text{for } k = 0, 1, \cdots, K-1, \tag{3}$$

where $\Pi_{\mathcal{C}_s}(\cdot)$ denotes the projection operator onto the set $\mathcal{C}_s$ and $\Theta_0 = \Theta_{0,0}$ is the initialization point. When $\Upsilon(\Theta) = \|\Theta\|_0$, $\Pi_{\mathcal{C}_s}(\cdot)$ performs hard thresholding by retaining the s entries with the largest magnitude and setting the rest to zero. When $\Upsilon(\Theta) = \operatorname{rank}(\Theta)$, $\Pi_{\mathcal{C}_s}(\cdot)$ returns the best rank-s approximation.

Assuming generalized sparse strong convexity and smoothness, and setting the step size $\eta_{t,k} = \eta \leq 1/L$ with L being the sparse strong smoothness parameter, our algorithm with an overparameterized projection set $s > \kappa^2 s^*$ converges geometrically to the ground-truth coefficients, up to the optimal statistical error of $s\sigma_x^2/t$, where σ_x^2 is a constant depending on the design and noise. Informally, we have

$$\mathbb{E}\left[\|\Theta_{t,K} - \Theta^*\|_2^2\right] \lesssim \underbrace{\delta^t \cdot \mathbb{E}\left[\|\Theta_0 - \Theta^*\|_2^2\right]}_{\text{geometric convergence}} + \underbrace{\frac{s^* \sigma_x^2}{t}}_{\text{opt. stat. error}}, \quad \text{for some } \delta \in (0,1). \tag{4}$$

1. Sparse condition number, sparse strong convexity, and smoothness parameters are defined formally in Section 3.

For logarithmically large round t, our algorithm achieves the optimal statistical error, as if one had solved the offline constrained program had been solved exactly using all data up to round t. Moreover, our algorithm achieves constant memory and storage complexity: $\mathcal{O}(d_1^2)$ for linear regression and $\mathcal{O}(d_1^2 d_2^2)$ for matrix sensing. Crucially, these complexities do not scale with the number of online rounds t. The result holds for arbitrarily large sparse condition numbers.

Remarkably, this is the first algorithm to simultaneously overcome all major challenges: eliminating the need for dynamic regularization, ensuring storage and memory efficiency, admitting closed-form per-round updates, and achieving optimal statistical guarantees under realistic assumptions. To establish generalized sparse strong convexity and smoothness, we initialize with a batch of samples and then stream on top of it; we also extend our results to scenarios where no such initial batch is available.

1.1 Related Work

We review additional works related to ours. Online regularized sparse optimization has been studied in several works such as Bertsekas (2011); Duchi et al. (2011); Xiao (2010), which analyzed the convergence behavior of various online methods. However, these studies assumed a fixed regularization parameter, and thus their results are not applicable to our setting. Han et al. (2024b) proposed a one-pass debiased stochastic gradient descent method for online confidence intervals. Its goal is statistical inference rather than maintaining a sparse estimator under a changing regularization path, so it does not address our constrained sparse-optimization problem. Recently, Fan et al. (2018a) proposed a two-stage algorithm that first identifies the support set by solving a penalized offline problem during a burn-in stage, and then performs truncated gradient descent restricted to that support. Their method, however, requires both a minimum signal strength assumption and a sufficiently large burn-in sample size, while our method does not rely on either. Our work is partly motivated by Liu and Foygel Barber (2020), who studied optimal thresholding algorithms for offline sparse linear regression. There are, however, two key differences: (i) we consider an online setting where the loss function changes from round to round, and (ii) we establish a new descent lemma and leverage it to provide last-iterate guarantees, from which we derive the optimal statistical rate for our algorithm. By contrast, Liu and Foygel Barber (2020) did not provide last-iterate guarantees, and hence their results cannot be applied in the online setting. Taken together, our work introduces the first algorithm that simultaneously addresses all four challenges: it eliminates the need for dynamic regularization, ensures storage and memory efficiency, admits closed-form updates, and achieves optimal statistical guarantees under realistic assumptions.

1.2 Notation

We summarize here the notation that will be used throughout the paper. We use c and C to denote generic constants which may change from line to line. For two sequences of real numbers $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, we write $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ if there exists some constant $C > 0$ such that $a_n \leq C b_n$ for all $n \geq 1$. We use $a_n \asymp b_n$ to denote $a_n \gtrsim b_n$ and $a_n \lesssim b_n$. The log operator is understood to be with respect to the base e. For a function $f(x)$, we use $\nabla f(x)$ to denote its derivative. For a vector u and any $p \geq 1$, we use $\|u\|_p$ to denote its p-th

norm, use $\|u\|_0$ to denote the cardinality of u, and use $\|a\|_\infty = \max_{j\le d}|a_j|$ to denote the maximum of absolute entries of $a \in \mathbb{R}^d$. For a matrix Θ, we use $\|\Theta\|_2$ to denote its spectral norm and use $\|\Theta\|_{\mathrm{F}}$ to denote its Frobenius norm, $\mathrm{tr}(\Theta)$ denotes its trace. We shall note that the Frobenius norm coincides with the ℓ_2 norm when Θ is a vector. For any vector $x \in \mathbb{R}^d$, we denote by $x_{\mathcal{S}} \in \mathbb{R}^d$ a vector whose entries within $\mathcal{S}$ are the same as those in x but the rest entries are all zeros. For any matrix $X \in \mathbb{R}^{d_1\times d_2}$, we denote by $X_{\mathcal{S}} \in \mathbb{R}^{d_1\times d_2}$ a matrix such that the columns within $\mathcal{S}$ are the same as those in X but the rest entries are all zeros. We use $\mathrm{vec}(X) \in \mathbb{R}^{d_1 d_2}$ to denote the vectorized representation of a matrix $X \in \mathbb{R}^{d_1\times d_2}$. We denote by $\mathcal{S}^*$ the support of the underlying coefficient Θ^*. For any two linear spaces S_1 and S_2, we denote by $S_1 + S_2 = \{x_1 + x_2 \mid x_1 \in S_1, x_2 \in S_2\}$. For any matrix A and any linear space S with the basis matrix being U (i.e., U consists of the basis vectors as its columns), let $P_S(A) = UU^\top A$, that is, P_S denotes the operator that projects the columns of A onto the linear space S. Constants C and c may vary from line to line.

2. Why Overparameterized Online Hard Thresholding Works

In this section, we formally define the problem and highlight the challenges involved in extending offline penalized regression to the online setting. We also discuss how overparameterization can help overcome the difficulty of solving a seemingly NP-hard problem. Finally, we introduce our proposed algorithm and provide an analysis of its runtime complexity.

2.1 Problem Setup

We consider a scenario where independent and identically distributed (i.i.d.) data arrive sequentially and are generated according to the following realizable linear model

$$Y_j = \langle X_j, \Theta^*\rangle + \epsilon_j,\ j \ge 1, \tag{5}$$

where $Y_j \in \mathbb{R}$ is the response, $X_j \in \mathbb{R}^{d_1\times d_2}$ is the covariate, ϵ_j is the random error, and $\Theta^* \in \mathbb{R}^{d_1\times d_2}$ is the ground-truth coefficient matrix with a generalized sparse structure. When $d_2 = 1$, the matrices X_j and Θ^* reduce to vectors. We assume that Θ^* is generalized s^*-sparse, i.e., $\Upsilon(\Theta^*) = s^*$, which corresponds to cardinality s^* in sparse linear regression and rank s^* in low-rank matrix regression.

We focus on the squared loss, denoting by ℓ_j the loss for the j-th data point:

$$\ell_j(\Theta) := \ell\Big(Y_j, \langle\Theta, X_j\rangle\Big) = \frac{1}{2}\Big(Y_j - \langle\Theta, X_j\rangle\Big)^2. \tag{6}$$

Suppose we have access to an initial batch of sample size t_0 consisting of i.i.d. samples $(X_{0_j}, Y_{0_j})_{1\le j\le t_0}$. Then the loss for the initial batch and the cumulative loss up to (including) round t are

$$\ell_0(\Theta) = \frac{1}{2}\sum_{j=1}^{t_0}\Big(Y_{0_j} - \langle\Theta, X_{0_j}\rangle\Big)^2 \ \text{ and } \ \mathcal{L}_t(\Theta) = \frac{1}{t_0+t}\Big(\ell_0(\Theta) + \sum_{j=1}^{t}\ell_j(\Theta)\Big), \ \text{ respectively.}$$

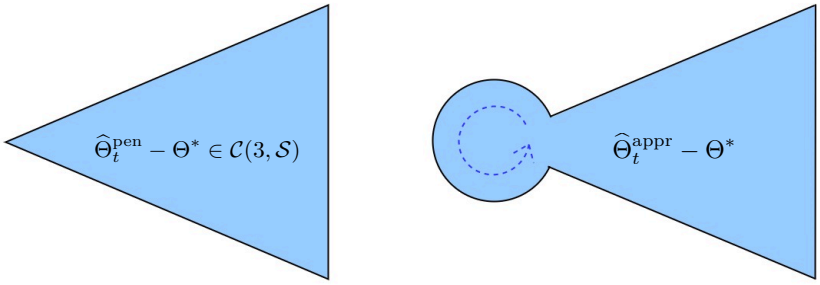

Figure 1: The left panel shows the cone-like set $\mathcal{C}(3, \mathcal{S})$ while the right panel depicts the region where $\widehat{\Theta}_t^{\text{appr}} - \Theta^*$ may fall.

2.2 Why Extending Offline Penalized Regression to the Online Setting Is Challenging

We focus on the linear regression problem with $d = d_1$ and $d_2 = 1$ in (5) as an illustration, though the same reasoning applies to low-rank matrix sensing. A natural online extension of the offline penalized regression (1) is

$$\widehat{\Theta}_t^{\text{pen}} = \underset{\Theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \mathcal{L}_t(\Theta) + \lambda_t \|\Theta\|_1 \right\}, \tag{7}$$

where $\|\Theta\|_1 = \sum_{j=1}^d |\Theta_j|$ is the vector ℓ_1 norm.

As discussed earlier, the main challenges in the online setting are: (1) the infeasibility of dynamic regularization, (2) managing memory and storage limitations, and (3) the requirement for real-time computation. Previous works (Han et al., 2024a; Yang et al., 2023) adopted related regularized streaming formulations and addressed the second challenge by storing only summary statistics. However, these approaches update regularization or tuning parameters over time, which becomes difficult especially when the errors are heterogeneous. In the case of i.i.d. errors, one can make λ_t adaptive to the sample size, reducing dynamic regularization to careful tuning of the standard error in early rounds. Moreover, these estimators are defined through regularized optimization problems at successive rounds or data batches. A naive alternative is to perform a few simple updates, such as sub-gradient steps, and use the last iterate as the solution for each round. Yet, this approach does not guarantee the desired statistical properties. We illustrate this issue below.

We denote the approximate solutions obtained after a few simple updates as $\widehat{\Theta}_t^{\text{appr}}$. Since the objective function in (7) changes from round to round, the problem is a moving target optimization problem, where the updates must converge sufficiently fast, ideally at a geometric rate, to keep up. However, geometric convergence is only guaranteed within a cone-like set

$$\mathcal{C}(3, \mathcal{S}) = \{\Theta : \|(\Theta - \Theta^*)_{\mathcal{S}^c}\|_1 \leq 3\|(\Theta - \Theta^*)_{\mathcal{S}}\|_1, \, \mathcal{S} \text{ is the support of } \Theta^*\}.$$

It is known that the exact solution at round t falls into the above set whenever $\lambda_t \geq 2\|\nabla \mathcal{L}_t(\Theta^*)\|_\infty$ (Yang et al., 2023, Lemma 3.1). In contrast, approximate solutions need not

remain in this set. In fact, $\widehat{\Theta}_t^{\text{appr}} - \Theta^*$, obtained from small perturbations of $\widehat{\Theta}_t^{\text{pen}} - \Theta^*$, can drift outside $\mathcal{C}(3, \mathcal{S})$ and instead form a neighborhood around it; see Figure 1. Once outside the set, the restricted strong convexity (Raskutti et al., 2010), fails to hold for $\widehat{\Theta}_t^{\text{appr}} - \Theta^*$ because it runs out of $\mathcal{C}(3, \mathcal{S})$, i.e., there may NOT exist a $\kappa > 0$ such that

$$\langle \nabla \mathcal{L}_t(\widehat{\Theta}_t) - \nabla \mathcal{L}_t(\Theta^*), \widehat{\Theta}_t^{\text{pen}} - \Theta^* \rangle \geq \kappa \, \|\widehat{\Theta}_t^{\text{pen}} - \Theta^*\|_2^2.$$

This leads to non-identifiability of approximate solutions, preventing geometric convergence or even desirable statistical guarantees. Related online sparse-regression procedures therefore rely on repeated global optimization: Kale et al. (2017) solve a linear program at periodic rounds, Yang et al. (2023) solve an ℓ_1-regularized problem at every round, and Han et al. (2024a) solve successive regularized estimating problems as data batches arrive. Moreover, dynamically updating λ_t is practically difficult, as its optimal value depends on unknown problem-dependent quantities. By contrast, online constrained optimization sidesteps this issue entirely, as it does not require dynamically updated regularization parameters.

2.3 Moving to Online Constrained Regression

We now turn to online constrained regression, which circumvents both the challenges of dynamic regularization and the violation of the restricted strong convexity condition. Upon receiving the t-th data point in round t, a straightforward approach is to run offline generalized-sparsity-constrained least squares to estimate the regression coefficients Θ^*:

$$\widehat{\Theta}_t = \underset{\Theta \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \left\{ \ell_0(\Theta) + \sum_{j=1}^{t} \ell_j(\Theta) \right\} \text{ s.t. } \Upsilon(\Theta) \leq s^*, \tag{8}$$

where $\Upsilon(\Theta) \leq s^*$ imposes a hard constraint on the generalized sparsity of Θ^*.

Two main challenges arise: (1) due to storage and memory limitations, it is infeasible to retain all past data to optimize (8) directly; and (2) solving cardinality-constrained regression is NP-hard in general (Natarajan, 1995; Foster et al., 2015). Challenge (1) can be addressed by maintaining suitable summary statistics, which we introduce later. To address (2), we first consider the case of linear regression. While Foster et al. (2015) established strong computational lower bounds for sparse linear regression under standard complexity assumptions, efficient globally convergent algorithms are possible when the design matrix is well-conditioned.

Now let us examine the optimization problem in the t-th round more carefully:

$$\widehat{\Theta}_t = \underset{\|\Theta\|_0 \leq s^*}{\operatorname{argmin}} \frac{\|\mathbb{Y} - \mathbb{X}\Theta\|_2^2}{t + t_0},$$

where $\mathbb{Y}$ consists of all labels and the rows of $\mathbb{X}$ consist of all covariates up to round t (we omit the dependence on t). As a warm-up, consider the special case where the feature matrix $\mathbb{X}$ is orthonormal, i.e., $\mathbb{X}^{\mathrm{T}}\mathbb{X} = (t + t_0)I$. In this setting, $\widehat{\Theta}_t$ has a closed form:

$$\widehat{\Theta}_t = \Pi_{\mathcal{C}_{s^*}} \left(\widehat{\Theta}^{\text{ols}} \right) = \Pi_{\mathcal{C}_{s^*}} \left((\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1} \mathbb{X}^{\mathrm{T}}\mathbb{Y} \right),$$

where $\mathcal{C}_{s^*}$ is the set of d-dimensional vectors with at most s^* nonzeros, and $\Pi_{\mathcal{C}_{s^*}}$ is the projection operator onto this set $\mathcal{C}_{s^*}$. In other words, $\widehat{\Theta}_t$ can be obtained by first computing the ordinary least square estimator and then hard thresholding to retain the s^* largest entries in magnitude.

Equivalently, $\widehat{\Theta}_t$ can be obtained as the limiting solution of the iterative hard thresholding algorithm:

$$\Theta_{t,k+1} = \Pi_{\mathcal{C}_{s^*}} \left(\Theta_{t,k} - \eta \cdot \frac{\mathbb{X}^{\mathrm{T}} \left(\mathbb{X}\Theta - \mathbb{Y} \right)}{t + t_0} \right), \tag{9}$$

starting from $\Theta_{t,0} = 0$, where η is the step size. Although this derivation assumes orthonormal designs, it suggests that this iterative algorithm may still converge to Θ^* under more general conditions.

Indeed, by assuming the restricted isometry property (RIP) constant $\delta_{3s^*} \leq 1/\sqrt{32}$, which bounds the sparse condition number to ≤ 1.43 and allows slightly more general designs than orthonormal ones, the seminal work by Blumensath and Davies (2009) proved that the iterative hard thresholding algorithm (9) with $\eta = 1$ recovers an approximate $\Theta_{t,k}$ satisfying

$$\|\Theta_{t,k} - \Theta^*\|_2 \leq \frac{6\|\epsilon\|_2}{\sqrt{t + t_0}}$$

for sufficiently large k, where $\epsilon \in \mathbb{R}^{t+t_0}$ is the noise vector[2].

For the realizable model (5) with i.i.d. data points, a slightly modified analysis yields, for $k \gtrsim \log\left(\|\Theta_{t,0} - \Theta^*\|_2 / \sqrt{(s \log d)/(t + t_0)} \right)$:

$$\begin{aligned} \|\Theta_{t,k} - \Theta^*\|_2 &\leq \frac{1}{2} \|\Theta_{t,k-1} - \Theta^*\|_2 + \frac{2\|\mathbb{X}_{\mathcal{S}_k}^{\mathrm{T}} \epsilon\|_2}{t + t_0} \\ &\leq 2^{-k} \|\Theta_{t,0} - \Theta^*\|_2 + \frac{4\sqrt{s} \cdot \|\mathbb{X}^{\mathrm{T}} \epsilon\|_\infty}{t + t_0} \\ &\lesssim \sqrt{\frac{4s \log d}{t + t_0}} \end{aligned}$$

where the last inequality holds if $\|\mathbb{X}^{\mathrm{T}} \epsilon\|_\infty \lesssim \sqrt{(t + t_0) \log d}$, a standard high-dimensional score bound (Fan et al., 2018b).

However, the condition number bound ≤ 1.43 is very restrictive, and real-world high-dimensional data often exhibit much larger condition numbers due to strongly correlated features. This motivates the natural question:

> *Is it possible to modify the hard thresholding algorithm to allow potentially arbitrary condition numbers while still achieving last-iterate convergence?*

2. As pointed out by Blumensath and Davies (2009), the step size is taken to be $\eta = 1$ for simplicity; using a more general step sizes require the RIP constant to depend on the choice of η.

Algorithm 1 An online hard thresholding algorithm.

Require: Θ_0, $\{\eta_{t,j}\}$, s, K.
Ensure: Θ_T

Initialization Phase: Generate t_0 samples
for $t = 1, \cdots, T$ **do**
  Generate the t-th sample outside the initialization phase. Set $\Theta_{t,0} = \Theta_{t-1}$.
  for $k = 0, \cdots, K-1$ **do**
    Update $\Theta_{t,k+1} = \Pi_{\mathcal{C}_s}\left\{\Theta_{t,k} - \eta_{t,k}\nabla\mathcal{L}_t(\Theta_{t,k})\right\}$.
  end for
  Set $\Theta_t = \Theta_{t,K}$.
end for

2.4 An Overparameterized Online Hard Thresholding Algorithm

Designing algorithms for the online case requires answering the above question first. Our key intuition is that when the linear system has a large condition number, restricting the projection set in (9) to exactly size s^* may be too limiting. A small projection set can miss true coefficients due to strong correlations among features. To address this, we overparameterize the projection set by using $\mathcal{C}_{s(\kappa,s^*)}$ instead of $\mathcal{C}_{s^*}$ in each hard-thresholding step, where $s = s(\kappa, s^*) \geq s^*$ depends on the condition number κ and the true sparsity s^*. The intuition is that a larger condition number warrants a larger projection set. This approach allows including a few additional features beyond the ground truth, mitigating the risk of missing important coefficients while keeping the estimation error nearly unaffected.

We now introduce our online hard thresholding (OHT) algorithm. First, we formally define the projection operator. Let $\mathcal{C}_s$ be the set of Φ that is at most generalized s-sparse:

$$\mathcal{C}_s = \{\Phi : \Upsilon(\Phi) \leq s\},$$

where $\mathcal{C}_s$ is nonconvex and represents the set of at most s-sparse vectors for cardinality-constrained linear regression, or the set of at most rank-s matrices for rank-constrained matrix sensing. The projection operator $\Pi_{\mathcal{C}_s}$ is defined as

$$\Pi_{\mathcal{C}_s}(\Theta) := \underset{\Phi\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}} \left\{\|\Phi - \Theta\|_{\mathrm{F}}^2 \mid \Phi \in \mathcal{C}_s\right\},$$

which returns the best generalized s-generalized sparse approximation of Θ. Concretely, this reduces to keeping the s largest entries of Θ in magnitude and truncating the rest to zero for cardinality-constrained optimization, and to returning the rank-s approximation for rank-constrained matrix sensing.

Our OHT algorithm is as follows. Upon receiving each observation (X_t, Y_t) in round $t \geq 0$, OHT performs K iterations of overparameterized hard thresholding with sparsity $s = s(\kappa, s^*)$:

$$\Theta_{t,j+1} = \Pi_{\mathcal{C}_s}\left\{\Theta_{t,j} - \eta_{t,j}\nabla\mathcal{L}_t(\Theta_{t,j})\right\}, \quad 0 \leq j \leq K-1 \tag{10}$$

where $\Theta_{t,0} = \Theta_{t-1,K}$ is the last iterate from the previous round, $\Theta_{0,0}$ is some initialization, and the gradient is

$$\nabla \mathcal{L}_t(\Theta) = \frac{1}{t+t_0} \sum_{i \in \mathcal{I}_t} X_i \left(\langle X_i, \Theta \rangle - Y_i \right), \tag{11}$$

with $\mathcal{I}_t = \{i : 0 \le i \le t_0\} \cup \{j : 1 \le j \le t\}$ indexing all data points up to round t. Algorithm 1 summarizes the details.

The main technical challenge in analyzing the sequence $\{\Theta_{t,k} : t \ge 0, k \ge 0\}$ comes from the nonconvex constraint set $\mathcal{C}_s$, which prevents the direct use of classical nonconvex optimization analysis. For comparison, if we replace $\mathcal{C}_s$ with a convex set $\mathcal{X}$, the projected gradient descent update

$$\Theta_{t,k+1} = \underset{\Theta \in \mathcal{X}}{\operatorname{argmin}} \left\{ \|\Theta - (\Theta_{t,k} - \eta_{t,k} \nabla \mathcal{L}_t(\Theta_{t,k}))\|_2^2 \right\}, \ 0 \le k \le K-1 \tag{12}$$

satisfies the first-order optimality condition:

$$\langle \eta_{t,k} \nabla \mathcal{L}_t(\Theta_{t,k}) + \Theta_{t,k+1} - \Theta_{t,k}, \Theta - \Theta_{t,k+1} \rangle \ge 0, \ \forall \ \Theta \in \mathcal{X}. \tag{13}$$

By setting $\Theta = \Theta_{t,k}$, we obtain

$$\eta_{t,k} \langle \nabla \mathcal{L}_t(\Theta_{t,k}), \Theta_{t,k+1} - \Theta_{t,k} \rangle \le -\|\Theta_{t,k+1} - \Theta_{t,k}\|_2^2.$$

Using the smoothness of $\mathcal{L}_t$, the function value satisfies

$$\begin{aligned} \mathcal{L}_t(\Theta_{t,k+1}) - \mathcal{L}_t(\Theta_{t,k}) &\le \langle \nabla \mathcal{L}_t(\Theta_{t,k}), \Theta_{t,k+1} - \Theta_{t,k} \rangle + \frac{L}{2} \|\Theta_{t,k+1} - \Theta_{t,k}\|_2^2 \\ &\le -\left(\frac{1}{\eta_{t,k}} - \frac{L}{2} \right) \|\Theta_{t,k+1} - \Theta_{t,k}\|_2^2 \le 0, \end{aligned} \tag{14}$$

for $\eta_{t,k} \le 2/L$, where L is the sparse strong smoothness parameter.

However, this argument fails for the nonconvex set $\mathcal{C}_s$, requireing a new understanding of the hard thresholding operator. Specifically, we establish descent lemmas, i.e., Lemmas 3 and 7:

$$\mathcal{L}_t(\Theta_{t,k+1}) - \mathcal{L}_t(\Theta_{t,k}) \le -\frac{\eta_{t,k}(1 - L\eta_{t,k})}{2} \|(\nabla \mathcal{L}_t(\Theta_{t,k}))_{\mathcal{S}_{t,k} \cup \mathcal{S}_{t,k+1}}\|_2^2 \le 0$$

for $\eta_{t,k} \le 1/L$. This new understanding enables a recursive relationship between outputs from successive online rounds, leading to the final convergence result (4).

To summarize, Algorithm 1 simultaneously addresses the four main challenges:

1. No dynamic regularization: The algorithm does not require updating any regularization parameter dynamically. Once the projection sparsity s is chosen, it remains fixed throughout all rounds.

2. Storage and memory efficiency: Each update only requires the previous solution Θ_{t-1} and the gradient $\nabla \mathcal{L}_t(\Theta_{t-1})$. For linear regression, we maintain $\sum_{i \in \mathcal{I}_t} X_i X_i^\top$ and $\sum_{i \in \mathcal{I}_t} X_i Y_i$; for matrix sensing, we store $\sum_{i \in \mathcal{I}_t} \mathrm{vec}(X_i)\mathrm{vec}(X_i)^\top$ and $\sum_{i \in \mathcal{I}_t} X_i Y_i$, with memory costs $\mathcal{O}(d^2)$ and $\mathcal{O}(d_1^2 d_2^2)$, respectively.

3. Real-time computation: Each round only requires a few projected gradient steps.

4. Realistic assumptions: We allow arbitrary sparse condition numbers.

Finally, the projection sparsity $s = s(\kappa, s^*)$ can be chosen as small as $s > \kappa^2 s^*$ when $\eta_{t,k} = 1/L$. For orthonormal designs ($\kappa = 1$), this reduces to $s > s^*$, which is nearly the minimal requirement to achieve good statistical accuracy, since any $s < s^*$ would miss true signals.

3. Global Convergence Analysis

In this section, we prove that our proposed online hard thresholding algorithm (Algorithm 1) achieves global convergence for both online sparse linear regression and low-rank matrix sensing. To begin, we assume that the loss functions $\mathcal{L}_t$ satisfy the following generalized sparse strong smoothness (GSSS) and generalized sparse strong convexity (GSSC) conditions with appropriate parameters.

Condition 1 (Generalized Sparse Strong Smoothness, GSSS)**.** *For each $t \geq 0$, the loss function $\mathcal{L}_t$ is (s_1, s_2)-generalized sparse strongly smooth, or (s_1, s_2)-GSSS for short, with parameter L_{s_1,s_2}, meaning that*

$$\mathcal{L}_t(\Theta') - \mathcal{L}_t(\Theta) \leq \left\langle \nabla \mathcal{L}_t(\Theta), \Theta' - \Theta \right\rangle + \frac{L_{s_1,s_2}}{2} \left\|\Theta' - \Theta\right\|_{\mathrm{F}}^2, \ \textit{for all } \Theta' \in \mathcal{C}_{s_1} \textit{ and } \Theta \in \mathcal{C}_{s_2}.$$

Condition 2 (Generalized Sparse Strong Convexity, GSSC)**.** *For each $t \geq 0$, the loss function $\mathcal{L}_t$ is (s_1, s_2)-generalized sparse strongly convex with parameter α_{s_1,s_2}, meaning that*

$$\mathcal{L}_t(\Theta') - \mathcal{L}_t(\Theta) \geq \left\langle \nabla \mathcal{L}_t(\Theta), \Theta' - \Theta \right\rangle + \frac{\alpha_{s_1,s_2}}{2} \left\|\Theta' - \Theta\right\|_{\mathrm{F}}^2, \ \textit{for all } \Theta' \in \mathcal{C}_{s_1} \textit{ and } \Theta \in \mathcal{C}_{s_2}.$$

These GSSS and GSSC conditions characterize the smoothness and convexity properties required for a wide range of problems. In the rest of this section, we will specify the generalized sparsity levels s_1 and s_2, the convexity parameter α_{s_1,s_2}, and the smoothness parameter L_{s_1,s_2} for both online sparse linear regression and online low-rank matrix sensing.

3.1 Online Sparse Linear Regression

We first investigate the performance of Algorithm 1 for online sparse linear regression, for which $d_2 = 1$ and we write $d = d_1$. Accordingly, we have $X_i \in \mathbb{R}^d$, $\Theta^* \in \mathbb{R}^d$, $\Upsilon(\Theta) = \|\Theta\|_0$, and

$$\Pi_{\mathcal{C}_s}(\Theta) = \operatorname*{argmin}_{\Phi \in \mathbb{R}^d} \Big\{ \|\Phi - \Theta\|_2^2 \mid \|\Phi\|_0 \leq s \Big\}, \tag{15}$$

which the hard thresholding operator that keeps the s largest entries of Θ in magnitude and truncates the rest to zeros.

For this problem, the GSSS and GSSC conditions reduce to the familiar sparse strong smoothness (SSS) and sparse strong convexity (SSC) conditions, respectively.

Assumption 1. *For all $t \geq 1$, the loss function $\mathcal{L}_t$ satisfies (s, s)-SSC and (s, s)-SSS conditions with parameters α and L, respectively. We define the sparse condition number as $\kappa = L/\alpha$.*

Assumption 1 is standard in the high-dimensional statistics literature (Bickel et al., 2009; Raskutti et al., 2010). By adapting the proofs of (Raskutti et al., 2010, Theorem 1 and Corollary 1), one can show that this assumption holds with high probability under Gaussian covariates, provided that $t_0 \geq cs \log d$ for some constant $c > 0$ and the relevant population sparse eigenvalues are bounded. For simplicity, we also assume that the random noises ϵ_i and covariates X_i satisfy the following standard assumption.

Assumption 2. *The random noises ϵ_i are mean-zero and have bounded second moments $\mathbb{E}[\epsilon_i^2] \leq \sigma^2$. Moreover, the scaled sum of noise-weighted covariates satisfies*

$$\mathbb{E}\left[\Big\| \sum_{i \in \mathcal{I}_t} \frac{\epsilon_i X_i}{\sqrt{t+t_0}} \Big\|_\infty^2\right] \leq \sigma_x^2.$$

The following result shows that $\sigma_x^2 = \mathcal{O}(\log d)$ when both X_i and ϵ_i are sub-Gaussian[3].

Proposition 1. *Suppose the coordinates of X_i are sub-Gaussian, $X_{ij} \sim subG(\sigma_1^2)$, and $\mathbb{E}\epsilon_i^2 \leq \sigma^2$. Then*

$$\mathbb{E}\left[\Big\| \sum_{i \in \mathcal{I}_t} \frac{\epsilon_i X_i}{\sqrt{t+t_0}} \Big\|_\infty^2\right] \leq C\sigma^2\sigma_1^2 \log d,$$

where C is some universal constant.

With these assumptions in place, we now characterize the performance of the best achievable estimator in each online round. Suppose we have received t online data points after the initial batch. Recall that $\widehat{\Theta}_t$ in (8) is the optimal s^*-sparse solution under the loss function $\mathcal{L}_t(\cdot)$. This estimator serves as the moving target that a globally convergent algorithm aims to track. Upon receiving an additional observation (X_{t+1}, Y_{t+1}), the loss function updates to $\mathcal{L}_{t+1}(\cdot)$, and the target estimator becomes $\widehat{\Theta}_{t+1}$. This gives rise to a moving-target optimization problem.

Lemma 2. *Suppose Assumptions 1 and 2 hold. Then, for all $t \geq 1$,*

$$\mathbb{E}\left[\|\widehat{\Theta}_t - \Theta^*\|_2^2\right] \leq \frac{8s^*\sigma_x^2}{\alpha^2(t+t_0)}.$$

The above expected mean squared error bound can be interpreted as the minimum statistical error achievable by any estimator under the loss function $\mathcal{L}_t(\cdot)$, given the data available up to round t. Intuitively, in round t, no estimator can outperform $\widehat{\Theta}_t$, since only $t + t_0$ observations are available. Having established this benchmark, we now analyze the performance of our OHT algorithm for online sparse linear regression, partially motivated by the analysis of the offline iterative hard thresholding algorithm in Liu and Foygel Barber (2020). For any s-sparse initialization $\Theta_{t,0}$ in round t, Liu and Foygel Barber (2020) established a linear convergence result under the projection sparsity level $s > \kappa^2 s^*$:

$$\min_{k=1,\cdots,K} \mathcal{L}_t(\Theta_{t,k}) - \mathcal{L}_t(\Theta^*) \leq \frac{L}{2}\left(\frac{1-\kappa}{1-2\gamma_{s,s^*}}\right)^K \|\Theta_{t,0} - \Theta^*\|_2^2, \tag{16}$$

3. A mean-zero random variable Z is said to be subG(σ^2) if the ψ_2 Orlicz norm of Z is bounded by σ, that is, $\|Z\|_{\psi_2} \leq \sigma$, where $\psi_2(x) = e^{x^2/t^2} - 1$.

where γ_{s,s^*} is a constant defined later, and the step size is taken as $1/L$. Actually, Liu and Foygel Barber (2020)'s result (16) can be easily extended to $\eta \geq 1/L$ but not $\eta < 1/L$, whereas our new results later hold for any $\eta \leq 1/L$. The requirement $\eta \leq 1/L$ is necessary for our descent lemma, Lemma 3, to hold.

However, the online and offline settings differ in ways that make extending this convergence result challenging:

(i) **Last-iterate convergence:** The result (16) does not guarantee last-iterate convergence. In the online setting, the output of round t serves as the initialization for round $t+1$. Without last-iterate guarantees for $\Theta_{t,K}$, establishing a recursive relationship between consecutive rounds is impossible.

(ii) **Moving target:** The offline result assumes a fixed loss, while in the online setting, $\mathcal{L}_t$ changes with each incoming observation. This leads to different target estimators $\widehat{\Theta}_t$ in each round, requiring a new analysis to understand how the moving target affects convergence.

(iii) **Metric mismatch:** The left and right sides of (16) involve different metrics, which do not naturally form a recursive relationship across multiple online rounds.

These differences necessitate a new understanding of the hard thresholding operator and a novel analysis to establish theoretical convergence guarantees for Algorithm 1 in the online setting.

To overcome challenge (i), we need to establish a last-iterate convergence result, which in turn requires a deeper understanding of the hard thresholding operator. To this end, we introduce a new descent lemma that guarantees the objective value decreases after each hard thresholding step.

Lemma 3 (Descent Property)**.** *Suppose $\mathcal{L}_t$ is (s,s)-SSS with smoothness parameter L, and the step sizes are set as $\eta_{t,k} \leq 1/L$. Let $\mathcal{S}_{t,k}$ and $\mathcal{S}_{t,k+1}$ denote the supports of $\Theta_{t,k}$ and $\Theta_{t,k+1}$ respectively. Then*

$$\mathcal{L}_t(\Theta_{t,k+1}) - \mathcal{L}_t(\Theta_{t,k}) \leq -\frac{\eta_{t,k}(1 - L\eta_{t,k})}{2} \|(\nabla \mathcal{L}_t(\Theta_{t,k}))_{\mathcal{S}_{t,k} \cup \mathcal{S}_{t,k+1}}\|_2^2 \leq 0.$$

It is worth emphasizing that this descent result holds for any projection sparsity level $s \geq 1$ and any loss function $\mathcal{L}_t(\cdot)$ satisfying the sparse smoothness assumption. This generalizes the descent property of projected gradient descent with convex constraints (see (Lan, 2020, Lemma 3.3) or (14)) to the nonconvex cardinality-constrained setting, showing that the loss $\mathcal{L}_t(\Theta_{t,k})$ is non-increasing across consecutive hard thresholding steps. Consequently, the last iterate $\Theta_{t,K}$ achieves the minimum loss among $\{\Theta_{t,k} : 0 \leq k \leq K\}$.

Leveraging this descent property, we can further quantify the loss incurred by the last-iterate solution $\Theta_{t,K}$:

$$\mathcal{L}_t(\Theta_{t,K}) - \mathcal{L}_t(\Theta^*) \leq \frac{1 - 2\gamma_{s,s^*}}{2\eta} \left(\frac{1 - \eta\alpha}{1 - 2\gamma_{s,s^*}} \right)^K \|\Theta_{t,0} - \Theta^*\|_2^2. \tag{17}$$

Here, the loss $\mathcal{L}_t$ can be directly related to the MSE $\|\Theta_{t,K} - \Theta^*\|_2^2$ at the beginning of round t. However, as highlightened in Challenges (ii) and (iii), issues remain: (1) the loss function

$\mathcal{L}_t$ changes dynamically as new observations arrive, yielding a moving optimization target, and (ii) the metrics on both sides of the inequality above are not naturally recursive, which prevents straightforward iteration across multiple online rounds.

To overcome these issues, we employ the (s, s)-SSC property to derive a lower bound for the loss difference $\mathcal{L}_t(\Theta_{t,K}) - \mathcal{L}_t(\Theta^*)$:

$$\begin{aligned}\mathcal{L}_t(\Theta_{t,K}) - \mathcal{L}_t(\Theta^*) &\geq \frac{\alpha}{2}\|\Theta_{t,K} - \Theta^*\|_2^2 + \langle \nabla \mathcal{L}_t(\Theta^*), \Theta_{t,K} - \Theta^* \rangle \\ &= \frac{\alpha}{2}\|\Theta_{t,K} - \Theta^*\|_2^2 - \frac{1}{t_0 + t}\sum_{i \in \mathcal{I}_t} \langle X_i \epsilon_i, \Theta_{t,K} - \Theta^* \rangle .\end{aligned}$$

Using the Cauchy-Schwarz inequality and the fact that both $\Theta_{t,K}$ and Θ^* are s-sparse, we have

$$\frac{1}{t_0 + t}\sum_{i \in \mathcal{I}_t} \langle X_i \epsilon_i, \Theta_{t,K} - \Theta^* \rangle \leq \frac{2s}{\alpha}\left\| \frac{\sum_{i \in \mathcal{I}_t} X_i \epsilon_i}{t_0 + t} \right\|_\infty^2 + \frac{\alpha}{4}\|\Theta_{t,K} - \Theta^*\|_2^2.$$

Consequently, the loss difference can be bounded from below in terms of the mean squared error and a statistical error term:

$$\mathcal{L}_t(\Theta_{t,K}) - \mathcal{L}_t(\Theta^*) \geq \frac{\alpha}{4}\|\Theta_{t,K} - \Theta^*\|_2^2 - \frac{2s}{\alpha}\left\| \frac{\sum_{i \in \mathcal{I}_t} X_i \epsilon_i}{t_0 + t} \right\|_\infty^2 . \tag{18}$$

Using the above results, we are now ready to establish a recursive relationship between the solutions obtained in successive online rounds. We summarize our result in the following lemma. We first formally define

$$\gamma_{s,s^*} := \sup\left\{ \frac{\langle Y - \Pi_{\mathcal{C}_s}(Z), Z - \Pi_{\mathcal{C}_s}(Z) \rangle}{\|Y - \Pi_{\mathcal{C}_s}(Z)\|_2^2} : Y, Z \in \mathbb{R}^d, \|Y\|_0 \leq s^*, Y \neq \Pi_{\mathcal{C}_s}(Z) \right\}$$

as the relative concavity of s-sparse hard-thresholding operator $\Pi_{\mathcal{C}_s}$ with respect to the ground truth sparsity level s^*.

Lemma 4. *Suppose Assumption 1 holds. We have $\gamma_{s,s^*} = \sqrt{s^*/s}/2$. Let $\eta_{t,k} = \eta \leq 1/L$ and choose K large enough such that*

$$K \geq K_0 := \frac{\log\{2(1 - 2\gamma_{s,s^*})/(\eta\alpha)\}}{\log\{(1 - 2\gamma_{s,s^*})/(1 - \eta\alpha)\}},$$

then

$$\delta := \frac{2(1 - 2\gamma_{s,s^*})}{\eta\alpha}\left(\frac{1 - \eta\alpha}{1 - 2\gamma_{s,s^*}}\right)^K < 1. \tag{19}$$

Consequently,

$$\|\Theta_{t+1,0} - \Theta^*\|_2^2 = \|\Theta_{t,K} - \Theta^*\|_2^2 \leq \underbrace{\delta \|\Theta_{t,0} - \Theta^*\|_2^2}_{I} + \underbrace{\frac{8s}{\alpha^2}\left\| \frac{\sum_{i \in \mathcal{I}_t} \epsilon_i X_i}{t + t_0} \right\|_\infty^2}_{II}. \tag{20}$$

The result above bounds the mean squared error $\|\Theta_{t+1,0} - \Theta^*\|_2^2$ by two terms. The first term (I) contracts the error from the previous round $\|\Theta_{t,0} - \Theta^*\|_2^2$ by a factor of δ through K iterations of hard-thresholded gradient descent. The second term (II) represents the optimal statistical error with $t + t_0$ samples in round t. Therefore, when η, K, and s are chosen appropriately, the successive solutions form a contraction, ensuring that the sequence of iterates $\Theta_{t,K} : t \geq 0$ converges to the true sparse coefficient Θ^*. By applying this result recursively, we establish the convergence rate of $\Theta_{t,K}$ for any round $t \geq 1$.

Theorem 5. *Suppose Assumptions 1 and 2 hold. Let $\{\Theta_{t,k} : t \geq 1, 1 \leq k \leq K\}$ denote the solution sequence generated by Algorithm 1 with step size $\eta_{t,k} = \eta \leq 1/L$, $K \geq K_0$, and projection sparsity $s > s^*/(\eta^2\alpha^2)$. Then*

$$\mathbb{E}[\|\Theta_{t+1,0} - \Theta^*\|_2^2] \leq \delta^{t+1}\mathbb{E}\left[\|\Theta_0 - \Theta^*\|_2^2\right] + \frac{8s\sigma_x^2}{\alpha^2\log(1/\delta)}\left(\frac{\delta^{t-1}}{t_0+1} + \frac{3}{\delta(t+1+t_0)} + \frac{\delta^{(t+1)/2-1}}{t_0+1}\right)$$

where $\delta < 1$ is defined in (19)*. Consequently, when*

$$t \gtrsim \frac{\log\left\{\mathbb{E}\|\Theta_0 - \Theta^*\|_2^2/(s^*\sigma_x^2)\right\}}{\log(1/\delta)} + \frac{\log(t+t_0)}{\log(1/\delta)},$$

we have $\mathbb{E}[\|\Theta_{t+1,0} - \Theta^*\|_2^2] \lesssim s\sigma_x^2/(t+t_0)$.

The above theorem indicates that, for projection sparsity satisfying $s^*/(\eta^2\alpha^2) < s \lesssim s^*/(\eta^2\alpha^2)$, Algorithm 1 achieves a convergence rate of $\mathcal{O}(s^*\sigma_x^2/(t+t_0))$, matching the optimal offline statistical error given in Lemma 2. Furthermore, if we select the step size $\eta_{t,k} = \eta = 1/L$, the projection sparsity requirement reduces to $s > \kappa^2 s^*$. When the sparse condition number κ is small, this requirement is minimal. For example, in isotropic designs where $\kappa = 1$, the OHT algorithm converges with $s > s^*$, whereas under-projection ($s < s^*$) will miss some of the true signals. We note, however, that the dependency of s on κ may be sub-optimal and could potentially be improved.

3.2 Online Low-Rank Matrix Sensing

In this section, we analyze the OHT algorithm for the online low-rank matrix sensing problem. Here, the rank function serves as the sparsity measure, i.e., $\Upsilon(\Theta) = \mathrm{rank}(\Theta)$, and the corresponding projection operator is

$$\Pi_{\mathcal{C}_s}(\Theta) = \underset{X \in \mathbb{R}^{d_1 \times d_2}}{\mathrm{argmin}} \Big\{\|X - \Theta\|_{\mathrm{F}}^2 \mid \mathrm{rank}(X) \leq s\Big\}. \tag{21}$$

Let $\Theta = U\Sigma V^\top$ denote the singular value decomposition (SVD) of $\Theta \in \mathbb{R}^{d_1 \times d_2}$, where the singular values in Σ are arranged in decreasing order, and let U_s be the first s columns of U. Then, the projection operator can be equivalently written as $\Pi_{\mathcal{C}_s}(\Theta) = U_s U_s^\top \Theta$.

In this context, we refer to the generalized sparse strong smoothness and generalized sparse strong convexity assumptions as low-rank strong smoothness (LowRankSS) and low-rank strong convexity (LowRankSC) assumptions, respectively. As in the online sparse linear regression setting, the following assumptions are required.

Assumption 3. *For all $t \geq 0$, the loss function $\mathcal{L}_t$ is (s,s)-LowRankSC and (s,s)-LowRankSS assumptions with parameter α and L, respectively. The corresponding (s,s)-low-rank condition number is defined as $\kappa = L/\alpha$.*

A sufficient condition for LowRankSC and LowRankSS is the RIP condition (Maros and Scutari, 2023). In particular, Theorem 2.3 of Candès and Plan (2011) shows that RIP holds with high probability for random measurement ensembles satisfying an appropriate concentration inequality. Properly normalized i.i.d. Gaussian, Bernoulli, and sub-Gaussian measurement ensembles are examples (Vershynin, 2000).

Assumption 4. *The noise terms ϵ_j are mean-zero with bounded second moments, i.e., $\mathbb{E}[\epsilon_j] = 0$ and $\mathbb{E}[\epsilon_j^2] \leq \sigma^2$. Moreover, there exists a constant $\sigma_X > 0$ such that*

$$\mathbb{E}\left[\Big\| \sum_{i \in \mathcal{I}_t} \frac{\epsilon_i X_i}{\sqrt{t_0 + t}} \Big\|_2^2\right] \leq \sigma_X^2.$$

With these assumptions, we first characterize the performance of $\widehat{\Theta}_t$, the target estimator for the OHT algorithm in round $t \geq 0$.

Lemma 6. *Suppose Assumptions 3 and 4 hold. For all $t \geq 1$,*

$$\mathbb{E}[\|\widehat{\Theta}_t - \Theta^*\|_{\mathrm{F}}^2] \leq \frac{8 s^* \sigma_X^2}{\alpha^2 (t_0 + t)}.$$

We point out that (Negahban and Wainwright, 2011, Corollary 1) established a similar high-probability bound. Our result here is in expectation, which is needed for our subsequent analysis.

Lemma 7. *Suppose $\mathcal{L}_t$ is (s,s)-LowRankSS with smoothness parameter L, and let $\eta_{t,k} \leq 1/L$. Let $S_{t,k}$ and $S_{t,k+1}$ denote the column spaces of $\Theta_{t,k}$ and $\Theta_{t,k+1}$, respectively. Then*

$$\mathcal{L}_t(\Theta_{t,k+1}) - \mathcal{L}_t(\Theta_{t,k}) \leq -\frac{\eta_{t,k}(1 - L\eta_{t,k})}{2} \|P_{S_{t,k} + S_{t,k+1}} \nabla \mathcal{L}_t(\Theta_{t,k}))\|_{\mathrm{F}}^2.$$

This lemma implies that the objective value is non-increasing between consecutive hard-thresholding gradient steps within each round, for any projection rank $s \geq 1$ and any low-rank strong smooth loss function. It generalizes the descent property of projected gradient descent with a convex constraint (Lan, 2020, Lemma 3.3) to accommodate rank-cardinality constraints.

The next two results establish the recursive relation between successive rounds and the overall convergence guarantee. Recall the definitions of K_0 and δ in Lemma 4, where α denotes the low-rank strong convexity parameter and γ_{s,s^*} is the relative concavity parameter

$$\gamma_{s,s^*} = \sup\left\{ \frac{\langle Y - \Pi_{\mathcal{C}_s}(Z), Z - \Pi_{\mathcal{C}_s}(Z)\rangle}{\|Y - \Pi_{\mathcal{C}_s}(Z)\|^2} : Y, Z \in \mathbb{R}^{d_1 \times d_2}, \mathrm{rank}(Y) \leq s^*, Y \neq \Pi_{\mathcal{C}_s}(Z) \right\}.$$

It can be shown that, for the low-rank hard-thresholding operator, $\gamma_{s,s^*} = \sqrt{s^*/s}/2$; see Lemma 18 in the appendix.

Lemma 8. *Suppose Assumption 3 holds. Take $\eta_{t,k} = \eta \leq 1/L$. If $s > s^*/(\eta^2\alpha^2)$ and $K \geq K_0 = \log\{2(1-2\gamma_{s,s^*})/(\eta\alpha)\}/\log\{(1-2\gamma_{s,s^*})/(1-\eta\alpha)\}$, then*

$$\|\Theta_{t+1,0} - \Theta^*\|_{\mathrm{F}}^2 \leq \delta \|\Theta_{t,0} - \Theta^*\|_{\mathrm{F}}^2 + \frac{8s}{\alpha^2}\left\|\frac{\sum_{i\in\mathcal{I}_t}\epsilon_i X_i}{t_0+t}\right\|_2^2,$$

where $\delta < 1$ is as in Lemma 4.

Theorem 9. *Suppose Assumptions 3 and 4 hold. Let $\{\Theta_{t,k}\}$ be generated by Algorithm 1 with $\eta_{t,k} = \eta \leq 1/L$, $K \geq K_0$, and projection rank $s > s^*/(\eta^2\alpha^2)$. Then*

$$\mathbb{E}\left[\|\Theta_{t+1,0} - \Theta^*\|_{\mathrm{F}}^2\right] \leq \delta^{t+1}\|\Theta_0 - \Theta^*\|_{\mathrm{F}}^2 + \frac{8s\sigma_X^2}{\alpha^2\log(1/\delta)}\left(\frac{\delta^t}{t_0+1} + \frac{3}{\delta(t+2+t_0)} + \frac{\delta^{t/2}}{t_0+1}\right),$$

for some $\delta < 1$. Consequently, when

$$t \gtrsim \frac{\log\left\{\mathbb{E}\|\Theta_0 - \Theta^*\|_{\mathrm{F}}^2/(s^*\sigma_x^2)\right\}}{\log(1/\delta)} + \frac{\log(t+t_0)}{\log(1/\delta)},$$

we have

$$\mathbb{E}[\|\Theta_{t+1,0} - \Theta^*\|_{\mathrm{F}}^2] \lesssim \frac{s\sigma_X^2}{\alpha^2(t+t_0)}.$$

This theorem indicates that Algorithm 1 produces a sequence that converges to the ground truth coefficient matrix at a rate of $\mathcal{O}(s^*\sigma_X^2/t)$ for online rank-constrained matrix sensing, under projection rank levels $s^*/(\eta^2\alpha^2) < s \lesssim s^*$. These results parallel those obtained for online sparse linear regression, with σ_x^2 replaced by σ_X^2.

4. No Batch Initialization

In previous sections, we introduced our online hard thresholding algorithm, Algorithm 1, for efficient generalized sparse online regression. The algorithm relies on the key assumption that the generalized sparse strong convexity and smoothness assumptions hold, which is only possible by starting with an initial batch of observations. This raises a natural question:

What if such an initialization batch is not available?

This section addresses this question by proposing a *single-phase* algorithm that computes a solution upon receiving each observation, without a batch initialization. Specifically, this corresponds to $t_0 = 0$ in the previous section, so that

$$\mathcal{L}_t(\Theta) = \frac{1}{t}\sum_{j=1}^{t}\ell_j(\Theta).$$

We assume that the GSSS and GSSC assumptions hold after receiving t_w observations, although the exact value of t_w may not be known in practice. For simplicity, we shall take $t_w = t_0$. With a slight abuse of notation, we denote the data pairs received up to round t as $(X_1, Y_1), \cdots, (X_t, Y_t)$ and the solution computed in round t as Θ_t. The proposed algorithm proceeds exactly as Algorithm 1, but without an initial batch.

4.1 Online Sparse Linear Regression

When analyzing the performance of Algorithm 1 without batch initialization, we assume that the $(2s, s)$-SSS assumption[4] with parameter L holds for all $t \geq 0$. A key challenge in this setting is that the SSC assumption may not hold during early rounds, potentially leading to poor early solutions. Specifically, without the SSC assumption, errors from the hard-thresholding steps for $t \leq t_0$ may propagate and amplify across rounds.

Proposition 10. *Suppose Assumption 2 holds, and $\mathcal{L}_t(\cdot)$ is $(2s, s)$-SSS with parameter L for $t \geq 1$. Considering running Algorithm 1 for t_0 rounds without an initial batch, with step size $\eta_t = \eta \leq \frac{1}{L}$ and arbitrary $s \geq 1, K \geq 1$. Then, there exists a constant C_1 independent of t such that, for all $1 \leq t \leq t_0$,*

$$\mathbb{E}[\mathcal{L}_t(\Theta_{t,K})] \leq C_1 4^{tK} + \sigma^2 \log(t+1),$$

where $C_1 = C_0(\|\Theta_{0,0} - \Theta^*\|_2 + (4s\eta^2\sigma_x^2 + \|\Theta^*\|_2^2)/3)$ and $C_0 = (s+s^*)\mathbb{E}\big[\|X_0\|_\infty^2\big]$.

Once t_0 observations are collected, the SSC assumption begins to hold. Then, $\|\Theta_{t_0,K} - \Theta^*\|^2$ can be treated as the initial gap. By applying Eq. (18), we see that the expected initial gap is bounded:

$$\begin{aligned}\frac{\alpha}{4}\mathbb{E}[\|\Theta_{t_0,K} - \Theta^*\|_2^2] &\leq \mathbb{E}[\mathcal{L}_{t_0}(\Theta_{t_0,K}) - \mathcal{L}_{t_0}(\Theta^*)] + \frac{2s}{\alpha}\mathbb{E}\left[\Big\|\frac{\sum_{i\in\mathcal{I}_{t_0}} X_i\epsilon_i}{t_0}\Big\|_\infty^2\right] \\ &\leq C_1 4^{t_0 K} + \sigma^2\log(t_0+1) + \mathcal{O}\Big(\frac{s\sigma_x^2}{t_0}\Big),\end{aligned}$$

and this gap is geometrically discounted in each subsequent round $t > t_0$ by Theorem 5.

Corollary 11. *Suppose $\mathcal{L}_t$ is (s, s)-SSC with parameter α for all $t > t_0$, $(2s, s)$-SSS with parameter L for all $t \geq 1$, and Assumption 2 holds. Let $\{\Theta_{t,k}, t \geq 1, 1 \leq k \leq K, \}$ be the sequence generated by Algorithm 1 with $\eta_{t,k} = \eta \leq 1/L$, $s > s^*/(\eta^2\alpha^2)$, and*

$$K \geq K_0 := \frac{\log\{2(1-2\gamma_{s,s^*})/(\eta\alpha)\}}{\log\{(1-2\gamma_{s,s^*})/(1-\eta\alpha)\}}, \quad \gamma_{s,s^*} = \sqrt{s^*/s}/2,$$

but without an initial batch. Then for all $t \geq t_0$,

$$\mathbb{E}[\|\Theta_{t,K} - \Theta^*\|_2^2] \leq \delta^{t-t_0}\mathbb{E}[\|\Theta_{t_0,K} - \Theta^*\|_2^2] + \mathcal{O}\Big(\frac{s\sigma_x^2}{t}\Big) \leq \frac{4C_1\delta^{t-t_0}4^{t_0K}}{\alpha} + \mathcal{O}\Big(\frac{s\sigma_x^2}{t}\Big),$$

where $\delta < 1$ is defined in (19).

The above result indicates that Algorithm 1 without batch initialization still achieves the $\mathcal{O}(s^*\sigma_x^2/t)$ convergence rate once t is only logarithmically large.

4. This is slightly stronger than the (s, s)-SSS assumption in the setting with an initial batch; it can be relaxed to (s, s)-SSS assumption with appropriately defined higher-moment bounds on the covariates X_i.

4.2 Online Low-Rank Matrix Sensing

We further study the performance of Algorithm 1 for rank-constrained matrix sensing without batch initialization.

Proposition 12. *Suppose Assumption 4 holds and $\mathcal{L}_t(\cdot)$ is $(2s, s)$-LowRankSS with parameter L for all $t \geq 1$. Suppose we run Algorithm 1 for t_0 rounds without an initial batch, using step size $\eta_t = \eta \in [0, 1/L]$ and arbitrary $s \geq 1, K \geq 1$. Then, there exists a constant*

$$C_1 = 3C_0(\|\Theta_0 - \Theta^*\|_{\mathrm{F}}^2 + (4s\eta^2\sigma_X^2 + \|\Theta^*\|_{\mathrm{F}}^2)/3)$$

independent of t such that for all $1 \leq t \leq t_0$

$$\mathbb{E}[\mathcal{L}_t(\Theta_{t,K})] \leq C_1 4^{tK} + \sigma^2 \log(t+1),$$

where $C_0 = (s + s^*)\mathbb{E}\big[\|X_0\|_2^2\big]$.

Corollary 13. *Suppose Assumption 4 holds, $\mathcal{L}_t(\cdot)$ is $(2s, s)$-LowRankSS with parameter L for all $t \geq 1$, and $\mathcal{L}_t(\cdot)$ is (s, s)-LowRankSC with parameter α for all $t \geq t_0$. Let $\{\Theta_{t,k}, t \geq 1, 1 \leq k \leq K\}$ be the sequence generated by Algorithm 1 with $s > s^*/(\eta^2\alpha^2)$, $\eta_t = \eta \leq 1/L$, and*

$$K \geq K_0 := \frac{\log\{2(1 - 2\gamma_{s,s^*})/(\eta\alpha)\}}{\log\{(1 - 2\gamma_{s,s^*})/(1 - \eta\alpha)\}}, \quad \gamma_{s,s^*} = \sqrt{s^*/s}/2,$$

but without an initial batch. Then, for all $t \geq t_0$,

$$\mathbb{E}[\|\Theta_{t,K} - \Theta^*\|_{\mathrm{F}}^2] \leq \delta^{t-t_0}\mathbb{E}[\|\Theta_{t_0,K} - \Theta^*\|_{\mathrm{F}}^2] + \mathcal{O}\Big(\frac{s\sigma_X^2}{t}\Big) \leq \frac{4C_1\delta^{t-t_0}4^{t_0K}}{\alpha} + \mathcal{O}\Big(\frac{s\sigma_X^2}{t}\Big),$$

where $\delta < 1$ is defined in (19).

We can also see that Algorithm 1 exhibits the $\mathcal{O}(s^*\sigma_X^2/t)$ rate of convergence for rank-constrained matrix sensing without an initial batch.

5. Numerical Experiments

This section presents numerical experiments to validate the practical performance of our proposed algorithms.

5.1 Online Sparse Linear Regression

For the linear regression problem, let $d = d_1 = 1000$ and $d_2 = 1$. We consider a *weak* signal setting, where the sparse coefficient vector $\Theta^* \in \mathbb{R}^d$ has its first 15 entries drawn from a normal distribution $\Theta_j^* \sim \mathcal{N}(0, 0.2)$ for $j = 1, \cdots, 15$, and rest entries set to zero. Hence, the true sparsity is $s^* = \|\Theta^*\|_0 = 15$.

We evaluate the performance of our `OHT` algorithm, Algorithm 1, using different numbers of hard thresholding steps $K = 1, 5, 10, 20$ per round. As baselines, we implement `OS_LASSO_K=1` and `OS_LASSO_K=20` (Fan et al., 2018a), which first determine the support via LASSO on an initial batch of size t_0, and then perform K gradient descent steps over the pre-determined support. All algorithms share the same initial batch and online data.

We consider two covariate correlation structures.

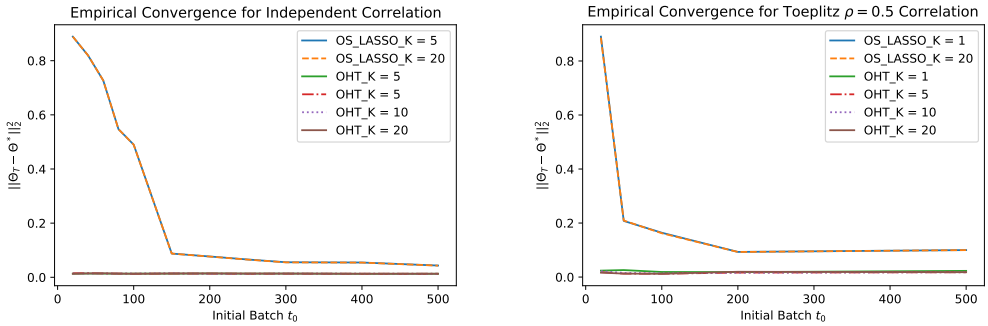

Figure 2: Online sparse linear regression under weak signal setup for $t_0 \in [20, 500]$ and $t = 10000$ online learning rounds for independent and Toeplitz $\rho = 0.5$ covariate designs.

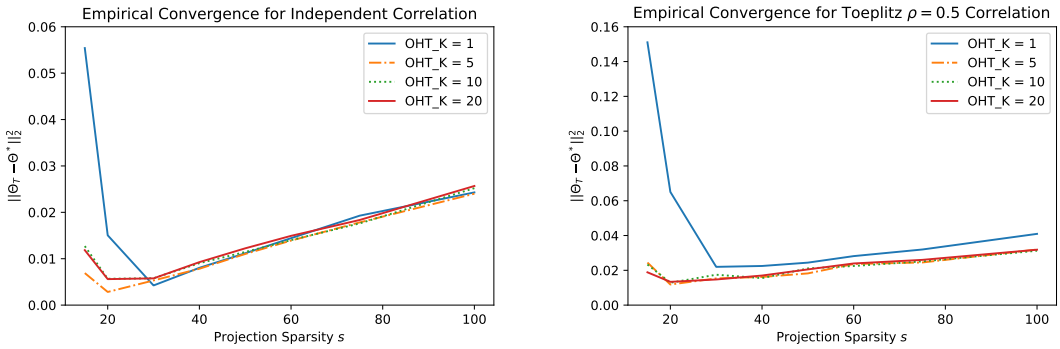

Figure 3: Online sparse linear regression under weak signal setup for projection sparsity level $s \in [15, 100]$, $t_0 = 100$, and $T = 10^4$ online rounds, under independent and Toeplitz $\rho = 0.5$ covariate correlation designs.

(i) Independent: each entry $[X_i]_j \sim \mathcal{N}(0, 1)$ independently for $1 \le j \le d$;

(ii) Toeplitz with $\rho = 0.5$: $X_i \sim \mathcal{N}(0, \Sigma)$, where $\Sigma_{jk} = \rho^{|j-k|}$ for $1 \le j, k \le d$.

For each covariate design, the response is generated as $Y_i = X_i^\top \Theta^* + \epsilon_i$ with independent noise $\epsilon_i \sim \mathcal{N}(0, 1)$. Each experiment first generates an initial batch of size t_0, followed by 10^4 online rounds, where each round observes a single data point (X_t, Y_t). For `OHT_K=1`, `OHT_K=5`, `OHT_K=10`, and `OHT_K=20`, we update Θ_t via K hard-thresholding gradient descent steps with projection sparsity s.

For the baseline algorithms `OS_LASSO_K=1` and `OS_LASSO_K=20`, we first estimate the support by running LASSO over the initial batch, where the regularization parameter is selected using cross-validation. Subsequently, in each online round t, we update Θ_t by conducting K truncated gradient descent steps that keep only the components in the estimated

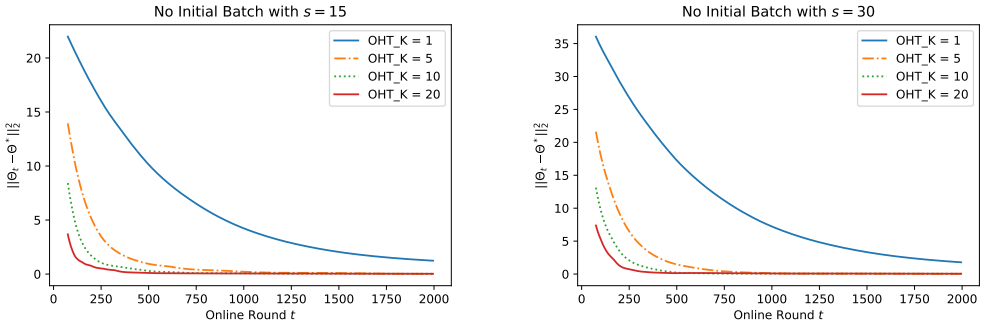

Figure 4: Online sparse linear regression under weak signal setup without the initial batch under projection sparsity level $s \in \{15, 30\}$.

support and truncate the rest to zero. From now on, we shall write Θ_t as the last iterate solution $\Theta_{t,K}$ for simplicity. We consider the following three sets of experiments.

1. We test the algorithms over different initial batch sizes $t_0 = 20, 40, \cdots, 500$ and $T = 10^4$ online rounds under both covariate designs. We report the MSE of the output solution, $\|\Theta_t - \Theta^*\|_2^2$, against t_0 in Figure 2. The projection sparsity is set to $s = 50$ and the step size to $\eta = 0.001$.

2. We test our algorithms for $s = 15, 20, 30, \cdots, 100$ under both covariate correlation designs, with $\eta = 0.001$, $t_0 = 100$, and $T = 10^4$ online rounds. Figure 3 summarizes the results.

3. We evaluate the algorithm for $s \in \{15, 30\}$ under the independent covariate design, with $\eta = 0.001$, $t_0 = 100$, and $T = 10^4$ online learning rounds. The MSE $\|\Theta_t - \Theta^*\|_2^2$ is plotted against t in Figure 4.

4. Finally, we test our OHT algorithms on highly correlated covariates with a large sparse condition number κ. We generate an orthonormal basis $U \in \mathbb{R}^{p \times p}$ and a diagonal positive-definite matrix Λ, then generate $X_i \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = U \Lambda U^\top$. Specifically, the first 12 diagonal entries of Λ are $500, 100, 10, 9, 8, \ldots, 1$ and the remaining entries are 1, giving $\kappa = 500$. The signal Θ^* is generated as before, with an initial batch size $t_0 = 100$ and $T = 10^4$ online rounds. We test sparsity levels $s = 14, 15, 20, 25, 30, 40$, and results are reported in Table 1.

From Figure 2, we observe that the performance of `OS_LASSO_K=1` and `OS_LASSO_K=20` is sensitive to the initial batch size t_0 under both covariate designs. In particular, both algorithms converge to solutions far away from the true coefficients Θ^* when the initial batch size is small ($t_0 \leq 150$) and their convergence performance improves with larger initial batches. By contrast, our OHT algorithms are robust to the initial batch size, achieving comparable performance for all $t_0 \in [20, 500]$.

From Figure 3, when $s \leq 30$, `OHT_K=5`, `OHT_K=10`, and `OHT_K=20` outperform `OHT_K=1`. When $s > 30$, all OHT variants perform comparably under the independent design, while

MSE	$s = 14$	$s = 15$	$s = 20$	$s = 25$	$s = 30$	$s = 40$
OHT_K=1	0.519131	0.529622	0.376415	0.239696	0.138760	0.126439
OHT_K=5	0.115403	0.070367	0.011384	0.007039	0.010385	0.012848
OHT_K=10	0.070169	0.046840	0.007254	0.009965	0.011081	0.013396
OHT_K=20	0.046217	0.009727	0.007280	0.008915	0.011370	0.014919

Table 1: Empirical last-iterate MSE $\|\Theta_T - \Theta^*\|_2^2$ for $t_0 = 100$ and $T = 10^4$ under projection sparsity levels $s = 14, 15, 20, 25, 30, 40$ and $[\Lambda_{i,i}, 1 \le i \le p] = [500, 100, 10, 8, \cdots, 1, 1, \cdots, 1]$.

OHT_K=5,10,20 slightly outperform OHT_K=1 under the Toeplitz design. These observations are consistent with Theorem 5, which requires sufficiently large s and K to ensure global convergence. Furthermore, the performance of OHT_K=5,10,20 is comparable, suggesting that $K = 5$ already exceeds the intrinsic K_0, and increasing K further does not improve accuracy further.

From Figure 4, under both sparsity levels $s = 15, 30$, our OHT algorithms with $K = 5, 10, 20$ generate sequences Θ_t that converge to the true sparse coefficient Θ^* even without an initial batch, while OHT_=1 converges much more slowly. At $t = 10^4$, a performance gap between OHT_=1 and OHT_=5,10,20 remains evident.

Based on Table 1, we observe that Algorithm 1 performs well and only requires a slightly over-selected projection sparsity level, $s = 20$, even when the sparse condition number $\kappa = 500$ is large. This suggests that the dependence of the projection sparsity on κ in Theorem 5 may not be tight.

These numerical experiments indicate the superior performance of our algorithm compared to OS_LASSO, regardless of the covariate design, particularly in scenarios with weak signals or small initial batches. Additional experiments in Appendix A further demonstrate empirical convergence across online rounds and under various design and signal strength settings, highlighting the efficiency and robustness of our approach.

5.2 Online Low-Rank Matrix Sensing

We evaluate the performance of our algorithm for the online low-rank matrix sensing problem. We consider $d_1 = d_2 = 50$ and an underlying coefficient matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ of rank $s^* = 5$. In our experiments, Θ^* is obtained as the rank-5 approximation of a matrix $\widetilde{\Theta} \in \mathbb{R}^{d_1 \times d_2}$ with entries $\widetilde{\Theta}_{i,j} \sim \mathcal{N}(0, 1)$ independently. Each feature-label pair (X_t, Y_t) is generated by independently sampling $X_t \in \mathbb{R}^{d_1 \times d_2}$ with entries $[X_t]_{i,j} \sim \mathcal{N}(0, 1)$, followed by $Y_t = \langle X_t, \Theta^* \rangle + \epsilon_t$ with $\epsilon_t \sim \mathcal{N}(0, 1)$. We conduct the following experiments.

1. With an initial batch of size $t_0 = 100$, we run $T = 10^4$ online rounds for various projection ranks $s \in \{3, 4, 5, 10, 15, 20\}$. For $s = 5$, we plot the MSE $\|\Theta_t - \Theta^*\|_{\mathrm{F}}^2$ versus the round t in Figure 5, and the log-MSE $\log(\|\Theta_t - \Theta^*\|_{\mathrm{F}}^2)$ versus $\log t$ to assess empirical convergence rates, including a reference line of slope -1. Similar results for $s = 15$ are presented in Figure 6. The MSE of the last-iterate solution Θ_T at $T = 10^4$ is reported for all algorithms in Table 2.

MSE	$s = 3$	$s = 4$	$s = 5$	$s = 10$	$s = 15$	$s = 20$
`OHT_K=1`	10.8453	5.6856	0.0546	0.1265	0.1848	0.2395
`OHT_K=5`	10.8447	5.6460	0.0508	0.1357	0.1968	0.2499
`OHT_K=10`	10.8367	5.6276	0.0503	0.1367	0.1958	0.2451
`OHT_K=20`	10.8299	5.2613	0.0499	0.1350	0.1990	0.2516

Table 2: Empirical last-iterate MSE $\|\Theta_T - \Theta^*\|_{\mathrm{F}}^2$ for $t_0 = 100$ and $T = 10^4$ under projection rank levels $s = 3, 4, 5, 10, 15, 20$.

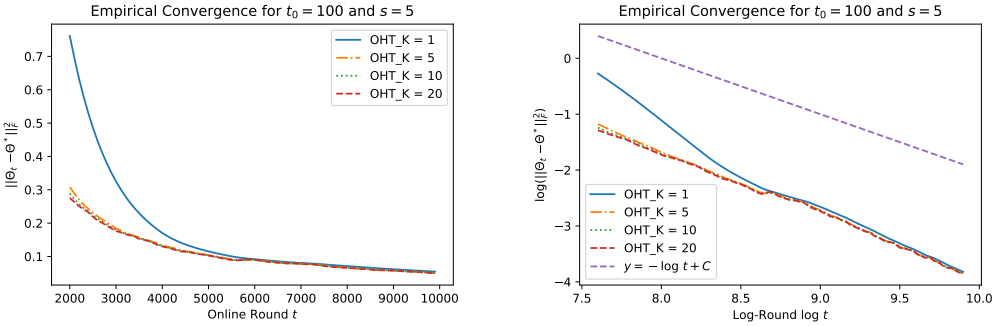

Figure 5: Online low-rank matrix sensing for $t_0 = 100$ and $T = 10000$ online learning rounds with projection sparsity level $s = 5$.

2. Without an initial batch ($t_0 = 0$), we run $T = 2000$ online rounds for projection ranks $s \in \{5, 15\}$ and hard-thresholding gradient descent steps $K \in \{1, 5, 10, 20\}$, reporting $\|\Theta_t - \Theta^*\|_{\mathrm{F}}^2$ versus t and $\log t$ in Figure 7.

From Figures 5 and 6, Algorithm 1 generates solution sequences converging to Θ^* for $s = 5, 15$ and $K = 1, 5, 10, 20$, with an initial batch of size $t_0 = 100$. Since the covariance matrix is isotropic and $s^* = 5$, only a slightly larger projection rank is required to ensure convergence. The slopes of $\log(\|\Theta_t - \Theta^*\|_{\mathrm{F}}^2)$ versus $\log t$ are approximately -1 for both $s = 5, 15$ when $K \geq 5$, consistent with Theorem 9, which predicts an $\mathcal{O}(s/(t + t_0))$ convergence rate.

Table 2 shows that under-picking the projection level (e.g., $s = 3, 4$) prevents convergence. Additionally, the MSE for $s = 20$ is roughly four times that for $s = 5$, confirming the linear dependence of the MSE on the projection sparsity s as indicated in Theorem 9.

Finally, Figure 7 demonstrates that our algorithms perform well even without an initial batch, with Θ_t converging to Θ^* for $s \in \{5, 15\}$. These results validate the practical efficiency and robustness of our method and corroborate the theoretical guarantees in Section 4.2.

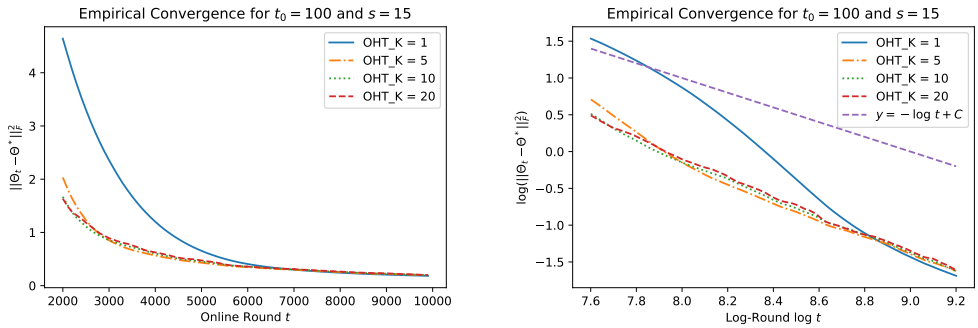

Figure 6: Online low-rank matrix sensing for $t_0 = 100$ and $T = 10000$ online learning rounds with projection sparsity level $s = 15$.

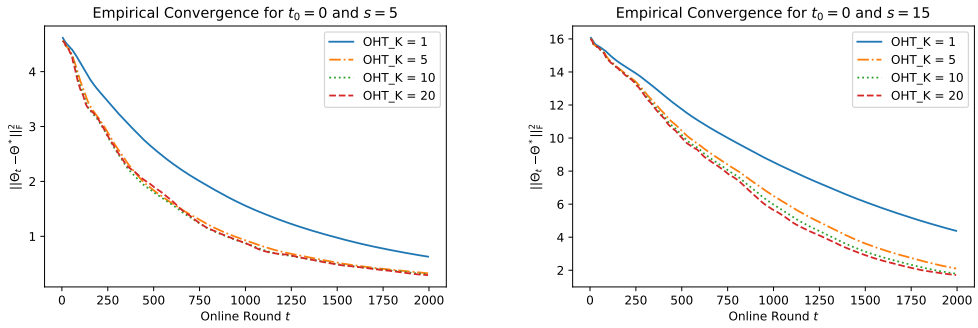

Figure 7: Online low-rank matrix sensing *without* the initial batch ($t_0 = 0$) and $T = 2000$ online learning rounds with projection rank levels $s \in \{5, 15\}$.

6. Conclusions

This paper proposes an online hard thresholding algorithm for online generalized sparse regression problems, with a focus on online sparse linear regression and low-rank matrix sensing. The algorithm does not need dynamic regularization, features closed-form updates in each online round, requires only the storage of summary statistics, making it memory and storage efficient, and converges globally to the ground-truth coefficients at the optimal statistical rate under realistic assumptions. To our knowledge, it is the first algorithm that simultaneously achieves all four of these properties. One practical limitation is that key quantities such as the per-round iteration number K and projection sparsity s are typically unavailable in real-world applications. In practice, the projection sparsity level can be chosen via cross-validation using an initial batch. If no initial batch is available, cross-validation can be performed during the early online rounds, after which the algorithm proceeds in a streaming fashion. Moreover, numerical experiments demonstrate that the algorithm is robust to the choices of s and K, provided that s is not under-picked.

Acknowledgments

Qiang Sun is supported in part by an NSERC Discovery Grant (RGPIN-2018-06484), a Data Sciences Institute Catalyst Grant, and a computing grant from Compute Canada.

Appendix

This appendix presents additional numerical experiments, and collects proofs for the main results and technical lemmas. Additional numerical experiments are presented in Appendix A. Appendix B collects proofs for online sparse linear regression in Section 3.1, and Appendix C collects the proofs for online low-rank matrix sensing in Section 3.2. Appendix D collects proofs for the no initial batch case in Section 4.

Notation. For any two spaces S_1 and S_2, with a slight abuse of notation, we denote by $S_1 \cap S_2^{\perp}$ the intersection between S_1 and the orthogonal complement of S_2.

Appendix A. Additional Numerical Results

In this section, we conduct additional numerical experiments to investigate the performance of our OHT algorithms under various signal strength and covariate correlation setups. We consider the following two experiments with sparsity level $s = 50$ and step-size $\eta = 0.001$:

1. **Weak signal under different covariate correlations:** We test the algorithms for initial batch size $t_0 = 100, 500$ and 10^4 online learning rounds, under both independent and Toeplitz $\rho = 0.3, 0.5, 0.7$ covariate correlation designs provided in Section 5. We report the MSE $\|\Theta_t - \Theta^*\|_{\mathrm{F}}^2$ against online learning rounds t in Figures 8, 9, 10, and 11.

2. **Strong signal under independent covariate correlation:** We generate a vector $\widehat{\Theta}^* \in \mathbb{R}^d$ where each entry is normally distributed such that $\widehat{\Theta}_i^* \sim \mathcal{N}(0, 0.4)$; we conduct hard thresholding truncation such that $\Theta_i^* = \widehat{\Theta}_i^*$ if $|\widehat{\Theta}_i^*| \geq 1$ and $\Theta_i^* = 0$ otherwise. The cardinality of the generating parameter Θ^* is 12. We then test our algorithms for initial batch size $t_0 = 100, 500$ and 10^4 subsequent online learning rounds, under the independent covariate correlation design. We plot the MSE $\|\Theta_t - \Theta^*\|_{\mathrm{F}}$ against online learning rounds t in Figure 12.

From Figures 8, 9, 10, and 11, we observe that our OHT algorithms outperform `OS_LASSO_K=1` and `OS_LASSO_K=20` in the weak signal setting for $t_0 = 100, 500$, regardless of the covariate correlation design. Further, we observe that `OS_LASSO_K=1` and `OS_LASSO_K=20` exhibit comparable performance to our algorithms for strong-signal instances. This is because under strong-signal settings, `OS_LASSO_K=1` and `OS_LASSO_K=20` are able to identify the correct set of nonzero entries $\mathcal{S}$ by using a small initial batch $t_0 = 100$. Consequently, the solutions are truncated to the correct set in each subsequent online learning round. These numerical observations further demonstrate the efficiency and robustness of our algorithm in the weak-signal and small-initial-batch scenarios.

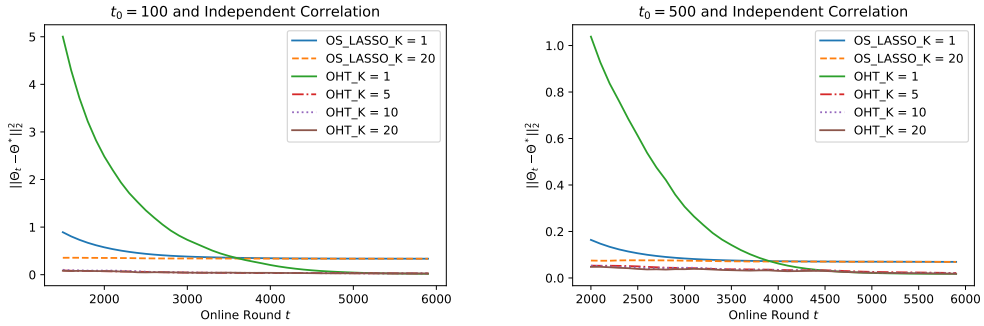

Figure 8: Empirical convergence of MSE $\|\Theta_t - \Theta^*\|_F^2$ against online learning round t under weak signal setup and independent covariate correlation

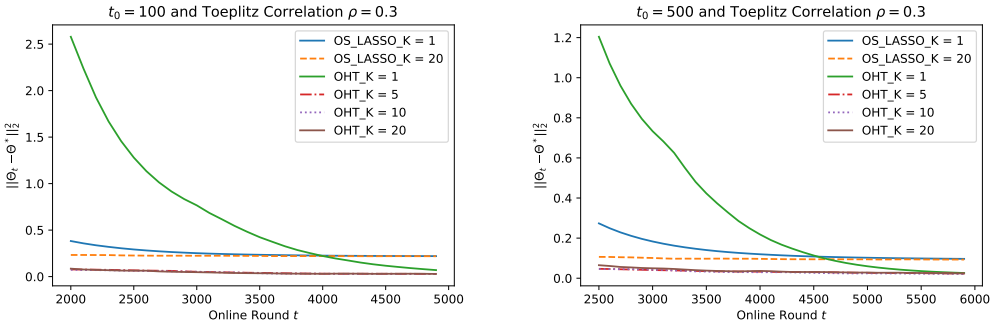

Figure 9: Empirical convergence of MSE $\|\Theta_t - \Theta^*\|_F^2$ against online learning round t and weak signal setup with Toeplitz $\rho = 0.3$ covariate correlation

Appendix B. Proofs for Section 3.1

B.1 Proof of Proposition 1

Proof Conditional on ϵ_i, we apply (Vershynin, 2018, Proposition 2.6.1) to obtain

$$\|Z_j\|_{\psi_2} = \left\| \left[\sum_{i \in \mathcal{I}_t} \frac{\epsilon_i X_i}{\sqrt{t+t_0}} \right]_j \right\|_{\psi_2} \leq \sqrt{\frac{C}{t+t_0} \sum_{i \in I_t} \epsilon_i^2 \|X_{ij}\|_{\psi_2}^2} \leq \sqrt{\frac{C\sigma_x^2}{t+t_0} \sum_{i \in I_t} \epsilon_i^2},$$

where C is some constant.

Since Z_j is sub-Gaussian with $\|Z_j\|_{\psi_2}$, conditional on ϵ_i, we have

$$\|Z_j^2\|_{\psi_1} = \|Z_j\|_{\psi_2}^2 = \frac{C\sigma_x^2}{t+t_0} \sum_{i \in I_t} \epsilon_i^2.$$

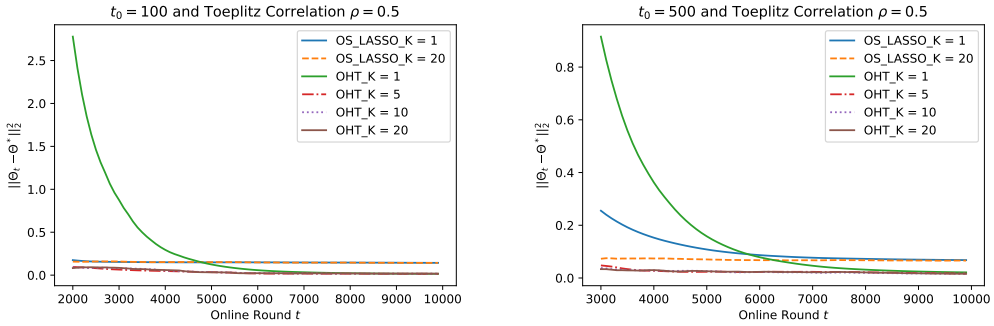

Figure 10: Empirical convergence of MSE $\|\Theta_t - \Theta^*\|_{\mathrm{F}}^2$ against online learning round t under weak signal setup and Toeplitz $\rho = 0.5$ covariate correlation

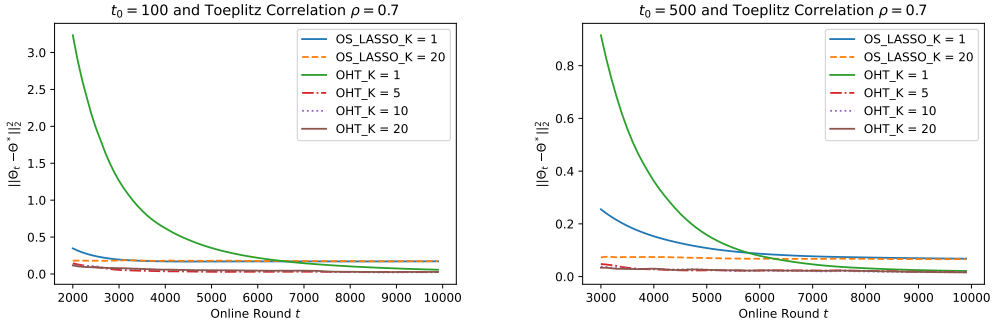

Figure 11: Empirical convergence of MSE $\|\Theta_t - \Theta^*\|_{\mathrm{F}}^2$ against online learning round t under weak signal setup and Toeplitz $\rho = 0.7$ covariate correlation

Let $V_i = \epsilon_i X_{ij}$. Then, we apply the maximum inequality (van der Vaart and Wellner, 1996, Lemma 2.2.2), and obtain

$$\mathbb{E}\left[\max_{1\le j\le d} Z_j^2 \;\middle|\; \epsilon_i, i \in I_t\right] \le \left\|\max_{1\le j\le d} Z_j^2\right\|_{\psi_1}$$
$$\le C\psi_1^{-1}(d)\left\|\frac{1}{t+t_0}\sum_{i\in\mathcal{I}_t}\epsilon_i^2 X_{ij}^2\right\|_{\psi_1} \le C\log d\cdot\frac{\sigma_x^2}{t+t_0}\sum_{i\in I_t}\epsilon_i^2,$$

and thus

$$\mathbb{E}\left[\max_{1\le j\le d} Z_j^2\right] \le C\sigma^2\sigma_x^2\log d,$$

where C is some constant. ■

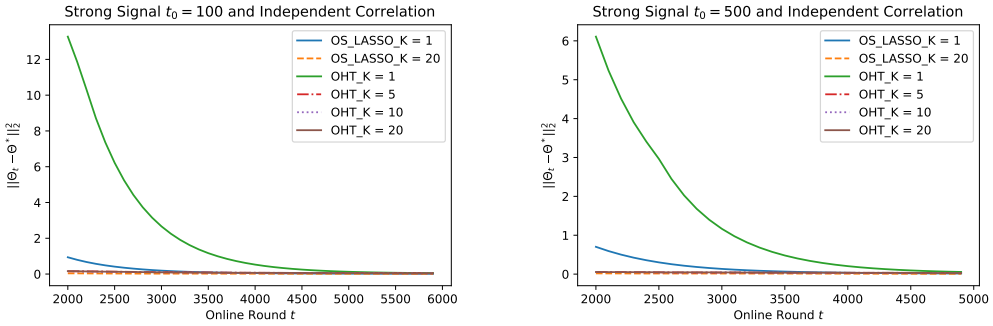

Figure 12: Empirical convergence of MSE $\|\Theta_t - \Theta^*\|_{\mathrm{F}}^2$ against online learning round t under strong signal setup and independent covariate correlation

B.2 Proof of Lemma 2

Proof Recall that $\widehat{\Theta}_t = \arg\min_{\|\Theta\|_0 \le s^*} \mathcal{L}_t(\Theta)$ for cardinality-constrained linear regression where $\Upsilon(\Theta) = \|\Theta\|_0$, we have

$$\begin{aligned}\mathcal{L}_t(\widehat{\Theta}_t) &= \frac{1}{2(t_0+t)} \sum_{i \in \mathcal{I}_t} \Big(\Big\langle X_i, \widehat{\Theta}_t \Big\rangle - Y_i \Big)^2 = \frac{1}{2(t_0+t)} \sum_{i \in \mathcal{I}_t} \Big(\Big\langle X_i, \widehat{\Theta}_t - \Theta^* \Big\rangle - \epsilon_i \Big)^2 \\ &\le \frac{1}{2(t_0+t)} \sum_{i \in \mathcal{I}_t} \Big(\langle X_i, \Theta^* \rangle - Y_i \Big)^2 = \frac{1}{2(t_0+t)} \sum_{i \in \mathcal{I}_t} \epsilon_i^2 = \mathcal{L}_t(\Theta^*),\end{aligned}$$

implying that

$$\frac{1}{2(t_0+t)} \sum_{i \in \mathcal{I}_t} \Big\langle X_i, \widehat{\Theta}_t - \Theta^* \Big\rangle^2 \le \frac{1}{t_0+t} \sum_{i \in \mathcal{I}_t} \epsilon_i \Big\langle X_i, \widehat{\Theta}_t - \Theta^* \Big\rangle .$$

Because $\mathcal{L}_t(\cdot)$ satisfies the (s,s)-SSC condition under Assumpption (1), using the fact that $\|\widehat{\Theta}_t - \Theta^*\|_0 \le 2s^*$ and applying Proposition 14 in Appendix Section B.6 arrives at

$$\begin{aligned}\frac{\alpha}{2}\|\widehat{\Theta}_t - \Theta^*\|_2^2 &\le \frac{1}{2(t_0+t)} \sum_{i \in \mathcal{I}_t} \Big\langle X_i, \widehat{\Theta}_t - \Theta^* \Big\rangle^2 \le \frac{1}{t_0+t} \sum_{i \in \mathcal{I}_t} \epsilon_i \Big\langle X_i, \widehat{\Theta}_t - \Theta^* \Big\rangle \\ &\le \sqrt{2s^*}\|\Theta' - \Theta^*\|_2 \Big\| \frac{1}{t_0+t} \sum_{i \in \mathcal{I}_t} \epsilon_i X_i \Big\|_\infty \\ &\le \frac{2s^*}{\alpha} \Big\| \frac{1}{t_0+t} \sum_{i \in \mathcal{I}_t} \epsilon_i X_i \Big\|_\infty^2 + \frac{\alpha}{4}\|\widehat{\Theta}_t - \Theta^*\|_2^2.\end{aligned} \tag{22}$$

Rearranging the terms, we obtain

$$\|\widehat{\Theta}_t - \Theta^*\|_2^2 \le \frac{8s^*}{\alpha^2} \Big\| \frac{1}{t_0+t} \sum_{i \in \mathcal{I}_t} \epsilon_i X_i \Big\|_\infty^2 .$$

By using Assumption 2 that $\mathbb{E}\Big[\Big\|\sum_{i\in\mathcal{I}_t}\epsilon_i X_i\Big\|_\infty^2\Big] \leq (t_0+t)\sigma_x^2$, and taking expectations on both sides of the above inequality, we conclude that

$$\mathbb{E}[\|\widehat{\Theta}_t - \Theta^*\|_2^2] \leq \frac{8s^*}{\alpha^2}\mathbb{E}\Big[\Big\|\frac{1}{t_0+t}\sum_{i\in\mathcal{I}_t}\epsilon_i X_i\Big\|_\infty^2\Big] \leq \frac{8s^*\sigma_x^2}{\alpha^2(t_0+t)}.$$

This completes the proof. ■

B.3 Proof of Lemma 3

Proof When $\Upsilon(\Theta) = \|\Theta\|_0$, we consider a hard thresholding step of Algorithm 1 that

$$\begin{aligned}\Theta_{t,k+1} &= \Pi_{\mathcal{C}_s}\left(\Theta_{t,k} - \eta_{t,k}\nabla\mathcal{L}_t(\Theta_{t,k})\right)\\ &= \underset{z\in\mathbb{R}^d,\|z\|_0\leq s}{\operatorname{argmin}}\left\{\|z - (\Theta_{t,k} - \eta_{t,k}\nabla\mathcal{L}_t(\Theta_{t,k}))\|_2^2\right\}.\end{aligned}$$

By Assumption 1, $\mathcal{L}_t(\cdot)$ satisfies the (s,s)-SSS condition (1) with parameter L. Let $g_k = \nabla\mathcal{L}_t(\Theta_{t,k})$. We then analyze the change of objective value from $\Theta_{t,k}$ to $\Theta_{t,k+1}$ w.r.t. $\mathcal{L}_t(\cdot)$ as

$$\begin{aligned}\mathcal{L}_t(\Theta_{t,k+1}) - \mathcal{L}_t(\Theta_{t,k}) &\leq \langle\nabla\mathcal{L}_t(\Theta_{t,k}), \Theta_{t,k+1} - \Theta_{t,k}\rangle + \frac{L}{2}\|\Theta_{t,k+1} - \Theta_{t,k}\|_2^2\\ &= \langle g_k, \Theta_{t,k+1} - \Theta_{t,k}\rangle + \frac{L}{2}\|\Theta_{t,k+1} - \Theta_{t,k}\|_2^2 = \mathrm{I} + \mathrm{II},\end{aligned} \tag{23}$$

where $\mathrm{I} := \langle g_k, \Theta_{t,k+1} - \Theta_{t,k}\rangle$ and $\mathrm{II} := \frac{L}{2}\|\Theta_{t,k+1} - \Theta_{t,k}\|_2^2$. For simplicity, we write $\eta = \eta_{t,k}$, $\Theta_k = \Theta_{t,k}$, $\Theta_{k+1} = \Theta_{t,k+1}$, $\mathcal{S}_{k-1} = \mathcal{S}_{t,k-1}$ and $\mathcal{S}_k = \mathcal{S}_{t,k}$. We proceed to bound I and II respectively. We start with I:

$$\begin{aligned}\mathrm{I} &= \langle g_k, \Theta_{k+1} - \Theta_k\rangle\\ &= -\left\langle\Theta^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}}, g^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}}\right\rangle + \left\langle\Theta^{k+1}_{\mathcal{S}_{k+1}} - \Theta^k_{\mathcal{S}_{k+1}}, g^k_{\mathcal{S}_{k+1}}\right\rangle\\ &= -\left\langle\Theta^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}}, g^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}}\right\rangle - \eta\|g^k_{\mathcal{S}_{k+1}}\|_2^2\\ &= -\left\langle\Theta^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}} - \eta g^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}}, g^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}}\right\rangle - \eta\|g^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}}\|_2^2 - \eta\|g^k_{\mathcal{S}_{k+1}}\|_2^2\\ &\leq \frac{1}{2\eta}\|\Theta^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}} - \eta g^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}}\|_2^2 - \frac{\eta}{2}\|g^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}}\|_2^2 - \eta\|g^k_{\mathcal{S}_{k+1}}\|_2^2\\ &\leq \frac{\eta}{2}\|g^k_{\mathcal{S}_{k+1}\setminus\mathcal{S}_k}\|_2^2 - \frac{\eta}{2}\|g^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}}\|_2^2 - \eta\|g^k_{\mathcal{S}_{k+1}}\|_2^2\\ &\leq -\frac{\eta}{2}\|g^k_{\mathcal{S}_k\cup\mathcal{S}_{k+1}}\|_2^2,\end{aligned}$$

where the last second inequality follows from the fact

$$\|\Theta^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}} - \eta g^k_{\mathcal{S}_k\setminus\mathcal{S}_{k+1}}\|_2^2 \leq \|\Theta^{k+1}_{\mathcal{S}_{k+1}\setminus\mathcal{S}_k}\|_2^2 = \eta^2\|g^k_{\mathcal{S}_{k+1}\setminus\mathcal{S}_k}\|_2^2.$$

For II, by using the fact that $\|a\|_2^2 = \|a+b\|_2^2 - \|b\|_2^2 - 2\langle a, b\rangle$, we have

$$\begin{aligned}
\mathrm{II} &= \frac{L}{2}\|\Theta_{k+1} - \Theta_k\|_2^2 \\
&= \frac{L}{2}\left\|\Theta_{k+1} - \Theta_k + \eta g^k_{\mathcal{S}_{k+1}\cup\mathcal{S}_t}\right\|_2^2 - \frac{L}{2}\left\|\eta g^k_{\mathcal{S}_{k+1}\cup\mathcal{S}_k}\right\|_2^2 - L\eta\cdot\left\langle g^k_{\mathcal{S}_{k+1}\cup\mathcal{S}_k}, \Theta_{k+1} - \Theta_k\right\rangle \\
&= \frac{L}{2}\inf_{z\in\mathcal{C}_s}\left\|z - \Theta_k + \eta g^k_{\mathcal{S}_{k+1}\cup\mathcal{S}_k}\right\|_2^2 - \frac{L}{2}\left\|\eta g^k_{\mathcal{S}_{k+1}\cup\mathcal{S}_k}\right\|_2^2 - L\eta\cdot\mathrm{I} \\
&\le -L\eta\cdot\mathrm{I}.
\end{aligned}$$

Plugging the above bounds for I and II into (23) acquires

$$\mathcal{L}(\Theta_{k+1}) - \mathcal{L}(\Theta_k) \le (1-L\eta)\cdot\mathrm{I} \le -\frac{\eta(1-L\eta)}{2}\|g^k_{\mathcal{S}_k\cup\mathcal{S}_{k+1}}\|_2^2,$$

as desired. ■

B.4 Proof of Lemma 4

Proof By using the (s,s)-SSC and (s,s)-SSS conditions (1) with $\Upsilon(\Theta) = \|\Theta\|_0$ and the fact that $\|\Theta_{t,k-1}\|_0, \|\Theta_{t,k}\|_0, \|\Theta^*\|_0 \le s$, we have the followings:

$$\begin{aligned}
\mathcal{L}_t(\Theta^*) &\ge \mathcal{L}_t(\Theta_{t,k-1}) + \langle\nabla\mathcal{L}_t(\Theta_{t,k-1}), \Theta^* - \Theta_{t,k-1}\rangle + \frac{\alpha}{2}\|\Theta^* - \Theta_{t,k-1}\|_2^2, \\
\mathcal{L}_t(\Theta_{t,k}) &\le \mathcal{L}_t(\Theta_{t,k-1}) + \langle\nabla\mathcal{L}_t(\Theta_{t,k-1}), \Theta_{t,k} - \Theta_{t,k-1}\rangle + \frac{L}{2}\|\Theta_{t,k} - \Theta_{t,k-1}\|_2^2 \\
&\le \mathcal{L}_t(\Theta_{t,k-1}) + \langle\nabla\mathcal{L}_t(\Theta_{t,k-1}), \Theta_{t,k} - \Theta_{t,k-1}\rangle + \frac{1}{2\eta}\|\Theta_{t,k} - \Theta_{t,k-1}\|_2^2.
\end{aligned}$$

The above two inequalities suggest that

$$\mathcal{L}_t(\Theta_{t,k}) - \mathcal{L}_t(\Theta^*) \le \langle\nabla\mathcal{L}_t(\Theta_{t,k-1}), \Theta_{t,k} - \Theta^*\rangle + \frac{1}{2\eta}\|\Theta_{t,k} - \Theta_{t,k-1}\|_2^2 - \frac{\alpha}{2}\|\Theta^* - \Theta_{t,k-1}\|_2^2.$$

Meanwhile, we have

$$\begin{aligned}
&\frac{1}{2\eta}\|\Theta_{t,k} - \Theta^*\|_2^2 \\
&= \frac{1}{2\eta}\|\Theta_{t,k-1} - \Theta^*\|_2^2 - \frac{1}{2\eta}\|\Theta_{t,k} - \Theta_{t,k-1}\|_2^2 + \frac{1}{\eta}\langle\Theta_{t,k-1} - \Theta_{t,k}, \Theta^* - \Theta_{t,k}\rangle \\
&= \frac{1}{2\eta}\|\Theta_{t,k-1} - \Theta^*\|_2^2 - \frac{1}{2\eta}\|\Theta_{t,k} - \Theta_{t,k-1}\|_2^2 + \frac{1}{\eta}\langle\Theta_{t,k-1} - \eta\nabla\mathcal{L}_t(\Theta_{t,k-1}) - \Theta_{t,k}, \Theta^* - \Theta_{t,k}\rangle \\
&\quad + \langle\nabla\mathcal{L}_t(\Theta_{t,k-1}), \Theta^* - \Theta_{t,k}\rangle \\
&\le \frac{1}{2\eta}\|\Theta_{t,k-1} - \Theta^*\|_2^2 - \frac{1}{2\eta}\|\Theta_{t,k} - \Theta_{t,k-1}\|_2^2 + \frac{\gamma_{s,s^*}}{\eta}\|\Theta^* - \Theta_{t,k}\|_2^2 + \langle\nabla\mathcal{L}_t(\Theta_{t,k-1}), \Theta^* - \Theta_{t,k}\rangle,
\end{aligned}$$

where we use the update rule that $\Theta_{t,k} = \Pi_{\mathcal{C}_s}(\Theta_{t,k-1} - \eta\nabla\mathcal{L}_t(\Theta_{t,k-1}))$ and

$$\gamma_{s,s^*} = \sup\left\{\frac{\langle Z - \Pi_{\mathcal{C}_s}(Z), Y - \Pi_{\mathcal{C}_s}(Z)\rangle}{\|Y - \Pi_{\mathcal{C}_s}(Z)\|_2^2}, Y, Z\in\mathbb{R}^d, \|Y\|_0 \le s^*, Y \ne \Pi_{\mathcal{C}_s}(Z)\right\}$$

for any sparsity pair (s^*, s). By combining the above two inequalities, we have

$$\mathcal{L}_t(\Theta_{t,k}) - \mathcal{L}_t(\Theta^*) \leq \frac{1}{2\eta}\left[(1-\eta\alpha)\|\Theta_{t,k-1} - \Theta^*\|_2^2 - (1-2\gamma_{s,s^*})\|\Theta_{t,k} - \Theta^*\|_2^2\right].$$

By multiplying $2\eta\left(\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}}\right)^{K-k}$ to both sides of the above inequality and summing over $k = 1, \cdots, K$, we obtain

$$\begin{aligned}
&\sum_{k=1}^{K} 2\eta\left(\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}}\right)^{K-k}(\mathcal{L}_t(\Theta_{t,k}) - \mathcal{L}_t(\Theta^*)) \\
&\leq \sum_{k=1}^{K}\left(\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}}\right)^{K-k}\left[(1-\eta\alpha)\|\Theta_{t,k-1} - \Theta^*\|_2^2 - (1-2\gamma_{s,s^*})\|\Theta_{t,k} - \Theta^*\|_2^2\right] \\
&\leq (1-2\gamma_{s,s^*})\left(\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}}\right)^{K}\|\Theta_{t,0} - \Theta^*\|_2^2.
\end{aligned}$$

By using the descent lemma, aka Lemma 3, under the step-size that $\eta_{t,k} = \eta \leq \frac{1}{L}$, we have $\mathcal{L}_t(\Theta_{t,k+1}) \leq \mathcal{L}_t(\Theta_{t,k})$ for all $k = 0, \cdots, K-1$. Consequently, the above inequality implies that

$$\begin{aligned}
&\sum_{k=1}^{K}\left(\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}}\right)^{K-k}(\mathcal{L}_t(\Theta_{t,K}) - \mathcal{L}_t(\Theta^*)) \\
&\leq \frac{1-2\gamma_{s,s^*}}{2\eta}\left(\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}}\right)^{K}\|\Theta_{t,0} - \Theta^*\|_2^2.
\end{aligned}$$

Together with the fact that

$$\sum_{k=1}^{K}\left(\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}}\right)^{K-k} \geq 1,$$

we obtain

$$\mathcal{L}_t(\Theta_{t,K}) - \mathcal{L}_t(\Theta^*) \leq \frac{1-2\gamma_{s,s^*}}{2\eta}\left(\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}}\right)^{K}\|\Theta_{t,0} - \Theta^*\|_2^2. \tag{24}$$

Note that the above inequality holds regardless of $\mathrm{sgn}(\mathcal{L}_t(\Theta_{t,K}) - \mathcal{L}_t(\Theta^*))$. Recall that $\mathcal{S}_{\Theta_{t,K}}$ and $\mathcal{S}^*$ are the collection of non-zero entries of $\Theta_{t,K}$ and Θ^*, respectively. By using the (s,s)-SSC condition of $\mathcal{L}_t$, we further have

$$\begin{aligned}
\frac{\alpha}{2}\|\Theta_{t,K} - \Theta^*\|_2^2 &\leq \mathcal{L}_t(\Theta_{t,K}) - \mathcal{L}_t(\Theta^*) - \langle\nabla\mathcal{L}_t(\Theta^*), \Theta_{t,K} - \Theta^*\rangle \\
&= \mathcal{L}_t(\Theta_{t,K}) - \mathcal{L}_t(\Theta^*) + \frac{1}{t_0+t}\sum_{i\in\mathcal{I}_t}\langle X_i\epsilon_i, \Theta_{t,K} - \Theta^*\rangle.
\end{aligned} \tag{25}$$

By using the fact that $\|\Theta_{t,K} - \Theta^*\|_0 \leq s + s^* \leq 2s$, we further have

$$\begin{aligned}\frac{\alpha}{2}\|\Theta_{t,K} - \Theta^*\|_2^2 &\leq \mathcal{L}_t(\Theta_{t,K}) - \mathcal{L}_t(\Theta^*) + \left\|\frac{\sum_{i\in\mathcal{I}_t} X_i\epsilon_i}{t_0+t}\right\|_\infty \|\Theta_{t,K} - \Theta^*\|_1 \\ &\leq \mathcal{L}_t(\Theta_{t,K}) - \mathcal{L}_t(\Theta^*) + \sqrt{2s}\left\|\frac{\sum_{i\in\mathcal{I}_t} X_i\epsilon_i}{t_0+t}\right\|_\infty \|\Theta_{t,K} - \Theta^*\|_2 \\ &\leq \mathcal{L}_t(\Theta_{t,K}) - \mathcal{L}_t(\Theta^*) + \frac{2s}{\alpha}\left\|\frac{\sum_{i\in\mathcal{I}_t} X_i\epsilon_i}{t_0+t}\right\|_\infty^2 + \frac{\alpha}{4}\|\Theta_{t,K} - \Theta^*\|_2^2,\end{aligned}$$

which further implies that

$$\frac{\alpha}{4}\|\Theta_{t,K} - \Theta^*\|_2^2 \leq \mathcal{L}_t(\Theta_{t,K}) - \mathcal{L}_t(\Theta^*) + \frac{2s}{\alpha}\left\|\frac{\sum_{i\in\mathcal{I}_t} X_i\epsilon_i}{t_0+t}\right\|_\infty^2. \tag{26}$$

By substituting the above inequality into (24), we conclude that

$$\frac{\alpha}{4}\|\Theta_{t,K} - \Theta^*\|_2^2 \leq \frac{1-2\gamma_{s,s^*}}{2\eta}\left(\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}}\right)^K \|\Theta_{t,0} - \Theta^*\|_2^2 + \frac{2s}{\alpha}\left\|\frac{\sum_{i\in\mathcal{I}_t} X_i\epsilon_i}{t_0+t}\right\|_\infty^2.$$

We divide both sides of the above inequality by $\frac{\alpha}{4}$ and acquire the desired result. Further, by using Lemma 15 in Appendix Section B.6 that $\gamma_{s,s^*} = \sqrt{s^*/s}/2$ when $\Upsilon(\Theta) = \|\Theta\|_0$ and recalling the step-size $\eta_{t,k} = \eta \leq 1/L$ and sparsity level $s > \frac{s^*}{\eta^2\alpha^2}$, we obtain

$$2\gamma_{s,s^*} = \sqrt{s^*/s} < \eta\alpha,$$

implying that

$$\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}} < 1.$$

Skipping some algebra, we conclude that

$$\frac{2(1-2\gamma_{s,s^*})}{\alpha\eta}\left(\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}}\right)^K < 1 \iff K \geq \frac{\log\{2(1-2\gamma_{s,s^*})/(\eta\alpha)\}}{\log\{(1-2\gamma_{s,s^*})/(1-\eta\alpha)\}}.$$

This completes the proof. ■

B.5 Proof of Theorem 5

Proof Here we observe $\delta < 1$ according to (19). By taking expectations on both sides of (20) and using Assumption 2 and the fact that $\Theta_{t,0} = \Theta_{t-1,K}$, we have

$$\mathbb{E}[\|\Theta_{t,K} - \Theta^*\|_2^2] \leq \delta\mathbb{E}[\|\Theta_{t-1,K} - \Theta^*\|_2^2] + \frac{8s\sigma_x^2}{\alpha^2(t_0+t)}.$$

By recursively applying the above inequality and using Lemma 16, we have that

$$\begin{aligned}\mathbb{E}[\|\Theta_{t,K} - \Theta^*\|_2^2] &\leq \delta^{t+1}\|\Theta_0 - \Theta^*\|_2^2 + \frac{8s\sigma_x^2}{\alpha^2}\left(\sum_{j=1}^{t+1}\frac{\delta^{t-j}}{t_0+j}\right) \\ &\leq \delta^{t+1}\|\Theta_0 - \Theta^*\|_2^2 + \frac{8s\sigma_x^2}{\alpha^2\log(1/\delta)}\left(\frac{\delta^t}{t_0+1} + \frac{3}{\delta(t+2+t_0)} + \frac{\delta^{t/2}}{t_0+1}\right).\end{aligned}$$

This concludes the proof. ■

B.6 Supporting Lemmas

Proposition 14 (In-sample prediction error)**.** *Suppose $\mathcal{L}_t$ satisfies Assumption 1 under the sparsity level $s \geq s^*$, then for any $\Theta \in \mathbb{R}^d$ such that $\|\Theta\|_0 \leq s$, we have*
(a)

$$\alpha\|\Theta - \Theta^*\|_2^2 \leq \frac{1}{t_0+t}\sum_{i\in\mathcal{I}_t}\langle X_i, \Theta - \Theta^*\rangle^2 \leq L\|\Theta - \Theta^*\|_2^2. \tag{27}$$

(b)

$$\frac{\alpha}{4}\|\Theta - \Theta^*\|_2^2 \leq \mathcal{L}_t(\Theta) - \mathcal{L}_t(\widehat{\Theta}_t) + \frac{2s}{\alpha}\Big\|\sum_{j\in\mathcal{I}_t}\frac{\epsilon_j X_j}{t_0+t}\Big\|_\infty^2. \tag{28}$$

Proof (a) Suppose $\|\Theta\|_0 \leq s$ and $\|\Theta^*\|_0 \leq s$. Under the (s,s)-SSS condition, we have

$$\begin{aligned}
\mathcal{L}_t(\Theta) - \mathcal{L}_t(\Theta^*) &= \frac{1}{2(t+t_0)}\sum_{i\in\mathcal{I}_t}\Big((\langle X_i,\Theta\rangle - Y_i)^2 - \big(\langle X_i,\Theta^*\rangle - Y_i\big)^2\Big)\\
&= \frac{1}{2(t+t_0)}\sum_{i\in\mathcal{I}_t}\Big(\big(\langle X_i,\Theta-\Theta^*\rangle - \epsilon_i\big)^2 - \epsilon_i^2\Big)\\
&= \frac{1}{2(t+t_0)}\sum_{i\in\mathcal{I}_t}\Big(\langle X_i,\Theta-\Theta^*\rangle^2 - 2\epsilon_i\langle X_i,\Theta-\Theta^*\rangle\Big)\\
&\geq \langle\nabla\mathcal{L}_t(\Theta^*),\Theta-\Theta^*\rangle + \frac{\alpha}{2}\|\Theta-\Theta^*\|_2^2.
\end{aligned}$$

Using the fact that $\nabla\mathcal{L}_t(\Theta^*) = -\frac{1}{t_0+t}\sum_{i\in\mathcal{I}_t}X_i\epsilon_i$, the above inequality implies

$$\frac{1}{t_0+t}\sum_{i\in\mathcal{I}_t}\langle X_i,\Theta-\Theta^*\rangle^2 \geq \alpha\|\Theta-\Theta^*\|_2^2.$$

Similarly, under Assumption 1 and using the facts that $\|\Theta\|_0 \leq s$ and $\|\Theta^*\|_0 \leq s$, we have

$$\begin{aligned}
\mathcal{L}_t(\Theta) - \mathcal{L}_t(\Theta^*) &= \frac{1}{2(t+t_0)}\sum_{i\in\mathcal{I}_t}\Big(\langle X_i,\Theta-\Theta^*\rangle^2 - 2\epsilon_i\langle X_i,\Theta-\Theta^*\rangle\Big)\\
&\leq \langle\nabla\mathcal{L}_t(\Theta^*),\Theta-\Theta^*\rangle + \frac{L}{2}\|\Theta-\Theta^*\|_2^2.
\end{aligned}$$

This implies

$$\frac{1}{t_0+t}\sum_{i\in\mathcal{I}_t}\langle X_i,\Theta-\Theta^*\rangle^2 \leq L\|\Theta-\Theta^*\|_2^2,$$

completing the proof of part (a).

(b) We observe that

$$\begin{aligned}
&\mathcal{L}_t(\Theta) - \mathcal{L}_t(\widehat{\Theta}_t) \geq \mathcal{L}_t(\Theta) - \mathcal{L}_t(\Theta^*) \\
&= \frac{1}{2(t+t_0)} \sum_{i \in \mathcal{I}_t} \Big(\Big(\langle X_j, \Theta \rangle - Y_j \Big)^2 - (\langle X_j, \Theta_j^* \rangle - Y_j)^2 \Big) \\
&= \frac{1}{2(t+t_0)} \sum_{i \in \mathcal{I}_t} \Big(\Big(\langle X_j, \Theta - \Theta^* \rangle - \epsilon_j \Big)^2 - \epsilon_j^2 \Big) \\
&= \frac{1}{2(t+t_0)} \sum_{i \in \mathcal{I}_t} \Big(\langle X_j, \Theta - \Theta^* \rangle^2 - 2\epsilon_j \langle X_j, \Theta - \Theta^* \rangle \Big) \\
&\geq \frac{\alpha}{2} \|\Theta - \Theta^*\|_2^2 - \sum_{j \in \mathcal{I}_t} \frac{\epsilon_j}{t+t_0} \langle X_j, \Theta - \Theta^* \rangle \\
&\geq \frac{\alpha}{2} \|\Theta - \Theta^*\|_2^2 - \frac{2s}{\alpha} \Big\| \sum_{j \in \mathcal{I}_t} \frac{\epsilon_j X_j}{t_0 + t} \Big\|_\infty - \frac{\alpha}{4} \|\Theta - \Theta^*\|_2^2
\end{aligned}$$

where the second last inequality uses part (a) and the last inequality uses the fact that $\|\Theta - \Theta^*\|_0 \leq 2s$ and

$$\langle a, b \rangle \leq \sqrt{2s} \|a\|_2 \|b\|_\infty \leq \frac{\alpha}{4} \|a\|_2^2 + \frac{2s}{\alpha} \|b\|_\infty^2$$

when $\|a\|_0 \leq 2s$. As a result, we conclude that

$$\frac{\alpha}{4} \|\Theta - \Theta^*\|_2^2 \leq \mathcal{L}_t(\Theta) - \mathcal{L}_t(\widehat{\Theta}_t) + \frac{2s}{\alpha} \Big\| \sum_{j \in \mathcal{I}_t} \frac{\epsilon_j X_j}{t_0 + t} \Big\|_\infty^2,$$

completing the proof. ■

Lemma 15. *(Liu and Foygel Barber, 2020, Lemma 4.1) When $\Upsilon(\Theta) = \|\Theta\|_0$ and a s-sparse hard thresholding operator $\Pi_{\mathcal{C}_s}$, for any sparsity pair (s, s^*) where $0 < s^* \leq s$, the relative concavity to sparsity level s_0 is*

$$\gamma_{s,s^*} = \sup \left\{ \frac{\langle Y - \Pi_{\mathcal{C}_s}(Z), Z - \Pi_{\mathcal{C}_s}(Z) \rangle}{\|Y - \Pi_{\mathcal{C}_s}(Z)\|_2^2}, Y, Z \in \mathbb{R}^d, \|Y\|_0 \leq s^*, Y \neq \Pi_{\mathcal{C}_s}(Z) \right\} = \frac{\sqrt{s^*/s}}{2}.$$

Lemma 16. *For any $\gamma \in (0, 1)$, the following holds*

$$\sum_{t=1}^{T} \frac{\gamma^{T-t}}{t + t_0} \leq \frac{\gamma^{T-1}}{(t_0 + 1)} + \frac{3}{\gamma(T + 1 + t_0)} + \frac{\gamma^{(T+1)/2-1}}{1 + t_0}.$$

Proof Our proof consists of two steps. In Step 1, we provide an upper bound for the term $\sum_{t=1}^{T} \frac{\gamma^{T-t}}{t+t_0}$ such that

$$\sum_{t=1}^{T} \frac{\gamma^{T-t}}{t_0 + t} \leq \frac{\gamma^{T-1}}{(t_0 + 1) \log(1/\gamma)} + \gamma^T \int_1^{T+1} \frac{1}{(x + t_0)\gamma^x} dx.$$

In Step 2, we further provide a bound on the above integration and acquire the desired result.

Step 1: We note that $\sum_{t=1}^{T} \frac{\gamma^{T-t}}{t_0+t} = \gamma^T \sum_{t=1}^{T} \frac{1}{(t_0+t)\gamma^t}$. We denote by $f(t) := \log(\frac{1}{(t_0+t)\gamma^t}) = -\log(t_0+t) - t\log\gamma$. By utilizing the monotonicity of $\log(\cdot)$, we observe that $f(t)$ and $\frac{1}{(t_0+t)\gamma^t}$ share the same minimizer. By setting $f'(t)$ to be zero, we have

$$f'(\widehat{t}) = -\frac{1}{t_0+\widehat{t}} - \log\gamma = 0 \iff \widehat{t} = (\log(1/\gamma))^{-1} - t_0.$$

We can see that $f(t)$ is decreasing for $t \le \widehat{t}$ and increasing for $t \ge \widehat{t}$. We consider two scenarios.

- **Scenario 1:** Suppose $\widehat{t} = (\log(1/\gamma))^{-1} - t_0 \le 1$. Then we observe that $f(t)$ is monotonically increasing over $[1,\infty)$. In this case, clearly, we have

$$\sum_{t=1}^{T} \frac{\gamma^{T-t}}{t_0+t} \le \gamma^T \int_1^{T+1} \frac{1}{(t_0+x)\gamma^x} dx.$$

- **Scenario 2:** Suppose $\widehat{t} = (\log(1/\gamma))^{-1} - t_0 > 1$. We observe that $f(t)$ is decreasing within $t \in [1, \lfloor\widehat{t}\rfloor]$ and increasing within $[\lceil\widehat{t}\rceil, \infty)$. By utilizing this observation, we have

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\gamma^{T-t}}{t_0+t} &\le \gamma^T \left[\sum_{t=1}^{\lfloor\widehat{t}\rfloor} \frac{1}{(t_0+t)\gamma^t} + \int_{\lceil\widehat{t}\rceil}^{T+1} \frac{1}{(x+t_0)\gamma^x} dx \right] \\
&\le \gamma^T \left[\frac{\lfloor\widehat{t}\rfloor}{(t_0+1)\gamma} + \int_{\lceil\widehat{t}\rceil}^{T+1} \frac{1}{(x+t_0)\gamma^x} dx \right] \\
&\le \gamma^T \left[\left(\frac{1}{\log(1/\gamma)} - t_0 \right) \frac{1}{(t_0+1)\gamma} + \int_1^{T+1} \frac{1}{(x+t_0)\gamma^x} dx \right],
\end{aligned}$$

  where the second inequality comes from the fact that $f(t)$ is decreasing for $t \in [1, \lfloor\widehat{t}\rfloor]$ so that $\frac{1}{(t+t_0)\gamma^t} \le \frac{1}{(t_0+1)\gamma}$.

By combining the above two scenarios, we conclude that

$$\sum_{t=1}^{T} \frac{\gamma^{T-t}}{t_0+t} \le \frac{\gamma^{T-1}}{(t_0+1)\log(1/\gamma)} + \gamma^T \int_1^{T+1} \frac{1}{(x+t_0)\gamma^x} dx. \tag{29}$$

Step 2: By setting $c = \log(1/\gamma) > 0$, we have $\exp(c) = \frac{1}{\gamma}$ and obtain the following.

$$\begin{aligned}
&\gamma^{T+1} \int_1^{T+1} \frac{1}{(x+t_0)\gamma^x} dx \\
&= \gamma^{T+1} \int_1^{T+1} (x+t_0)^{-1} \exp(cx) dx \\
&= \gamma^{T+1} \Big(\frac{\exp(cx)}{c(x+t_0)} \Big|_{x=1}^{x=T+1} + \int_1^{T+1} \frac{e^{cx}}{c(x+t_0)^2} dx \Big) \\
&= \frac{1}{c(T+1+t_0)} - \frac{\gamma^{T+1} e^c}{c(t_0+1)} + \gamma^{T+1} \int_1^{T+1} \frac{e^{cx}}{c(x+t_0)^2} dx \\
&\le \frac{1}{c(T+1+t_0)} - \frac{\gamma^T}{c(t_0+1)} + \underbrace{\frac{\gamma^{T+1}}{c} \int_{\frac{T+1}{2}}^{T+1} \frac{e^{cx}}{(x+t_0)^2} dx}_{\mathrm{I}} + \underbrace{\frac{\gamma^{T+1}}{c} \int_1^{\frac{T+1}{2}} \frac{e^{cx}}{(x+t_0)^2} dx}_{\mathrm{II}}.
\end{aligned}$$

For the last two terms, by using the fact that $e^{c(T+1)} = \frac{1}{\gamma^{T+1}}$, we can see that

$$\mathrm{I} \le \frac{\gamma^{T+1}}{c} \int_{\frac{T+1}{2}}^{T+1} \frac{e^{c(T+1)}}{(x+t_0)^2} dx \le \int_{\frac{T+1}{2}}^{T+1} \frac{1}{c(x+t_0)^2} dx = -\frac{1}{c(x+t_0)} \Big|_{x=\frac{T+1}{2}}^{x=T+1} \le \frac{2}{c(T+1+2t_0)},$$

and

$$\begin{aligned}
\mathrm{II} &\le \frac{\gamma^{T+1}}{c} \int_1^{\frac{T+1}{2}} \frac{e^{cx}}{(x+t_0)^2} dx \\
&\le \frac{\gamma^{T+1}}{c} \int_1^{\frac{T+1}{2}} \frac{\gamma^{-(T+1)/2}}{(x+t_0)^2} dx = \frac{\gamma^{(T+1)/2}}{c} \int_1^{\frac{T+1}{2}} \frac{1}{(x+t_0)^2} dx \\
&= -\frac{\gamma^{(T+1)/2}}{c(x+t_0)} \Big|_{x=1}^{x=\frac{T+1}{2}} = -\frac{\gamma^{(T+1)/2}}{c} \Big(\frac{2}{T+1+2t_0} - \frac{1}{1+t_0} \Big) \\
&= \frac{\gamma^{(T+1)/2}}{c} \left(\frac{1}{1+t_0} - \frac{2}{T+1+2t_0} \right),
\end{aligned}$$

where the second inequality uses the fact that $e^{cx} = \gamma^{-x} \le \gamma^{-(T+1)/2}$ for $x \in [1, \frac{T+1}{2}]$. By combining the above terms, we obtain

$$\begin{aligned}
&\gamma^{T+1} \int_1^{T+1} \frac{1}{(x+t_0)\gamma^x} dx \\
&\le \frac{1}{c(T+1+t_0)} - \frac{\gamma^T}{c(t_0+1)} + \frac{2}{c(T+1+2t_0)} + \frac{\gamma^{(T+1)/2}}{c} \left(\frac{1}{1+t_0} - \frac{2}{T+1+2t_0} \right) \\
&\le \frac{3}{c(T+1+t_0)} + \frac{\gamma^{(T+1)/2}}{c(1+t_0)}.
\end{aligned}$$

By substituting the above inequality into (29), we obtain

$$\sum_{t=1}^{T} \frac{\gamma^{T-t}}{t+t_0} \le \frac{1}{\log(1/\gamma)} \left(\frac{\gamma^{T-1}}{(t_0+1)} + \frac{3}{\gamma(T+1+t_0)} + \frac{\gamma^{(T+1)/2-1}}{1+t_0} \right),$$

completing the proof. ■

Appendix C. Proofs for Section 3.2

C.1 Proof of Lemma 6

Proof Recall that $\widehat{\Theta}_t = \arg\min_{\mathrm{rank}(\Theta)\leq s^*} \mathcal{L}_t(\Theta)$ and $\mathrm{rank}(\widehat{\Theta}_t - \Theta^*) \leq 2s^*$. Suppose $\mathcal{L}_t(\cdot)$ satisfies the $2s$-LowRankSC condition. By following the analysis of Lemma 2 for sparse linear regression, we note that (22) still holds for low-rank matrix sensing. Specifically, we have Proposition 17 (a) that

$$\frac{\alpha}{2}\|\widehat{\Theta}_t - \Theta^*\|_{\mathrm{F}}^2 \leq \frac{1}{2(t_0+t)}\sum_{i\in\mathcal{I}_t}\left\langle X_i, \widehat{\Theta}_t - \Theta^*\right\rangle^2 \leq \frac{1}{t_0+t}\sum_{i\in\mathcal{I}_t}\epsilon_i\left\langle X_i, \widehat{\Theta}_t - \Theta^*\right\rangle.$$

For any matrices $A, B \in \mathbb{R}^{d_1\times d_2}$ where $\mathrm{rank}(A) \leq 2s$, we can see that

$$\langle A, B\rangle \leq \|A\|_{\mathrm{s}_1}\|B\|_{\mathrm{s}_\infty} \leq \sqrt{2s}\|A\|_{\mathrm{F}}\|B\|_2 \leq \frac{\alpha}{4}\|A\|_{\mathrm{F}}^2 + \frac{2s}{\alpha}\|B\|_2^2, \tag{30}$$

where $\|A\|_{\mathrm{s}_p}$ is the Schatten p-norm of A and the second inequality uses $\|A\|_{\mathrm{s}_1} \leq \sqrt{\mathrm{rank}(A)}\|A\|_{\mathrm{F}}$ for any matrix A and the last inequality uses the fact that $ab \leq \frac{a^2}{\alpha} + \frac{\alpha b^2}{4}$. Combining the above inequalities and using the fact that $\mathrm{rank}(\widehat{\Theta}_t - \Theta^*) \leq 2s^*$, we have

$$\begin{aligned}
\frac{\alpha}{2}\|\widehat{\Theta}_t - \Theta^*\|_{\mathrm{F}}^2 &\leq \frac{1}{t_0+t}\sum_{i\in\mathcal{I}_t}\epsilon_i\left\langle X_i, \widehat{\Theta}_t - \Theta^*\right\rangle \\
&\leq \Big\|\frac{1}{t_0+t}\sum_{i\in\mathcal{I}_t}\epsilon_i X_i\Big\|_{\mathrm{s}_\infty}\|\widehat{\Theta}_t - \Theta^*\|_{\mathrm{s}_1} \\
&\leq \sqrt{2s^*}\Big\|\frac{1}{t_0+t}\sum_{i\in\mathcal{I}_t}\epsilon_i X_i\Big\|_2\|\widehat{\Theta}_t - \Theta^*\|_{\mathrm{F}} \\
&\leq \frac{2s^*}{\alpha}\Big\|\frac{1}{t_0+t}\sum_{i\in\mathcal{I}_t}\epsilon_i X_i\Big\|_2^2 + \frac{\alpha}{4}\|\widehat{\Theta}_t - \Theta^*\|_{\mathrm{F}}^2.
\end{aligned}$$

Rearranging the terms, we have

$$\|\widehat{\Theta}_t - \Theta^*\|_{\mathrm{F}}^2 \leq \frac{8s^*}{\alpha^2}\Big\|\sum_{i\in\mathcal{I}_t}\frac{\epsilon_i X_i}{t_0+t}\Big\|_2^2.$$

By using Assumption 4 that $\mathbb{E}\big[\big\|\sum_{i\in\mathcal{I}_t}\epsilon_i X_i\big\|_2^2\big] \leq (t_0+t)\sigma_X^2$ and taking expectations on both sides of the above inequality, we conclude that

$$\mathbb{E}[\|\widehat{\Theta}_t - \Theta^*\|_{\mathrm{F}}^2] \leq \frac{8s^*}{\alpha^2}\mathbb{E}\Big[\Big\|\sum_{i\in\mathcal{I}_t}\frac{\epsilon_i X_i}{t_0+t}\Big\|_2^2\Big] \leq \frac{8s^*\sigma_X^2}{\alpha^2(t_0+t)}.$$

This completes the proof. ■

C.2 Proof of Lemma 7

Proof We denote by $\Theta_{t,k}$ and $\Theta_{t,k+1}$ two consecutive rank-s solutions generated by Algorithm 1 with projection operator $\Pi_{\mathcal{C}_s}(\Theta)$ returning the rank-s approximation of Θ and denote by $\mathcal{L}_t(\cdot)$ the corresponding loss function. By Assumption 1, $\mathcal{L}_t(\cdot)$ satisfies the $2s$-LowRankSS condition (1) with parameter L. For notational convenience, we write $\eta_{t,k} = \eta$ and rewrite the update rule as

$$\Theta_{t,k+1} = \underset{Z\in\mathbb{R}^{d_1\times d_2},\mathrm{rank}(Z)\leq s}{\operatorname{argmin}} \|Z - (\Theta_{t,k} - \eta\nabla\mathcal{L}_t(\Theta_{t,k}))\|_{\mathrm{F}}^2 .$$

For simplicity, we denote by $S_{t,k}$ and $S_{t,k+1}$ the column spaces of $\Theta_{t,k}$ and $\Theta_{t,k+1}$, respectively. Let $g_k = \nabla\mathcal{L}(\Theta_{t,k})$. We then bound the change of the objective function value from $\mathcal{L}_t(\Theta_{t,k})$ to $\mathcal{L}_t(\Theta_{t,k+1})$ as

$$\begin{aligned}\mathcal{L}_t(\Theta_{t,k+1}) - \mathcal{L}_t(\Theta_{t,k}) &\leq \langle\nabla\mathcal{L}_t(\Theta_{t,k+1}), \Theta_{t,k+1} - \Theta_{t,k}\rangle_{\mathrm{F}} + \frac{L}{2}\|\Theta_{t,k+1} - \Theta_{t,k}\|_{\mathrm{F}}^2 \\ &= \langle g_k, \Theta_{t,k+1} - \Theta_{t,k}\rangle_{\mathrm{F}} + \frac{L}{2}\|\Theta_{t,k+1} - \Theta_{t,k}\|_{\mathrm{F}}^2 \\ &= \mathrm{I} + \mathrm{II}, \end{aligned} \tag{31}$$

where $\mathrm{I} = \langle g_k, \Theta_{t,k+1} - \Theta_{t,k}\rangle_{\mathrm{F}}$ and $\mathrm{II} = \frac{L}{2}\|\Theta_{t,k+1} - \Theta_{t,k}\|_{\mathrm{F}}^2$. We proceed to bound I and II respectively. We start with I and have

$$\begin{aligned}\mathrm{I} &= \left\langle P_{S_k+S_{k+1}} g_k, P_{S_k+S_{k+1}}(\Theta_{t,k+1} - \Theta_{t,k})\right\rangle_{\mathrm{F}} \\ &= -\left\langle P_{S_{t,k}\cap S_{t,k+1}^\perp}\Theta_{t,k}, P_{S_t\cap S_{t,k+1}^\perp} g_t\right\rangle + \left\langle P_{S_{t,k+1}}(\Theta_{t,k+1} - \Theta_{t,k}), P_{S_{t,k+1}} g_k\right\rangle \\ &= -\left\langle P_{S_{t,k}\cap S_{t,k+1}^\perp}\Theta_{t,k}, P_{S_k\cap S_{t,k+1}^\perp} g_k\right\rangle + \left\langle P_{S_{t,k+1}}(P_{S_{t,k+1}}(\Theta_{t,k} - \eta g_k) - \Theta_{t,k}), P_{S_{t,k+1}} g_t\right\rangle \\ &= -\left\langle P_{S_{t,k}\cap S_{t,k+1}^\perp}\Theta_{t,k}, P_{S_{t,k}\cap S_{t,k+1}^\perp} g_k\right\rangle - \eta\|P_{S_{t,k+1}} g_k\|_{\mathrm{F}}^2 \\ &= -\left\langle P_{S_{t,k}\cap S_{t,k+1}^\perp}\Theta_{t,k} - \eta P_{S_{t,k}\cap S_{t,k+1}^\perp} g_k, P_{S_{t,k}\cap S_{t,k+1}^\perp} g_k\right\rangle - \eta\|P_{S_{t,k}\cap S_{t,k+1}^\perp} g_k\|_{\mathrm{F}}^2 - \eta\|P_{S_{t,k+1}} g_k\|_{\mathrm{F}}^2 \\ &\leq \frac{1}{2\eta}\|P_{S_{t,k}\cap S_{t,k+1}^\perp}\Theta_{t,k} - \eta P_{S_{t,k}\cap S_{t,k+1}^\perp} g_k\|_{\mathrm{F}}^2 - \frac{\eta}{2}\|P_{S_{t,k}\cap S_{t,k+1}^\perp} g_k\|_{\mathrm{F}}^2 - \eta\|P_{S_{t,k+1}} g_k\|_{\mathrm{F}}^2 \\ &\leq \frac{\eta}{2}\|P_{S_{t,k+1}\cap S_{t,k}^\perp} g_k\|_{\mathrm{F}}^2 - \frac{\eta}{2}\|P_{S_{t,k}\cap S_{t,k+1}^\perp} g_k\|_{\mathrm{F}}^2 - \eta\|P_{S_{t,k+1}} g_k\|_{\mathrm{F}}^2 \\ &\leq -\frac{\eta}{2}\|P_{S_{t,k}+S_{t,k+1}} g_k\|_{\mathrm{F}}^2, \end{aligned}$$

where the third equality comes from $\Theta_{t,k+1} = P_{S_{t,k+1}}(\Theta_{t,k} - \eta g_k)$, and the last second inequality follows from the fact

$$\|P_{S_{t,k}\cap S_{t,k+1}^\perp}(\Theta_{t,k} - \eta g_k)\|_{\mathrm{F}}^2 \leq \|P_{S_{t,k+1}\cap S_{t,k}^\perp}\Theta_{t,k+1}\|_{\mathrm{F}}^2 = \eta^2\|P_{S_{t,k+1}\cap S_{t,k}^\perp} g_k\|_{\mathrm{F}}^2.$$

For II, by using the fact that $\|a\|_{\mathrm{F}}^2 = \|a+b\|_{\mathrm{F}}^2 - \|b\|_{\mathrm{F}}^2 - 2\langle a, b\rangle$, we note that

$$\begin{aligned}\Theta_{t,k+1} &= \underset{Z:\mathrm{rank}(Z)\leq r}{\operatorname{argmin}} \left\{\|Z - \Theta_{t,k} + \eta g_k\|_{\mathrm{F}}^2\right\} \\ &= \operatorname{argmin}\left\{\|Z - \Theta_{t,k} + \eta g_k\|_{\mathrm{F}}^2 \;\middle|\; \mathrm{rank}(Z)\leq r, Z\in S_{t,k} + S_{t,k+1}\right\} \\ &= \underset{Z:\mathrm{rank}(Z)\leq r}{\operatorname{argmin}} \|P_{S_{t,k}+S_{t,k+1}}(Z - \Theta_{t,k} + \eta g_k)\|_{\mathrm{F}}^2 \leq \|\eta P_{S_{t,k}+S_{t,k+1}} g_k\|_{\mathrm{F}}^2, \end{aligned}$$

which yields that

$$\begin{aligned}
\mathrm{II} &= \frac{L}{2}\left\|\Theta_{t,k+1}-\Theta_{t,k}\right\|_{\mathrm{F}}^2 \\
&= \frac{L}{2}\left\|\Theta_{t,k+1}-\Theta_{t,k}+\eta P_{S_{t,k}+S_{t,k+1}}g_k\right\|_{\mathrm{F}}^2-\frac{L}{2}\left\|\eta P_{S_{t,k}+S_{t,k+1}}g_k\right\|_{\mathrm{F}}^2 \\
&\quad -L\eta\cdot\left\langle P_{S_{t,k}+S_{t,k+1}}g_k,\Theta_{t,k+1}-\Theta_{t,k}\right\rangle \\
&= \frac{L}{2}\inf_{Z:\mathrm{rank}(Z)\le r}\left\|Z-\Theta_{t,k}+\eta P_{S_{t,k}+S_{t,k+1}}g_t\right\|_{\mathrm{F}}^2-\frac{L}{2}\left\|\eta P_{S_{t,k}+S_{t,k+1}}g_k\right\|_{\mathrm{F}}^2-L\eta\cdot\mathrm{I} \\
&\le -L\eta\cdot\mathrm{I}.
\end{aligned}$$

Plugging the above bounds for I and II into (31), we obtain that

$$\mathcal{L}_t(\Theta_{t,k+1})-\mathcal{L}_t(\Theta_{t,k})\le(1-L\eta)\cdot\mathrm{I}\le-\frac{\eta(1-L\eta)}{2}\|P_{S_{t,k}+S_{t,k+1}}g_k\|_{\mathrm{F}}^2,$$

completing the proof. ■

C.3 Proof of Lemma 8

Proof (a) We consider the case where we conduct K gradient descent steps upon receiving the t-th data. Recall the descent Lemma 7, when $\eta_{t,k}=\eta\le 1/L$, we have

$$\mathcal{L}_t(\Theta_{t,k+1})-\mathcal{L}_t(\Theta_{t,k})\le-\frac{\eta(1-L\eta)}{2}\|P_{\mathcal{S}_{\Theta_{t,k}}+\mathcal{S}_{\Theta_{t,k+1}}}\nabla\mathcal{L}_t(\Theta_{t,k})\|_{\mathrm{F}}^2\le 0.$$

Using the $2s$-LowRank SC and SS conditions and the fact that $\mathrm{rank}(\Theta_{t,k-1}),\mathrm{rank}(\Theta_{t,k})$, $\mathrm{rank}(\Theta^*)\le s$, we see that (24) and (25) still hold for low-rank matrix sensing. Specifically, for low-rank matrix regression, we rewrite (24) as

$$\mathcal{L}_t(\Theta_{t,K})-\mathcal{L}_t(\Theta^*)\le\frac{1-2\gamma_{s,s^*}}{2\eta}\left(\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}}\right)^K\|\Theta_{t,0}-\Theta^*\|_{\mathrm{F}}^2, \tag{32}$$

where

$$\gamma_{s,s^*}=\sup\left\{\frac{\langle Z-\Pi_{\mathcal{C}_s}(Z),Y-\Pi_{\mathcal{C}_s}(Z)\rangle}{\|Y-\Pi_{\mathcal{C}_s}(Z)\|_{\mathrm{F}}^2},Y,Z\in\mathbb{R}^{d_1\times d_2},\mathrm{rank}(Y)\le s^*,Y\neq\Pi_{\mathcal{C}_s}(Z)\right\}$$

for any sparsity pair (s^*,s). As in (25), using (30) and the fact that $\mathrm{rank}(\Theta_{t,K}-\Theta^*)\le 2s$, we have

$$\begin{aligned}
\frac{\alpha}{2}\|\Theta_{t,K}-\Theta^*\|_{\mathrm{F}}^2 &\le \mathcal{L}_t(\Theta_{t,K})-\mathcal{L}_t(\Theta^*)+\frac{1}{t_0+t}\sum_{i\in\mathcal{I}_t}\langle\epsilon_iX_i,\Theta_{t,K}-\Theta^*\rangle \\
&\le \mathcal{L}_t(\Theta_{t,K})-\mathcal{L}_t(\Theta^*)+\frac{2s}{\alpha}\left\|\frac{\sum_{i\in\mathcal{I}_t}\epsilon_iX_i}{t_0+t}\right\|_2^2+\frac{\alpha}{4}\|\Theta_{t,K}-\Theta^*\|_{\mathrm{F}}^2,
\end{aligned}$$

which further implies that

$$\frac{\alpha}{4}\|\Theta_{t,K}-\Theta^*\|_{\mathrm{F}}^2\le\mathcal{L}_t(\Theta_{t,K})-\mathcal{L}_t(\Theta^*)+\frac{2s}{\alpha}\left\|\frac{\sum_{i\in\mathcal{I}_t}\epsilon_iX_i}{t_0+t}\right\|_2^2. \tag{33}$$

By combining (32) and (33), we conclude that

$$\frac{\alpha}{4}\|\Theta_{t,K}-\Theta^*\|_{\mathrm{F}}^2 \le \frac{1-2\gamma_{s,s^*}}{2\eta}\left(\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}}\right)^K \|\Theta_{t,0}-\Theta^*\|_{\mathrm{F}}^2 + \frac{2s}{\alpha}\left\|\frac{\sum_{i\in\mathcal{I}_t}\epsilon_i X_i}{t_0+t}\right\|_2^2.$$

We divide both sides of the above inequality by $\frac{\alpha}{4}$ and obtain that

$$\|\Theta_{t,K}-\Theta^*\|_{\mathrm{F}}^2 \le \frac{2(1-2\gamma_{s,s^*})}{\alpha\eta}\left(\frac{1-\eta\alpha}{1-2\gamma_{s,s^*}}\right)^K \|\Theta_{t,0}-\Theta^*\|_{\mathrm{F}}^2 + \frac{8s}{\alpha^2}\left\|\frac{\sum_{i\in\mathcal{I}_t}\epsilon_i X_i}{t_0+t}\right\|_2^2.$$

(b) Further, by using Lemma 18 that $\gamma_{s,s^*} = \sqrt{s^*/s}/2$ when $\Upsilon(\Theta) = \mathrm{rank}(\Theta)$ and recalling the step-size $\eta \le 1/L$ and rank level $s > \frac{s^*}{\eta^2\alpha^2}$, we obtain the desired result by following the analysis of Lemma 4 (b). ■

C.4 Proof of Theorem 9

Proof Here we observe $\widetilde{\delta} < 1$ by recalling (19). By applying Lemma B.4 (a), taking expectations on both sides, and using Assumption 4 and the fact that $\Theta_{t,0} = \Theta_{t-1,K}$, we have

$$\mathbb{E}[\|\Theta_{t,K}-\Theta^*\|_{\mathrm{F}}^2] \le \widetilde{\delta}\,\mathbb{E}[\|\Theta_{t-1,K}-\Theta^*\|_{\mathrm{F}}^2] + \frac{8s\sigma_X^2}{\alpha^2(t_0+t)}.$$

By recursively applying the above inequality and using Lemma B.6, we have that

$$\begin{aligned}\mathbb{E}[\|\Theta_{t,K}-\Theta^*\|_{\mathrm{F}}^2] &\le \widetilde{\delta}^{t+1}\|\Theta_0-\Theta^*\|_{\mathrm{F}}^2 + \frac{8s\sigma_X^2}{\alpha^2}\left(\sum_{j=1}^{t+1}\frac{\widetilde{\delta}^{t-j}}{t_0+j}\right)\\ &\le \widetilde{\delta}^{t+1}\|\Theta_0-\Theta^*\|_{\mathrm{F}}^2 + \frac{8s\sigma_x^2}{\alpha^2\log(1/\widetilde{\delta})}\left(\frac{\widetilde{\delta}^t}{t_0+1} + \frac{3}{\widetilde{\delta}(t+2+t_0)} + \frac{\widetilde{\delta}^{t/2}}{t_0+1}\right).\end{aligned}$$

This concludes the proof. ■

C.5 Supporting Lemmas

Proposition 17. *Suppose $\mathcal{L}_t$ satisfies Assumption 3 under the rank level $s \ge s^*$, then for any $\Theta \in \mathbb{R}^{d_1\times d_2}$ such that rank$(\Theta) \le s$, we have*
(a)

$$\alpha\|\Theta-\Theta^*\|_{\mathrm{F}}^2 \le \frac{1}{t_0+t}\sum_{i\in\mathcal{I}_t}\langle X_i, \Theta-\Theta^*\rangle^2 \le L\|\Theta-\Theta^*\|_{\mathrm{F}}^2. \tag{34}$$

(b)

$$\frac{\alpha}{4}\|\Theta-\Theta^*\|_{\mathrm{F}}^2 \le \mathcal{L}_t(\Theta) - \mathcal{L}_t(\widehat{\Theta}_t) + \frac{2s}{\alpha}\Big\|\sum_{j\in\mathcal{I}_t}\frac{\epsilon_j X_j}{t_0+t}\Big\|_2^2. \tag{35}$$

Proof (a) Clearly, we can see that $\text{rank}(\Theta) \leq s$ and $\text{rank}(\Theta^*) \leq s^*$. Under Assumption 1, part (a) can be acquired by using a similar analysis to that of of Proposition 14.
(b) We observe that

$$\begin{aligned}
&\mathcal{L}_t(\Theta) - \mathcal{L}_t(\widehat{\Theta}_t) \geq \mathcal{L}_t(\Theta) - \mathcal{L}_t(\Theta^*) \\
&= \frac{1}{2(t+t_0)} \sum_{i \in \mathcal{I}_t} \Big(\Big(\langle X_j, \Theta \rangle - Y_j \Big)^2 - (\langle X_j, \Theta_j^* \rangle - Y_j)^2 \Big) \\
&= \frac{1}{2(t+t_0)} \sum_{i \in \mathcal{I}_t} \Big(\Big(\langle X_j, \Theta - \Theta^* \rangle - \epsilon_j \Big)^2 - \epsilon_j^2 \Big) \\
&= \frac{1}{2(t+t_0)} \sum_{i \in \mathcal{I}_t} \Big(\langle X_j, \Theta - \Theta^* \rangle^2 - 2\epsilon_j \langle X_j, \Theta - \Theta^* \rangle \Big) \\
&\geq \frac{\alpha}{2} \|\Theta - \Theta^*\|_{\text{F}}^2 - \sum_{j \in \mathcal{I}_t} \frac{\epsilon_j}{t+t_0} \langle X_j, \Theta - \Theta^* \rangle \\
&\geq \frac{\alpha}{2} \|\Theta - \Theta^*\|_{\text{F}}^2 - \frac{2s}{\alpha} \Big\| \sum_{j \in \mathcal{I}_t} \frac{\epsilon_j X_j}{t_0 + t} \Big\|_2^2 - \frac{\alpha}{4} \|\Theta - \Theta^*\|_{\text{F}}^2,
\end{aligned}$$

where the second last inequality uses part (a) and the last inequality uses the fact that $\text{rank}(\Theta - \Theta^*) \leq 2s$ and applying (30) that

$$\langle A, B \rangle \leq \|A\|_{\text{s}_1} \|B\|_{\text{s}_\infty} \leq \sqrt{2s} \|A\|_{\text{F}} \|B\|_2 \leq \frac{\alpha}{4} \|A\|_{\text{F}}^2 + \frac{2s}{\alpha} \|B\|_2^2,$$

where $\|A\|_{\text{s}_p}$ is the Schatten p-norm of A and the second inequality uses $\|A\|_{\text{s}_1} \leq \sqrt{\text{rank}(A)} \|A\|_{\text{F}}$. As a result, we conclude that

$$\frac{\alpha}{4} \|\Theta - \Theta^*\|_{\text{F}}^2 \leq \mathcal{L}_t(\Theta) - \mathcal{L}_t(\widehat{\Theta}_t) + \frac{2s}{\alpha} \Big\| \sum_{j \in \mathcal{I}_t} \frac{\epsilon_j X_j}{t_0 + t} \Big\|_2^2,$$

completing the proof. ■

Lemma 18. *(Liu and Foygel Barber, 2020, Lemma 5.2) For any rank pair (s, s^*), the relative concavity of the low-rank projection operator is*

$$\gamma_{s,s^*} = \sup \left\{ \frac{\langle Y - \Pi_{\mathcal{C}_s}(Z), Z - \Pi_{\mathcal{C}_s}(Z) \rangle}{\|Y - \Pi_{\mathcal{C}_s}(Z)\|^2}, Y, Z \in \mathbb{R}^{d_1 \times d_2}, \textit{rank}(Y) \leq s^*, Y \neq \Pi_{\mathcal{C}_s}(Z) \right\} = \frac{\sqrt{s^*/s}}{2},$$

where $0 < s^* \leq s$ and $\Pi_{\mathcal{C}_s}(\Theta)$ is an operator that returns the rank-s approximation of a matrix $\Theta \in \mathbb{R}^{d_1 \times d_2}$.

Appendix D. Proofs for Section 4

D.1 Proof of Proposition 10

Proof We assume that $\eta_{t,k} = \eta$. We first analyze the MSE $\|\Theta_{t,k} - \Theta^*\|^2$ for $t \leq t_0, 1 \leq k \leq K$. Recall that $\Theta_{t,k} = \Pi_{\mathcal{C}_s}\{\Theta_{t,k-1} - \eta \nabla \mathcal{L}_t(\Theta_{t,k-1})\}$. Let $\mathcal{S}^* = \text{support}(\Theta^*)$, $\mathcal{S}_{t,k} =$

support$(\Theta_{t,k})$, and $g = \nabla \mathcal{L}_t(\Theta_{t,k-1})$, we have

$$\begin{aligned}\|\Theta_{t,k} - \Theta^*\|_2^2 &= \|\Pi_{\mathcal{C}_s}\{\Theta_{t,k-1} - \eta \nabla \mathcal{L}_t(\Theta_{t,k-1})\} - \Theta^*\|_2^2 \\ &= \|(\Theta_{t,k-1} - \eta g - \Theta^*)_{\mathcal{S}_{t,k}} - \Theta^*_{\mathcal{S}^*\setminus \mathcal{S}_{t,k}}\|_2^2 \\ &\le \|\Theta^{t,k-1}_{\mathcal{S}_{t,k}} - \Theta^*_{\mathcal{S}_{t,k}}\|_2^2 + \eta^2 \|g_{\mathcal{S}_{t,k}}\|_2^2 - 2\langle \Theta^{t,k-1}_{\mathcal{S}_{t,k}} - \Theta^*_{\mathcal{S}_{t,k}}, \eta g_{\mathcal{S}_{t,k}}\rangle + \|\Theta^*_{\mathcal{S}^*\setminus \mathcal{S}_{t,k}}\|_2^2 \\ &\le 2\|\Theta^{t,k-1}_{\mathcal{S}_{t,k}} - \Theta^*_{\mathcal{S}_{t,k}}\|_2^2 + 2\eta^2 \|g_{\mathcal{S}_{t,k}}\|_2^2 + \|\Theta^*_{\mathcal{S}^*\setminus \mathcal{S}_{t,k}}\|_2^2.\end{aligned} \tag{36}$$

We write $\widetilde{X}_t = [X_1, X_2, \cdots, X_t]^\top \in \mathbb{R}^{t\times d}$. Taking expectation and using the fact

$$g_{\mathcal{S}_{t,k}} = t^{-1}\Big[\widetilde{X}_t^\top \widetilde{X}_t(\Theta_{t,k-1} - \Theta^*) + \widetilde{X}_t^\top \widetilde{\epsilon}_t\Big]_{\mathcal{S}_{t,k}} = [A(\Theta_{t,k-1} - \Theta^*)]_{\mathcal{S}_{t,k}} + \frac{[\widetilde{X}_t^\top \widetilde{\epsilon}_t]_{\mathcal{S}_{t,k}}}{t},$$

where $A = t^{-1}\widetilde{X}_t^\top \widetilde{X}_t$. We have

$$[A(\Theta_{t,k-1} - \Theta^*)]_{\mathcal{S}_{t,k}} = A_{\mathcal{S}_{t,k}}(\Theta_{t,k-1} - \Theta^*),$$

where $A_{\mathcal{S}_{t,k}}$ agrees with A when the row index belongs to $\mathcal{S}_{t,k}$ with other rows being zeroes. Let $\mathcal{A} = \text{support}\{\Theta^* - \Theta_{t,k-1}\}$ and $\mathcal{D} = \mathcal{A}\cup \mathcal{S}_{t,k}$, we can see that the cardinality of $\mathcal{D}$ is no more than $2s + s*$, and

$$\begin{aligned}\Big\|\big[A(\Theta_{t,k-1} - \Theta^*)\big]_{\mathcal{S}_{t,k}}\Big\|_2^2 &= (\Theta_{t,k-1} - \Theta^*)^\top A^\top_{\mathcal{S}_{t,k}} A_{\mathcal{S}_{t,k}}(\Theta_{t,k-1} - \Theta^*) \\ &= (\Theta_{t,k-1} - \Theta^*)^\top A^\top_{\mathcal{S}_{t,k}\mathcal{A}} A_{\mathcal{S}_{t,k}\mathcal{A}}(\Theta_{t,k-1} - \Theta^*) \\ &= (\Theta_{t,k-1} - \Theta^*)^\top_{\mathcal{D}} A^\top_{\mathcal{D}\mathcal{D}} A_{\mathcal{D}\mathcal{D}}(\Theta_{t,k-1} - \Theta^*)_{\mathcal{D}},\end{aligned}$$

where $A_{\mathcal{S}_{t,k}\mathcal{A}}$ denotes the collection of entries within $\mathcal{A}$ whose row index belong to $\mathcal{S}_{t,k}$ and column index belong to $\mathcal{A}$, and $\mathcal{A}_{\mathcal{D}\mathcal{D}}$ is defined in a similar manner.

Note that when $\mathcal{L}_t(\cdot)$ satisfies the $(2s, s)$-SSS condition with constant L, (27) states that for any $\Theta \in \mathbb{R}^d$ such that $\|\Theta\|_0 \le 2s$,

$$\frac{1}{t}\sum_{i=1}^t \langle X_i, \Theta - \Theta^*\rangle^2 = (\Theta - \Theta^*)^\top A(\Theta - \Theta^*) \le L\|\Theta - \Theta^*\|_2^2,$$

which further implies

$$\|A_{\mathcal{D}\mathcal{D}}\|_2 \le L \text{ and } \|A_{\mathcal{D}\mathcal{D}}\|_2^2 \le L^2.$$

As a result, we obtain that

$$\Big\|\big[A(\Theta_{t,k-1} - \Theta^*)\big]_{\mathcal{S}_{t,k}}\Big\|_2^2 \le L^2\|\Theta_{t,k-1} - \Theta^*\|_2^2.$$

This together with the fact that ϵ_i's are i.i.d. distributed further imply

$$\begin{aligned}\mathbb{E}\left[\|g_{\mathcal{S}_{t,k}}\|_2^2\right] &\le \frac{2}{t^2}\mathbb{E}\Big[\text{tr}\Big((\Theta_{t,k-1} - \Theta^*)^\top (\widetilde{X}_t^\top \widetilde{X}_t)^2(\Theta_{t,k-1} - \Theta^*)\Big)\Big] + \frac{2}{t^2}\mathbb{E}\Big[\Big\|\Big(\sum_{i=1}^t \epsilon_i X_i\Big)_{\mathcal{S}_{t,k}}\Big\|_2^2\Big] \\ &\le \frac{2}{t^2}\mathbb{E}\Big[\text{tr}\Big((\Theta_{t,k-1} - \Theta^*)^\top (\widetilde{X}_t^\top \widetilde{X}_t)^2(\Theta_{t,k-1} - \Theta^*)\Big)\Big] + \frac{2s}{t^2}\mathbb{E}\Big[\Big\|\sum_{i=1}^t \epsilon_i X_i\Big\|_\infty^2\Big] \\ &\le 2L^2\mathbb{E}[\|\Theta_{t,k-1} - \Theta^*\|_2^2] + 2s\sigma_x^2/t.\end{aligned}$$

We then obtain

$$\begin{aligned}\mathbb{E}[\|\Theta_{t,k}-\Theta^*\|_2^2] &\le 2\mathbb{E}[\|\Theta^{t,k-1}_{\mathcal{S}_{t,k}}-\Theta^*_{\mathcal{S}_{t,k}}\|_2^2]+2\eta^2\mathbb{E}[\|g_{\mathcal{S}_{t,k}}\|_2^2]+\mathbb{E}[\|\Theta^*_{\mathcal{S}^*\setminus\mathcal{S}_{t,k}}\|_2^2]\\ &\le 2(1+\eta^2L^2)\mathbb{E}[\|\Theta_{t,k-1}-\Theta^*\|_2^2]+4s\eta^2\sigma_x^2/t+\|\Theta^*\|_2^2\\ &\le 2(1+\eta^2L^2)\mathbb{E}[\|\Theta_{t,k-1}-\Theta^*\|_2^2]+4s\eta^2\sigma_x^2+\|\Theta^*\|_2^2.\end{aligned}$$

Recursively applying the above relationship under the condition that $\eta \le \frac{1}{L}$, we obtain

$$\mathbb{E}[\|\Theta_{t,k}-\Theta^*\|_2^2] \le 4^{(t-1)K+k}\|\Theta_{0,0}-\Theta^*\|_2+\left(4s\eta^2\sigma_x^2+\|\Theta^*\|_2^2\right)\frac{(4^{(t-1)K+k}-1)}{3}. \tag{37}$$

Next, using the descent lemma with $\eta \le 1/L$ acquires

$$\begin{aligned}&\mathcal{L}_t(\Theta_{t,K})-\mathcal{L}_t(\widehat{\Theta}_t) \le \mathcal{L}_t(\Theta_{t,0})-\mathcal{L}_t(\widehat{\Theta}_t)=\mathcal{L}_t(\Theta_{t-1,K})-\mathcal{L}_t(\widehat{\Theta}_t)\\ &\quad=\mathcal{L}_{t-1}(\Theta_{t-1,K})-\mathcal{L}_{t-1}(\widehat{\Theta}_{t-1})+\Big(\mathcal{L}_t(\Theta_{t-1,K})-\mathcal{L}_{t-1}(\Theta_{t-1,K})\Big)+u_t\\ &\quad\le \mathcal{L}_{t-1}(\Theta_{t-1,K})-\mathcal{L}_{t-1}(\widehat{\Theta}_{t-1})+\frac{1}{t}\ell_t(\Theta_{t-1,K})-\frac{1}{t(t-1)}\sum_{i=1}^{t-1}\ell_i(\Theta_{t-1,K})+u_t\end{aligned}$$

where $u_t=\mathcal{L}_{t-1}(\widehat{\Theta}_{t-1})-\mathcal{L}_t(\widehat{\Theta}_t)$. In addition, we have that for $t=1$, $\mathcal{L}_1(\Theta_{1,K})-\mathcal{L}_1(\widehat{\Theta}_1)\le \mathcal{L}_1(\Theta_{1,0})-\mathcal{L}_1(\widehat{\Theta}_1)=\ell_1(\Theta_{1,0})-\mathcal{L}_1(\widehat{\Theta}_1)=\ell_1(\Theta_0)-\mathcal{L}_1(\widehat{\Theta}_1)$. By recursively applying the above relation, we have

$$\mathcal{L}_t(\Theta_{t,K})-\mathcal{L}_t(\widehat{\Theta}_t)\le\sum_{j=1}^{t}\frac{1}{j}\ell_j(\Theta_{j-1,K})+\sum_{j=1}^{t}u_j=\sum_{j=1}^{t}\frac{1}{j}\ell_j(\Theta_{j-1,K})-\mathcal{L}_t(\widehat{\Theta}_t). \tag{38}$$

Because data are independently generated and collected, the t-th feature-label pair (X_t,Y_t) is independent of $\Theta_{t-1,K}$ at the $(t-1)$-th epoch. Using the fact that $\|\Theta_{t-1,K}-\Theta^*\|_0\le s+s^*$, we have

$$\begin{aligned}\mathbb{E}[\ell_t(\Theta_{t-1,K})]&=\mathbb{E}\Big[\Big(\langle X_t,\Theta_{t-1,K}\rangle-Y_t\Big)^2\Big]=\mathbb{E}\Big[\Big(\langle X_t,\Theta_{t-1,K}-\Theta^*\rangle-\epsilon_t\Big)^2\Big]\\ &=\mathbb{E}\Big[\langle X_t,\Theta_{t-1,K}-\Theta^*\rangle^2\Big]+\mathbb{E}[\epsilon_t^2]\\ &\le\mathbb{E}\Big[\|X_t\|_\infty^2\Big]\mathbb{E}[\|\Theta_{t-1,K}-\Theta^*\|_1^2\Big]+\mathbb{E}[\epsilon_t^2]\\ &\le(s+s^*)\mathbb{E}\Big[\|X_t\|_\infty^2\Big]\mathbb{E}[\|\Theta_{t-1,K}-\Theta^*\|_2^2\Big]+\mathbb{E}[\epsilon_t^2]\\ &\le C_0\mathbb{E}[\|\Theta_{t-1,K}-\Theta^*\|_2^2]+\sigma^2\end{aligned}$$

where $C_0=(s+s^*)\mathbb{E}\big[\|X_0\|_\infty^2\big]$. Clearly, rearranging (38) and taking expectations on both sides, we obtain

$$\mathbb{E}[\mathcal{L}_t(\Theta_{t,K})]\le\sum_{j=1}^{t}\frac{1}{j}\mathbb{E}[\ell_j(\Theta_{j-1,K})]\le C_0\sum_{1\le j\le t}\frac{1}{j}\mathbb{E}\big[\|\Theta_{j-1,K}-\Theta^*\|_2^2\big]+\sigma^2\log(t+1).$$

Finally, applying (37) to the above inequality, we conclude that

$$\begin{aligned}\mathbb{E}[\mathcal{L}_t(\Theta_{t,K})] &\le C_0 \sum_{j\le t} \frac{4^{(j-1)K}}{j} \times \left(\|\Theta_{0,0} - \Theta^*\|_2 + \frac{4s\eta^2\sigma_x^2 + \|\Theta^*\|_2^2}{3} \right) + \sigma^2 \log(t+1) \\ &\le C_1 4^{tK} + \sigma^2 \log(t+1)\end{aligned}$$

for all $t \le t_0$, where $C_1 = C_0(\|\Theta_{0,0} - \Theta^*\|_2 + (4s\eta^2\sigma_x^2 + \|\Theta^*\|_2^2)/3)$. This completes the proof. ■

D.2 Proof of Proposition 12

Proof We first analyze the solution sequence $\{\Theta_{t,k} - \Theta^*\}_{t\le t_0}$ up to time t_0. Recall that $\Theta_{t,k} = \Pi_{\mathcal{C}_s}\{\Theta_{t,k-1} - \eta\nabla\mathcal{L}_t(\Theta_{t,k-1})\}$. Let $S_{t,k}$ and $\widetilde{S}_{t,k}$ be the column spaces of $\Theta_{t,k}$ and $\Theta_{t,k} \cup \Theta^*$ respectively. Let $g = \nabla\mathcal{L}_t(\Theta_{t,k-1})$. We have

$$\begin{aligned}&\|\Theta_{t,k} - \Theta^*\|_{\mathrm{F}}^2 \\ &= \|\Pi_{\mathcal{C}_s}\{\Theta_{t,k-1} - \eta\nabla\mathcal{L}_t(\Theta_{t,k-1})\} - \Theta^*\|_{\mathrm{F}}^2 \\ &= \|P_{S_{t,k}}(\Theta_{t,k-1} - \eta g - \Theta^*) - P_{S^*\cap S_{t,k}^\perp}\Theta^*\|_{\mathrm{F}}^2 \\ &\le \|P_{S_{t,k}}(\Theta_{t,k-1} - \Theta^*)\|_{\mathrm{F}}^2 + \eta^2\|P_{S_{t,k}}(g)\|_{\mathrm{F}}^2 \\ &\qquad - 2\langle P_{S_{t,k}}(\Theta_{t,k-1} - \Theta^*), \eta P_{S_{t,k}}(g)\rangle + \|P_{S^*\cap S_{t,k}^\perp}\Theta^*\|_{\mathrm{F}}^2 \\ &\le \|P_{S_{t,k}}(\Theta_{t,k-1} - \Theta^*)\|_{\mathrm{F}}^2 + \eta^2\|P_{S_{t,k}}(g)\|_{\mathrm{F}}^2 \\ &\qquad + 2\eta\,\|P_{S_{t,k}}(\Theta_{t,k-1} - \Theta^*)\|_{\mathrm{F}}\,\|P_{S_{t,k}}(g)\|_{\mathrm{F}} + \|P_{S^*\cap S_{t,k}^\perp}\Theta^*\|_{\mathrm{F}}^2.\end{aligned} \tag{39}$$

For $P_{S_{t,k}}(g)$, we have

$$\begin{aligned}P_{S_{t,k}}(g) &= P_{S_{t,k}}\left[t^{-1}\sum_{i\le t} X_i\,(\langle X_i, \Theta\rangle - Y_i)\right] \\ &= P_{S_{t,k}}\Big[t^{-1}\sum_{i\le t} X_i\,\langle X_i, \Theta_{t,k-1} - \Theta^*\rangle - \epsilon_i)\Big] \\ &= t^{-1}\sum_{i\le t} P_{S_{t,k}}X_i\left\langle P_{S_{t,k}\cup\widetilde{S}_{t,k-1}}X_i, \Theta_{t,k-1} - \Theta^*\right\rangle - t^{-1}\sum_{i\le t} P_{S_{t,k}}X_i\epsilon_i,\end{aligned}$$

implying that

$$\|P_{S_{t,k}}(g)\|_{\mathrm{F}}^2 \le 2t^{-2}\left\|\sum_{i\le t} P_{S_{t,k}}X_i\left\langle P_{S_{t,k}\cup\widetilde{S}_{t,k-1}}X_i, \Theta_{t,k-1} - \Theta^*\right\rangle\right\|_{\mathrm{F}}^2 + 2t^{-2}\left\|\sum_{i\le t} P_{S_{t,k}}X_i\epsilon_i\right\|_{\mathrm{F}}^2.$$

Write $S_{t,k} \cup \widetilde{S}_{t,k-1} = \{S_{t,k}^\perp \cap \widetilde{S}_{t,k-1}\} \cup S_{t,k}$, where $\{S_{t,k}^\perp \cap \widetilde{S}_{t,k-1}\} \cap S_{t,k} = 0$. Then

$$P_{S_{t,k}\cup\widetilde{S}_{t,k-1}}X_i = P_{S_{t,k}}X_i + P_{S_{t,k}^\perp\cap\widetilde{S}_{t,k-1}}X_i \qquad \text{and} \qquad \left\langle P_{S_{t,k}}X_i, P_{S_{t,k}^\perp\cap\widetilde{S}_{t,k-1}}X_i\right\rangle = 0,$$

which further implies

$$\begin{aligned}
&\left\|\sum_{i\le t} P_{S_{t,k}} X_i \left\langle P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} X_i, \Theta_{t,k-1}-\Theta^*\right\rangle\right\|_{\mathrm{F}}^2 \\
&= \left\|\sum_{i\le t} P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} X_i \left\langle P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} X_i, \Theta_{t,k-1}-\Theta^*\right\rangle\right\|_{\mathrm{F}}^2 \\
&\quad - \left\|\sum_{i\le t} P_{S_{t,k}^{\perp}\cap\widetilde{S}_{t,k-1}} X_i \left\langle P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} X_i, \Theta_{t,k-1}-\Theta^*\right\rangle\right\|_{\mathrm{F}}^2 \\
&\le \left\|\sum_{i\le t} P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} X_i \left\langle P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} X_i, \Theta_{t,k-1}-\Theta^*\right\rangle\right\|_{\mathrm{F}}^2 .
\end{aligned}$$

Combining the above inequalities, we obtain

$$\|P_{S_{t,k}}(g)\|_{\mathrm{F}}^2 \le 2t^{-2}\left\|\sum_{i\le t} P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} X_i \left\langle P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} X_i, \Theta_{t,k-1}-\Theta^*\right\rangle\right\|_{\mathrm{F}}^2 + 2t^{-2}\left\|\sum_{i\le t} P_{S_{t,k}} X_i\epsilon_i\right\|_{\mathrm{F}}^2 . \tag{40}$$

Let $z_i = \mathrm{vec}(P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} X_i)$ and $\theta_{t,k-1} = \mathrm{vec}(\Theta_{t,k-1}-\Theta^*)$, then we can bound the first term as

$$\begin{aligned}
&\left\|t^{-1}\sum_{i\le t} P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} X_i \left\langle P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} X_i, \Theta_{t,k-1}-\Theta^*\right\rangle\right\|_{\mathrm{F}}^2 \\
&= \theta_{t,k-1}^{\top}\Big(\frac{1}{t}\sum_{i\le t} z_i z_i^{\top}\Big)\Big(\frac{1}{t}\sum_{i\le t} z_i z_i^{\top}\Big)\theta_{t,k-1} \\
&\le \left\|\frac{1}{t}\sum_{i\le t} z_i z_i^{\mathrm{T}}\right\|_2^2 \|\Theta_{t,k-1}-\Theta^*\|_{\mathrm{F}}^2 .
\end{aligned} \tag{41}$$

To further bound the right hand side (RHS) of the inequality above, we need to bound $\|\sum_{i\le t} z_i z_i^{\mathrm{T}}/t\|_2$. Because $\mathcal{L}_t(\cdot)$ satisfies the $(2s+s^*)$-LowRankSS with parameter L, following the proof of (34), we have

$$\frac{1}{t}\sum_{i\le t}\left\langle P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} X_i, P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} A\right\rangle^2 \le L\|P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} A\|_{\mathrm{F}}^2,$$

for any matrix $A\in\mathbb{R}^{d_1\times d_2}$. Let $a=\mathrm{vec}(A)$. This further implies

$$\begin{aligned}
a^{\mathrm{T}}\left(\frac{1}{t}\sum_{i\le t} z_i z_i^{\mathrm{T}}\right) a &= \frac{1}{t}\sum_{i\le t}\left\langle P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} X_i, P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} A\right\rangle^2 \\
&\le L\|P_{S_{t,k}\cup\widetilde{S}_{t,k-1}} A\|_{\mathrm{F}}^2 \le L\|P_{S_{t,k}\cup\widetilde{S}_{t,k-1}}\|_2^2\|A\|_{\mathrm{F}}^2 \le L\|a\|_2^2,
\end{aligned}$$

and thus

$$\left\| \frac{1}{t} \sum_{i \le t} z_i z_i^{\mathrm{T}} \right\|_2 \le L.$$

Plugging the above inequality into (41), we obtain

$$\left\| t^{-1} \sum_{i \le t} P_{S_{t,k} \cup \widetilde{S}_{t,k-1}} X_i \left\langle P_{S_{t,k} \cup \widetilde{S}_{t,k-1}} X_i, \Theta_{t,k-1} - \Theta^* \right\rangle \right\|_{\mathrm{F}}^2 \le L^2 \|\Theta_{t,k-1} - \Theta^*\|_{\mathrm{F}}^2.$$

Substituting the above inequality into (40) and taking expectations on both sides, we have

$$\begin{aligned}
\mathbb{E}[\|P_{S_{t,k}}(g)\|_{\mathrm{F}}^2] &\le 2L^2 \mathbb{E}\|\Theta_{t,k-1} - \Theta^*\|_{\mathrm{F}}^2 + \frac{2}{t^2} \mathbb{E} \left\| P_{S_{t,k}} \Big(\sum_{i=1}^{t} X_i \epsilon_i \Big) \right\|_{\mathrm{F}}^2 \\
&\le 2L^2 \mathbb{E}\|\Theta_{t,k-1} - \Theta^*\|_{\mathrm{F}}^2 + \frac{2s}{t^2} \mathbb{E} \left\| \sum_{i=1}^{t} X_i \epsilon_i \right\|_2^2 \\
&\le 2L^2 \mathbb{E}\|\Theta_{t,k-1} - \Theta^*\|_{\mathrm{F}}^2 + 2s\sigma_X^2/t,
\end{aligned}$$

where the second inequality uses the facts that $\operatorname{rank}(P_{S_{t,k}}(\sum_{i=1}^{t} X_i \epsilon_i)) \le s$ and $\|A\|_{\mathrm{F}}^2 \le \operatorname{rank}(A)\|A\|_2^2$, and the last inequality comes from Assumption 4. Following the analysis of Proposition 10, we conclude that

$$\mathbb{E}[\mathcal{L}_t(\Theta_t)] \le C_1 4^{tK} + \sigma^2 \log(t+1)$$

for all $t \le t_0$, where $C_1 = C_0(\|\Theta_0 - \Theta^*\|_{\mathrm{F}}^2 + (4s\eta^2\sigma_X^2 + \|\Theta^*\|_{\mathrm{F}}^2)/3)$ and $C_0 = (s+s^*)\mathbb{E}\big[\|X_0\|_2^2\big]$. This completes the proof. ■

References

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40 (5):2452–2482, 2012.

Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163–195, 2011.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

Emmanuel J. Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

Emmanuel J. Candès and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61): 2121–2159, 2011.

Jianqing Fan, Wenyan Gong, Chris Junchi Li, and Qiang Sun. Statistical Sparse Online Regression: A Diffusion Approximation Perspective. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS 2018). Proceedings of Machine Learning Research*, 84:1017–1026, 2018a.

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, 46(2):814–841, 2018b.

Dean Foster, Howard Karloff, and Justin Thaler. Variable Selection Is Hard. In *Proceedings of the 28th Conference on Learning Theory (COLT 2015). Proceedings of Machine Learning Research*, 40:696–709, 2015.

Dongxiao Han, Jinhan Xie, Jin Liu, Liuquan Sun, Jian Huang, Bei Jiang, and Linglong Kong. Inference on high-dimensional single-index models with streaming data. *Journal of Machine Learning Research*, 25(337):1–68, 2024a.

Ruijian Han, Lan Luo, Yuanyuan Lin, and Jian Huang. Online inference with debiased stochastic gradient descent. *Biometrika*, 111(1):93–108, 2024b.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On Iterative Hard Thresholding Methods for High-Dimensional M-Estimation. In *Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS 2014). Advances in Neural Information Processing Systems*, 27:685–693, 2014.

Satyen Kale, Zohar Karnin, Tengyuan Liang, and Dávid Pál. Adaptive Feature Selection: Computationally Efficient Online Sparse Linear Regression under RIP. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017). Proceedings of Machine Learning Research*, 70:1780–1788, 2017.

Harold J. Kushner and G. George Yin. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, 1997.

Guanghui Lan. *First-Order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.

Haoyang Liu and Rina Foygel Barber. Between hard and soft thresholding: Optimal iterative thresholding algorithms. *Information and Inference: A Journal of the IMA*, 9(4): 899–933, 2020.

Po-Ling Loh and Martin J. Wainwright. Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16(19):559–616, 2015.

Marie Maros and Gesualdo Scutari. Decentralized Matrix Sensing: Statistical Guarantees and Fast Convergence. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Advances in Neural Information Processing Systems*, 36:40154–40166, 2023.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*. Omnipress, 1571–1578, 2012.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11(78):2241–2259, 2010.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

Jacob Steinhardt, Stefan Wager, and Percy Liang. The statistics of streaming sparse regression. *arXiv preprint arXiv:1412.4182*, 2014.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

Roman Vershynin. On large random almost Euclidean bases. *Acta Mathematica Universitatis Comenianae*, 69(2):137–144, 2000.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(88):2543–2596, 2010.

Shuoguang Yang, Yuhao Yan, Xiuneng Zhu, and Qiang Sun. Online Linearized LASSO. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023). Proceedings of Machine Learning Research*, 206:7594–7610, 2023.

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7(90):2541–2563, 2006.