

Global Convergence of Sub-gradient Method for Robust Matrix Recovery: Small Initialization, Noisy Measurements, and Over-parameterization

Jianhao Ma

JIANHAO@UMICH.EDU

*Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48105, USA*

Salar Fattahi

FATTAHI@UMICH.EDU

*Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48105, USA*

Editor: Francis Bach

Abstract

In this work, we study the performance of sub-gradient method (SubGM) on a natural nonconvex and nonsmooth formulation of *low-rank matrix recovery* with ℓ_1 -loss, where the goal is to recover a low-rank matrix from a limited number of measurements, a subset of which may be grossly corrupted with noise. We study a scenario where the rank of the true solution is unknown and over-estimated instead. The over-estimation of the rank gives rise to an over-parameterized model in which there are more degrees of freedom than needed. Such over-parameterization may lead to overfitting, or adversely affect the performance of the algorithm. We prove that a simple SubGM with small initialization is *agnostic* to both over-parameterization and noise in the measurements. In particular, we show that small initialization nullifies the effect of over-parameterization on the performance of SubGM, leading to an exponential improvement in its convergence rate. Moreover, we provide the first unifying framework for analyzing the behavior of SubGM under both outlier and Gaussian noise models, showing that SubGM converges to the true solution, even under arbitrarily large and arbitrarily dense noise values, and, perhaps surprisingly, even if the globally optimal solutions *do not* correspond to the ground truth. At the core of our results is a robust variant of restricted isometry property, called Sign-RIP, which controls the deviation of the sub-differential of the ℓ_1 -loss from that of an ideal, expected loss. As a byproduct of our results, we consider a subclass of robust low-rank matrix recovery with Gaussian measurements, and show that the number of required samples to guarantee the global convergence of SubGM is *independent* of the over-parameterized rank.¹

Keywords: low-rank matrix recovery, nonconvex optimization, nonsmooth optimization, sub-gradient method, over-parameterization

1. A preliminary version of this paper has appeared in NeurIPS Workshop on Optimization for Machine Learning (Ma and Fattahi, 2021a) (without archival proceedings). The current journal submission is a significant extension of this work on multiple fronts. We extend the definition of Sign-RIP to approximately low-rank matrices. Using this extension, we prove a series of new results on the effect of over-parameterization and noisy measurements on the performance of SubGM.

1. Introduction

We study the problem of *robust matrix recovery*, where the goal is to recover a low-rank positive semidefinite matrix $X^* \in \mathbb{R}^{d \times d}$ from a limited number of linear measurements of the form $\mathbf{y} = \mathcal{A}(X^*) + \mathbf{s}$, where $\mathbf{y} = [y_1, y_2, \dots, y_m]^\top$ is a vector of measurements, \mathcal{A} is a linear operator defined as $\mathcal{A}(\cdot) = [\langle A_1, \cdot \rangle, \langle A_2, \cdot \rangle, \dots, \langle A_m, \cdot \rangle]^\top$ with measurement matrices $\{A_i\}_{i=1}^m$, and $\mathbf{s} = [s_1, s_2, \dots, s_m]^\top$ is a noise vector. More formally, the robust matrix recovery is defined as

$$\text{find } X^* \quad \text{subject to: } \mathbf{y} = \mathcal{A}(X^*) + \mathbf{s}, \quad \text{rank}(X^*) = r, \quad (1)$$

where $r \leq d$ is the rank of X^* . Robust matrix recovery plays a central role in many contemporary machine learning problems, including motion detection in video frames (Bouwmans and Zahzah, 2014), face recognition (Luan et al., 2014), and collaborative filtering in recommender systems (Luo et al., 2014). Despite its widespread applications, it is well-known that solving (1) is a daunting task since it amounts to an NP-hard problem in its worst case (Natarajan, 1995; Recht et al., 2010). What makes this problem particularly difficult is the nonconvexity stemming from the rank constraint. The classical methods for solving low-rank matrix recovery problem are based on convexification techniques, which suffer from notoriously high computational cost. To alleviate this issue, a far more practical approach is to resort to the following natural *nonconvex* and *nonsmooth* formulation

$$\min_{U \in \mathbb{R}^{d \times r'}} f_{\ell_1}(U) := \frac{1}{m} \left\| \mathbf{y} - \mathcal{A}(UU^\top) \right\|_1, \quad (2)$$

where $r' \geq r$ is the search rank. The ℓ_1 -loss is used to robustify the solution against noisy measurements. The above formulation is inspired by the celebrated Burer-Monteiro approach (Burer and Monteiro, 2003), which circumvents the explicit rank constraint by optimizing directly over the factorized model $X = UU^\top$.

Perhaps the most significant breakthrough result in this line of research was presented by Bhojanapalli et al. (2016), showing that, when the rank of the true solution is known and the measurements are noiseless, the nonconvex formulation of the problem with a smooth ℓ_2 -loss has a *benign landscape*, i.e., it is devoid of undesirable local solutions; as a result, simple local-search algorithms are guaranteed to converge to the globally optimal solution. Such benign landscape seems to be omnipresent in other variants of low-rank matrix recovery, including matrix completion (Ge et al., 2017, 2016), robust PCA (Ge et al., 2017; Fattahi and Sojoudi, 2020), sparse dictionary learning (Sun et al., 2016; Qu et al., 2019), linear neural networks (Kawaguchi, 2016), among others; see recent survey papers (Chi et al., 2019; Zhang et al., 2020).

A recurring assumption for the absence of spurious local minima is the exact parameterization of the rank: it is often presumed that the *exact* rank of the true solution is known *a priori*. However, the rank of the true solution is rarely known in many applications. Therefore, it is reasonable to choose the rank of UU^\top conservatively as $r' > r$, leading to an *over-parameterized* model. This challenge is further compounded in the noisy regime, where the injected noise in the measurements can be “absorbed” as a part of the solution, due to the additional degrees of freedom in the model. Evidently, the existing proof techniques face major breakdowns in this setting, *as the problem may no longer enjoy a benign*

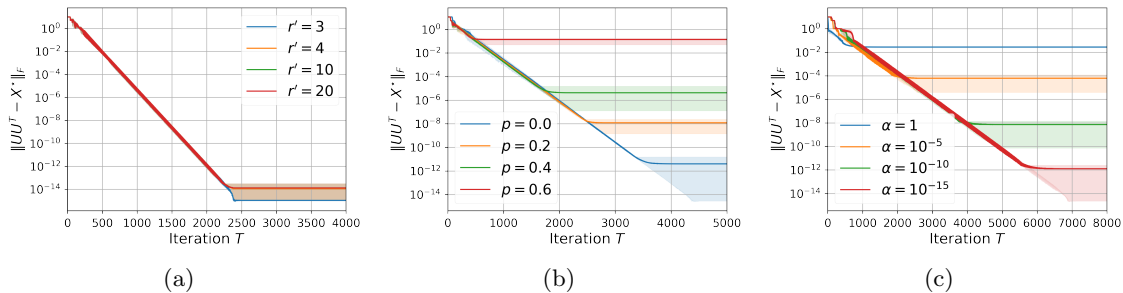


Figure 1: (a) The performance of SubGM for different search ranks r' . (b) The performance of SubGM for different corruption probabilities. (c) The performance of SubGM for different values of the initialization scale α . In all of the simulations, the initial point U_0 is chosen as αB , where B is obtained from Algorithm 2.

landscape. Moreover, over-parameterization may lead to a dramatic, exponential slow-down of the local-search algorithms—both theoretically and practically (Zhuo et al., 2021; Zhang et al., 2021, 2022).

In this work, we study the performance of a simple sub-gradient method (SubGM) on $f_{\ell_1}(U)$. We prove that small initialization nullifies the effect of over-parameterization on its performance—as if the search rank r' were set to the true (but unknown) rank r . Moreover, we show that SubGM converges to the ground truth at a near-linear rate even if local, or even global, spurious minima exist. Our proposed overarching framework is based on a novel signal-residual decomposition of the solution trajectory: we decompose the iterations of SubGM into *low-rank (signal)* and *residual* terms, and show that small initialization keeps the residual term small throughout the solution trajectory, while enabling the low-rank term to converge to the ground truth exponentially fast.

1.1 Power of Small Initialization

In this section, we shed light on the power of small initialization on the performance of SubGM for robust matrix recovery. Given an initial point U_0 and at every iteration t , SubGM selects an arbitrary direction D_t from the (Clarke) sub-differential (Clarke, 1990) of the ℓ_1 -loss function $\partial f_{\ell_1}(U_t)$. Due to local Lipschitzness of the ℓ_1 -loss, the Clarke sub-differential exists and can be obtained via chain rule (see (Clarke, 1990)):

$$\partial f_{\ell_1}(U_t) = \frac{1}{m} \sum_{i=1}^m \text{Sign} \left(\langle A_i, U_t U_t^\top - X^* \rangle \right) \left(A_i + A_i^\top \right) U_t. \quad (3)$$

At every iteration, SubGM updates the solution by moving towards $-D_t$, where $D_t \in \partial f_{\ell_1}(U_t)$, with a step-size η_t . To showcase the effect of small initialization on the performance of SubGM, we consider an instance of robust matrix recovery, where the true solution X^* is a randomly generated matrix with rank $r = 3$ and dimension $d = 20$. Furthermore, we consider $m = 500$ measurements, where the measurement matrices $\{A_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries.

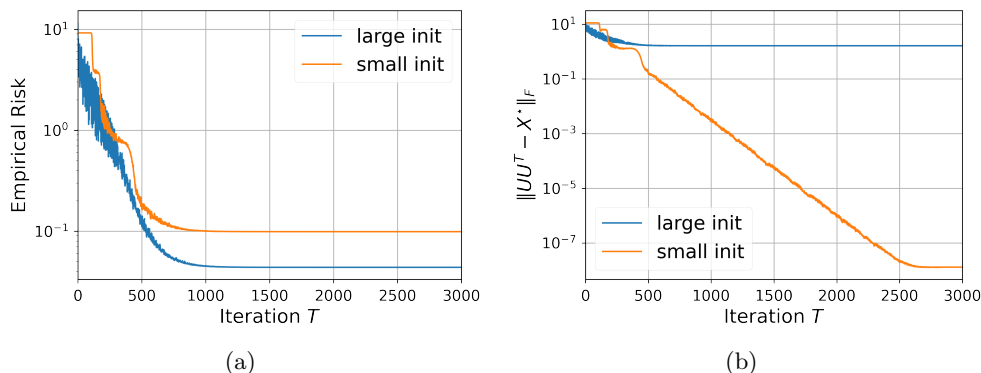


Figure 2: (a) The objective value of the solutions obtained via SubGM with and without small initialization. (b) The error of the solutions obtained via SubGM with and without small initialization. In both instances, the initial point is chosen as $U_0 = \alpha B$, where B is obtained from Algorithm 2. The initialization scale α is chosen as $\alpha = 10^{-15}$ and $\alpha = 1$, for SubGM with and without small initialization, respectively.

Property 1: Small initialization makes SubGM agnostic to over-parameterization.

Figure 1a shows the performance of SubGM with small initialization for both exact ($r' = 3$) and over-parameterized ($r' > 3$) settings, where 10% of the measurements are grossly corrupted with noise. Our simulations uncover an intriguing property of small initialization: neither the convergence rate nor the final error of SubGM is affected by the over-estimation of the rank. Moreover, Figure 1b depicts the performance of SubGM for the fully over-parameterized problem (i.e., $r' = d = 20$) with different levels of corruption probability (the fraction of measurements that are corrupted with noise). The added noise values are i.i.d. and drawn from a Gaussian distribution with mean and variance equal to 0 and 100, respectively. It can be seen that, even in the fully over-parameterized setting, SubGM is robust against large corruption probabilities.

Property 2: Small initialization improves convergence. It is known that different variants of the (sub-)gradient method converge linearly to the true solution, provided that the search rank coincides with the true rank ($r' = r$) (Tu et al., 2016; Zheng and Lafferty, 2015; Tong et al., 2021; Li et al., 2020a). However, these methods suffer from a dramatic, exponential slow-down in over-parameterized models with noisy measurements (Zhuo et al., 2021). Our simulations reveal that small initialization can restore the convergence back to linear, even in the over-parameterized and noisy settings. Figure 1c shows that SubGM converges linearly to an error that is proportional to the norm of the initial point: smaller initial points lead to more accurate solutions at the expense of a slightly larger number of iterations.

Property 3: Emergence of “spurious” global minima. Inspired by these simulations, a natural approach to explaining the desirable performance of SubGM is by showing that the robust matrix recovery problem enjoys a benign landscape. We refute this conjecture by showing that, not only does the robust matrix recovery with over-parameterized rank have sub-optimal solutions, but also its globally optimal solutions may be “spurious”, i.e., they do not correspond to the ground truth X^* . Figure 2 shows the performance of SubGM

with and without small initialization. It can be seen that SubGM converges to the ground truth which is a local solution for the ℓ_1 -loss with sub-optimal objective value. On the other hand, SubGM without small initialization converges to a high-rank solution with a strictly smaller objective value. In other words, the ground truth is not necessarily a globally optimal solution, and conversely, globally optimal solutions do not necessarily correspond to the ground truth.

From a statistical perspective, our simulations support the common empirical observation that first-order methods “generalize well”. In particular, SubGM converges to a low-rank solution that is close to the ground truth—i.e., has a better *generalization error*—rather than recovering a high-rank solution with a smaller objective value (or better *training error*). The smaller objective values for higher rank solutions are precisely due to the overfitting phenomenon: it is possible that the globally optimal solution to (2) achieves a zero objective value by absorbing the noise into its redundant ranks. To circumvent the issue of overfitting, a common approach is to regularize the high-rank solutions in favor of the low-rank ones via different regularization techniques. Therefore, the desirable performance of SubGM with small initialization can be attributed to its *implicit regularization* property. In particular, we show that small initialization of SubGM is akin to implicitly regularizing the redundant rank of the over-parameterized model, thereby avoiding overfitting; a recent work (Stöger and Soltanolkotabi, 2021) has shown a similar property for the gradient descent algorithm on the noiseless matrix recovery with ℓ_2 -loss.

1.2 Summary of Results

In this part, we present a summary of our results. Let σ_1 and σ_r be the largest and smallest (nonzero) eigenvalues of X^* , and define the condition number κ as σ_1/σ_r .

Theorem 1 (Convergence of SubGM; Informal) *Suppose that the measurements satisfy a direction-preserving property delineated in Section 3.2. Suppose that the initial point is chosen as $U_0 = \alpha B$, for a special choice of B and an initialization scale α . Consider the iterations $\{U_t\}_{t=0}^T$ generated by SubGM applied to the robust matrix recovery with step-size $\eta_t = \eta\rho^t$, for an appropriate choice of $0 < \rho < 1$ and sufficiently small η . Then, for any arbitrary accuracy $\varepsilon > 0$ and initialization scale $\alpha = \mathcal{O}((\varepsilon/d)^{1/\beta})$, we have*

$$\left\| U_T U_T^\top - X^* \right\|_F \leq \varepsilon \quad (4)$$

after $T = \mathcal{O}\left(\frac{\kappa \log^2(d/\varepsilon)}{\beta\eta}\right)$ iterations, where $0 < \beta \leq 2$ is a constant depending on the parameters of the problem.

The above result characterizes the performance of SubGM for the robust matrix recovery with ℓ_1 -loss. In particular, it shows that SubGM converges almost *linearly* to the true low-rank solution X^* , with a final error that is proportional to the initialization scale. Surprisingly, the required number of iterations is independent of the search rank r' and depends only logarithmically on d .

At the crux of our analysis lies a new restricted isometry property of the sub-differentials, which we call Sign-RIP. Under Sign-RIP, the sub-differentials of the ℓ_1 -loss are δ -away from the sub-differentials of an ideal, expected loss function (see Section 3.2 for precise

definitions). We will show that the classical notions of ℓ_2 -RIP (Recht et al., 2010) and ℓ_1/ℓ_2 -RIP (Li et al., 2020a) face major breakdowns in the presence of noise. In contrast, Sign-RIP provides much better robustness against noisy measurements, while being no more restrictive than its classical counterparts. We will show that, with Gaussian measurements, the Sign-RIP holds with an overwhelming probability under two popular noise models, namely *outlier noise model* and *Gaussian noise model*.

Our next theorem establishes the convergence of SubGM under the outlier noise model. To streamline the presentation, we use $\tilde{\mathcal{O}}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to hide the dependency on logarithmic factors.

Theorem 2 (Convergence of SubGM under Outlier Noise Model; Informal)

Suppose that the measurement matrices $\{A_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries, and a fraction $p < 1$ of the measurements are corrupted with arbitrarily large noise values. Suppose that the initial point is chosen as $U_0 = \alpha B$, for a special choice of B and a sufficiently small initialization scale α . Consider the iterations $\{U_t\}_{t=0}^T$ generated by SubGM applied to the robust matrix recovery with an exponentially decaying step-size $\eta_t = \eta\rho^t$, for an appropriate choice of $0 < \rho < 1$ and sufficiently small η . Finally, suppose that the number of measurements satisfies $m = \tilde{\Omega}(\kappa^4 dr^2/(1-p)^2)$. Then, for any arbitrary accuracy $\varepsilon > 0$ and initialization scale $\alpha = \varepsilon/d$, and with an overwhelming probability, we have

$$\left\| U_T U_T^\top - X^\star \right\|_F \leq \varepsilon, \quad (5)$$

after $T = \mathcal{O}\left(\frac{\kappa \log^2(d/\varepsilon)}{\eta}\right)$ iterations.

We compare Theorem 2 with a recent result by Ding et al. (2021) on the convergence of SubGM for robust matrix recovery. Theorem 2 shows that small initialization enables SubGM to converge almost linearly, which is exponentially faster than the sublinear rate $\tilde{\mathcal{O}}(1/\varepsilon)$ introduced by Ding et al. (2021). Second, Ding et al. (2021) show that SubGM requires $\tilde{\Omega}(\kappa^{12} dr'^3)$ samples to converge, which depends on the search rank r' . In the over-parameterized regime, where the true rank is small (i.e., $r = \mathcal{O}(1)$) and the search rank is large (i.e., $r' = \Omega(d)$), our result leads to *three orders of magnitude* improvement in the required number of samples (modulo the dependency on κ). Moreover, Ding et al. (2021) crucially rely on the equivalence between globally optimal solutions and the ground truth, which only holds when $p \leq 1/\sqrt{r'}$. We relax this assumption and show that SubGM converges to the ground truth, even if p is arbitrarily close to 1.

Next, we turn our attention to the Gaussian noise model, and show that SubGM converges even if the measurements are corrupted with dense Gaussian noises.

Theorem 3 (Convergence of SubGM under Gaussian Noise Model; Informal)

Suppose that the measurement matrices $\{A_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries, and each measurement is corrupted with a zero-mean Gaussian noise with a variance of at most ν^2 . Suppose that the initial point is chosen as $U_0 = \alpha B$, for a special choice of B and a sufficiently small initialization scale α . Consider the iterations $\{U_t\}_{t=0}^T$ generated by SubGM applied to the robust matrix recovery with exponentially decaying step-sizes $\eta_t = \eta\rho^t$, for an appropriate choice of $0 < \rho < 1$ and sufficiently small η . Finally, suppose

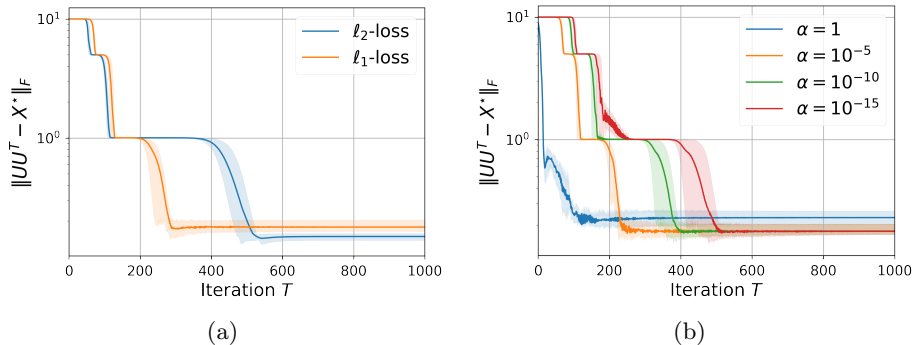


Figure 3: (a) Performance of SubGM and GD under the Gaussian noise model with ℓ_1 - and ℓ_2 -loss functions, respectively. (b) The effect of small initialization on the performance of SubGM under the Gaussian noise model.

that the number of measurements satisfies $m \gtrsim \tilde{\Omega}(\nu^2 \kappa^4 dr^2)$. Then, with an overwhelming probability, we have

$$\|U_t U_t^\top - X^*\|_F = \tilde{\mathcal{O}}\left(\sqrt{\frac{\nu^2 dr^2}{m}}\right), \quad (6)$$

after $T = \mathcal{O}\left(\frac{\kappa}{\eta} \log^2\left(\frac{m\kappa}{\nu r}\right)\right)$ iterations.

Traditionally, ℓ_2 -loss has been used for recovering the ground truth under the Gaussian noise model due to its correspondence to the so-called maximum likelihood estimation. Our paper extends the application of ℓ_1 -loss to this setting, proving that SubGM is robust against not only the outlier, but also Gaussian noise levels. More precisely, Theorem 3 shows that SubGM outputs a solution with an estimation error of $\tilde{\mathcal{O}}(\sqrt{\nu^2 dr^2/m})$, which is again independent of the search rank r' . To the best of our knowledge, the sharpest known estimation error for gradient descent (GD) (Zhuo et al., 2021) and its variants (Zhang et al., 2021) on ℓ_2 -loss is $\mathcal{O}(\sqrt{\nu^2 dr'/m})$, which scales with the search rank r' ; in the fully over-parameterized regime, our provided bound improves upon this error by a factor of $\mathcal{O}(\sqrt{d/r})$. Figure 3 compares the performance of SubGM and GD on ℓ_1 - and ℓ_2 -losses, when the measurements are corrupted with Gaussian noise. Candes and Plan (2011) showed that *any* estimate \hat{U} suffers from a minimax error of $\|\hat{U}\hat{U}^\top - X^*\|_F = \Omega(\sqrt{\nu^2 dr/m})$. Compared to this information-theoretic lower bound, our provided final error is sub-optimal only by a factor of \sqrt{r} .

Organization. To ensure the readability of the paper, we present the main theorems and the key ideas behind their proofs in the main body of the paper, and defer their technical proofs to the appendix. Appendix A presents the detailed proofs of our main results on SubGM; an interested reader would fully understand the proofs of our main theorems after reading the main body and Appendix A. Appendix B presents the proof of Sign-RIP, and Appendices C-F contain the proofs of the intermediate results used throughout the proofs of the main theorems.

2. Related Work

Landscape v.s. Trajectory Analysis: It has been recently shown that different variants of low-rank matrix recovery (e.g., matrix completion (Ge et al., 2016), matrix recovery (Ge et al., 2017), robust PCA (Fattahi and Sojoudi, 2020)) enjoy benign landscape. In particular, it is shown that low-rank matrix recovery with ℓ_2 -loss and noiseless measurements has a benign landscape in both exact (Ge et al., 2017, 2016; Zhang et al., 2019) and over-parameterized (Zhang, 2021) settings. On the other hand, it is known that ℓ_1 -loss possesses better robustness against outlier noise. However, there are far fewer results characterizing the landscape of low-rank matrix recovery with ℓ_1 -loss. Fattahi and Sojoudi (2020) and Josz et al. (2018) prove that robust matrix recovery with ℓ_1 -loss has no spurious local solution, provided that with $r' = r = 1$ and the measurement matrices correspond to element-wise projections. This result was later extended to Gaussian matrices by Ma and Fattahi (2021b). However, it is unclear whether these results extend to higher ranks or more general measurement matrices.

Despite its theoretical significance, benign landscape is too restrictive to hold in practice: Zhang et al. (2019) and Zhang (2021) show that spurious local minima are ubiquitous in the low-rank matrix recovery, even under fairly mild conditions. On the other hand, our experiments in Subsection 1.1 reveal that local-search algorithms may be able to avoid spurious local/global solutions with proper initialization. An alternative approach to explain the desirable performance of local-search algorithms is via *trajectory analysis*. It has been recently shown that the trajectories picked up by gradient-based algorithms benefit from implicit regularization (Gunasekar et al., 2018; Ma and Fattahi, 2022), and incremental learning (Ma et al., 2022), or behave non-monotonically over short timescales, yet consistently improve over long timescales (Cohen et al., 2021). In the context of over-parameterized low-rank matrix recovery with ℓ_2 -loss, Li et al. (2018) and Stöger and Soltanolkotabi (2021) use trajectory analysis to show that GD with small initialization can recover the ground truth, provided that the measurements are noiseless. Zhuo et al. (2021) extend this result to the noisy setting, showing that GD converges to a minimax optimal solution at a sublinear rate, and with a number of samples that scale with the search rank.

Iteration and Sample Complexity: Despite their guaranteed convergence, local-search algorithms may suffer from notoriously slow convergence rates: whereas 10 digits of accuracy can be expected in just a few hundred iterations of GD when $r' = r$, tens of thousands of iterations might produce just 1-2 accurate digits once $r' > r$ (Zhang et al., 2021). Table 1 shows the iteration complexity of the existing algorithms with different loss functions, compared to our proposed method. Evidently, under the outlier noise model, GD does not perform well due to the sensitivity of the ℓ_2 -loss to outliers. On the other hand, SubGM converges linearly in the exact setting ($r' = r$), and at a significantly slower (sublinear) rate in the over-parameterized regime ($r' > r$). In contrast, our proposed SubGM algorithm with small initialization converges near-linearly in *both* exact and over-parameterized regimes. In the Gaussian noise model, it is known that GD converges linearly to a minimax optimal solution in the exact setting, but suffers from a drastic, exponential slow-down in the over-parameterized regime. In contrast, our proposed SubGM algorithm with small initialization is not affected by the over-parameterization, and maintains its desirable convergence rate in both settings.

Algorithm	Outlier noise model		Gaussian noise model	
	$r' = r$	$r' \geq r$	$r' = r$	$r' \geq r$
GD+ ℓ_2 -loss	N/A	N/A	$\kappa \log\left(\frac{m}{\nu dr}\right)$ [6]	$\frac{\sigma_r}{\nu} \sqrt{\frac{m}{d}}$ [45]
SubGM+ ℓ_1 -loss	$\frac{r\kappa}{\sigma_r} \log\left(\frac{\sigma_r}{\varepsilon}\right)$ [20]	$\frac{\sigma_1 \kappa}{\varepsilon}$ [10]	N/A	N/A
Our results	$\kappa^2 \log^3\left(\frac{d}{\varepsilon}\right)$ (see Corollary 28)		$\kappa^2 \log^3\left(\frac{\kappa m}{\nu r}\right)$ (see Corollary 29)	

Table 1: A comparison between the **iteration complexity** of different techniques. We show that SubGM with small initialization and exponentially decaying step-size converges near-linearly to: (i) an arbitrary accuracy in the outlier noise model, and (ii) a nearly-minimax optimal error in the Gaussian noise model. Our derived iteration complexities are obtained from Theorems 2 and 3 after choosing an appropriate value for the step-size; see Corollaries 28 and 29 for the precise statements.

Algorithm	Outlier noise model		Gaussian noise model	
	$r' = r$	$r' \geq r$	$r' = r$	$r' \geq r$
GD+ ℓ_2 -loss	N/A	N/A	$\nu^2 \kappa^2 dr$ [6]	$\nu^2 \kappa^2 dr'$ [45]
SubGM+ ℓ_1 -loss	$\kappa^2 dr^2$ [20]*	$\kappa^{12} dr'^3$ [10]*	N/A	N/A
Our results	$\frac{\kappa^4 dr^2}{(1-p)^2}$ (see Corollary 28)		$\nu^2 \kappa^4 dr^2$ (see Corollary 29)	

Table 2: A comparison between the **sample complexity** of different techniques. Our results provide the best sample complexity bounds in the over-parameterized setting where $r' \gg r$, under both outlier and Gaussian noise models. For simplicity, we hide the dependency on the logarithmic factors. *Under an outlier noise model, the results of (Li et al., 2020a; Ding et al., 2021) holds under the assumption $p \lesssim 1/\sqrt{r'}$. In contrast, our result relaxes this assumption to $p < 1$.

Another important aspect of local-search algorithms is their sample complexity. Table 2 provides a comparison between the sample complexity of the existing algorithms, and our proposed method. In the outlier noise model, Li et al. (2019) show that SubGM with spectral initialization on ℓ_1 -loss requires $\mathcal{O}(dr^2)$ samples (modulo the condition number), provided that the true rank is known ($r' = r$), and the corruption probability is upper bounded as $p \lesssim 1/\sqrt{r'}$. Ding et al. (2021) extend this result to the over-parameterized regime, showing that SubGM with spectral initialization requires $\mathcal{O}(dr'^3)$ samples to converge, under the same assumption $p \lesssim 1/\sqrt{r'}$. In both works, the upper bound $p \lesssim 1/\sqrt{r'}$ is imposed to guarantee that the global minima of the ℓ_1 -loss correspond to the true solution. Our result relaxes this upper bound on the corruption probability by showing that SubGM converges to the ground truth, even if the ground truth is not globally optimal. In the Gaussian noise model, Zhuo et al. (2021) shows that GD recovers the true solution with $\mathcal{O}(dr')$ samples. In the over-parameterized regime, our result reduces the sample complexity to $\mathcal{O}(dr^2)$, showing that the sample complexity of SubGM is independent of the search rank r' .

Notations

For a rank- r matrix $M \in \mathbb{R}^{m \times n}$, its singular values are denoted as $\|M\| = \sigma_1(X) \geq \sigma_2(X) \geq \dots \geq \sigma_r(X) := \sigma_{\min}(X)$. For a square matrix $X \in \mathbb{R}^{n \times n}$, its eigenvalues are defined as $\lambda_1(X) \geq \lambda_2(X) \geq \dots \geq \lambda_n(X) := \lambda_{\min}(X)$. For two matrices X and Y of the same size, their inner product is defined as $\langle X, Y \rangle = \text{Tr}(X^\top Y)$, where $\text{Tr}(\cdot)$ is the trace operator. For a matrix X , its operator and Frobenius norms are denoted as $\|X\|$ and $\|X\|_F$, respectively. The unit rank- r sphere is defined as $\mathbb{S}_r^d = \{X \in \mathbb{R}^{d \times d} : \|X\|_F = 1, \text{rank}(X) \leq r\}$. We

also omit the superscript d when there is no ambiguity. We define P_V as the projection operator onto the row space of the matrix V . The notation $\mathbb{B}(X, \varepsilon)$ refers to a ball of radius ε centered at X , i.e., $\mathbb{B}(X, \varepsilon) = \{X' : \|X' - X\|_F \leq \varepsilon\}$. The ℓ_q norm of a vector x is defined as $\|x\|_q = (\sum |x_i|^q)^{1/q}$. Given two sequences $f(n)$ and $g(n)$, the notation $f(n) \lesssim g(n)$ implies that there exists a constant $C < \infty$ satisfying $f(n) \leq Cg(n)$. Moreover, the notation $f(n) \asymp g(n)$ implies that $f(n) \lesssim g(n)$ and $g(n) \lesssim f(n)$. Throughout the paper, the symbols C, c_1, c_2, \dots refer to universal constants whose precise value may change according to the context. The sign function $\text{Sign}(\cdot)$ is defined as $\text{Sign}(x) = x/|x|$ if $x \neq 0$, and $\text{Sign}(0) = [-1, 1]$. For two sets \mathcal{X} and \mathcal{Y} , the notation $\mathcal{X} + \mathcal{Y}$ refers to their Minkowski sum. Given two scalars a and b , the symbols $a \wedge b$ and $a \vee b$ are used to denote their minimum and maximum, respectively.

3. Our Overarching Framework

In this section, we present our overarching framework for the analysis of SubGM. To this goal, we first explain why the existing techniques for studying the smooth variants of the low-rank matrix recovery cannot be extended to their robust counterparts.

3.1 Failure of Existing Techniques

The majority of existing methods study the behavior of the gradient descent on ℓ_2 -loss $f_{\ell_2}(U) = \frac{1}{m} \|\mathbf{y} - \mathcal{A}(UU^\top)\|^2$ by analyzing its deviation from an “ideal”, noiseless loss function $\bar{f}_{\ell_2}(U) = \|UU^\top - X^*\|_F^2$. It is known that $\bar{f}_{\ell_2}(U) = \|UU^\top - X^*\|_F^2$ is devoid of spurious local minima, and its saddle points are strict, and hence, escapable (see (Zhang et al., 2020, Appendix A) for a simple proof). Therefore, by controlling the deviation of $f_{\ell_2}(U)$ and its gradients from $\bar{f}_{\ell_2}(U)$, one can show that $f_{\ell_2}(U)$ inherits the desirable properties of $\bar{f}_{\ell_2}(U)$. More concretely, the gradient of $f_{\ell_2}(U)$ can be written as $\nabla f_{\ell_2}(U) = Q(UU^\top - X^*)U$, where $Q(X) = (2/m) \sum_{i=1}^m (\langle A_i, X \rangle - s_i) (A_i + A_i^\top)$. One sufficient condition for $\nabla f_{\ell_2}(U) \approx \nabla \bar{f}_{\ell_2}(U)$ is to ensure that $Q(M)$ remains uniformly close to M for every rank- $(r + r')$ matrix X . In the noiseless setting, this condition can be guaranteed via ℓ_2 -RIP:

Definition 4 (ℓ_2 -RIP, Recht et al. (2010)) *The linear operator $\mathcal{A}(\cdot)$ satisfies ℓ_2 -RIP with parameters (k, δ) if, for every rank- k matrix M , we have $(1 - \delta)\|M\|_F^2 \leq \frac{1}{m}\|\mathcal{A}(M)\|^2 \leq (1 + \delta)\|M\|_F^2$.*

Roughly speaking, ℓ_2 -RIP entails that the linear operator $\mathcal{A}(\cdot)$ is nearly “norm-preserving” for every rank- k matrix. In the noiseless setting, this implies that $Q(UU^\top - X^*) \approx 4(UU^\top - X^*)$, which in turn leads to $\nabla f_{\ell_2}(U) \approx \nabla \bar{f}_{\ell_2}(U)$. On the other hand, it is known that ℓ_2 -RIP is satisfied under mild conditions. For instance, (k, δ) - ℓ_2 -RIP holds with

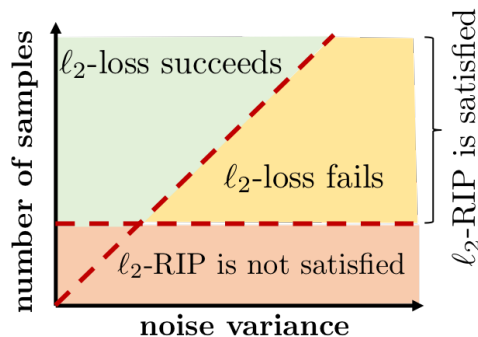


Figure 4: The number of samples to satisfy ℓ_2 -RIP is independent of the noise variance. However, the performance of ℓ_2 highly depends on the noise variance.

$m = \Omega(dk/\delta^2)$ Gaussian measurements (Recht et al., 2010). However, the next proposition shows that ℓ_2 -RIP is not enough to guarantee $Q(M) \approx 4M$ when the measurements are subject to noise.

Proposition 5 (Ma and Fattahi (2021a)) *Suppose that $r' = d$ and the measurement matrices $\{A_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries. Moreover, suppose that the noise vector \mathbf{s} satisfies $s_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ with probability p , and $s_i = 0$ with probability $1 - p$, for every $i = 1, \dots, m$. Then we have*

$$\mathbb{P} \left(\sup_{M \in \mathcal{S}} \|Q(M) - 4M\|_F \gtrsim \sqrt{\frac{(1 + p\sigma^2)d^2}{m}} \right) \geq \frac{1}{2}.$$

Proposition 5 sheds light on a fundamental shortcoming of ℓ_2 -RIP: in the presence of noise, it is possible for the measurements to satisfy ℓ_2 -RIP, yet $\nabla f_{\ell_2}(U)$ may be far from $\nabla \bar{f}_{\ell_2}(U)$. In particular, we show that, in order to have $\nabla f_{\ell_2}(U) \approx \nabla \bar{f}_{\ell_2}(U)$, the number of measurements must grow with the noise variance. On the other hand, for any fixed δ , ℓ_2 -RIP is guaranteed to be satisfied with a number of measurements that is *independent* of the noise variance. Figure 4 shows that ℓ_2 -RIP cannot capture the behavior of ℓ_2 -loss in the high noise regime. Other notions of RIP, such as ℓ_1/ℓ_2 -RIP (Li et al., 2020a), are also oblivious to the nature of the noise.

Algorithm 1 Subgradient Method

Input: measurement matrices $\{A_i\}_{i=1}^m$, measurement vector $\mathbf{y} = [y_1, \dots, y_m]^\top$, number of iterations T , the initial point U_0 ;

Output: Solution $\hat{X}_T = U_T U_T^\top$ to (2);

for $t \leq T$ **do**

Compute a sub-gradient $D_t \in \partial f_{\ell_1}(U_t)$;

Select the step-size η_t (see (10));

Set $U_{t+1} \leftarrow U_t - \eta_t D_t$;

end for

3.2 Sign-RIP: A New Robust Restricted Isometry Property

To address the aforementioned challenges, we argue that, while the measurements may not be norm-preserving in the presence of noise, they may still enjoy a “direction-preserving” property. At the heart of our analysis lies the following decomposition of the sub-differential of the ℓ_1 -loss:

$$\partial f_{\ell_1}(U) = \underbrace{\gamma \cdot \partial \bar{f}_{\ell_1}(U)}_{\text{expected sub-differential}} + \underbrace{(\partial f_{\ell_1}(U) - \gamma \cdot \partial \bar{f}_{\ell_1}(U))}_{\text{sub-differential deviation}},$$

where γ is a strictly positive number. In the above decomposition, the function $\bar{f}_{\ell_1}(U)$ is called the *expected loss*, and it is defined as $\|UU^\top - X^*\|_F$. As will be shown later, $\bar{f}_{\ell_1}(U)$ captures the expectation of the *empirical loss* $f_{\ell_1}(U)$, when the measurement matrices have i.i.d. Gaussian entries. To analyze the behavior of SubGM on $f_{\ell_1}(U)$ (Algorithm 1), we

first study the ideal scenario, where the loss deviation is zero, and hence, $f_{\ell_1}(U)$ coincides with its expectation. Under such an ideal scenario, we establish the global convergence of SubGM with small initialization. We then extend our result to the general case by carefully controlling the effect of sub-differential deviation. More specifically, we show that the desirable performance of SubGM extends to the empirical loss $f_{\ell_1}(U)$, provided that the sub-differentials are “direction-preserving”, that is, $D \approx \gamma \bar{D}$ for every $D \in \partial f_{\ell_1}(U)$ and $\bar{D} \in \partial \bar{f}_{\ell_1}(U)$, where

$$\partial f_{\ell_1}(U) = \left\{ (Q + Q^\top) U : Q \in \mathcal{Q}(UU^\top - X^*) \right\}, \text{ with } \mathcal{Q}(X) = \frac{1}{m} \sum_{i=1}^m \text{Sign}(\langle A_i, X \rangle - s_i) A_i. \quad (7)$$

Definition 6 (ε -approximate rank- k matrix) *We say matrix X is ε -approximate rank- k if there exists a matrix X' with $\text{rank}(X') \leq k$, such that $\|X - X'\|_F \leq \varepsilon$.*

Definition 7 (Sign-RIP) *The measurements are said to satisfy Sign-RIP with parameters $(k, \delta, \varepsilon, \mathcal{S})$ and a uniformly positive and bounded scaling function $\varphi : \mathcal{S} \rightarrow \mathbb{R}$ over the set \mathcal{S} if for every nonzero ε -approximate rank- k $X, Y \in \mathcal{S}$, and every $Q \in \mathcal{Q}(X)$, we have*

$$\left\langle Q - \varphi(X) \frac{X}{\|X\|_F}, \frac{Y}{\|Y\|_F} \right\rangle \leq \varphi(X) \delta. \quad (8)$$

According to our definition, the scaling function satisfies $\underline{\varphi} \leq \varphi(X) \leq \bar{\varphi}, \forall X \in \mathcal{S}$, for some constants $0 < \underline{\varphi} \leq \bar{\varphi} < \infty$. Without loss of generality, we assume that $\underline{\varphi} \leq 1 \leq \bar{\varphi}$. Later, we will show that this assumption is satisfied for Gaussian measurements and different noise models. Whenever there is no ambiguity, we say the measurements satisfy $(k, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP if they satisfy Sign-RIP with parameters $(k, \delta, \varepsilon, \mathcal{S})$ and a (possibly unknown) uniformly positive and bounded scaling function $\varphi : \mathcal{S} \rightarrow \mathbb{R}$.

Next, we provide the intuition behind Sign-RIP. For any $U \in \mathbb{R}^{d \times r'}$, the rank of $UU^\top - X^*$ is at most $r + r'$. Now, suppose that the measurements satisfy $(r' + r, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP with small δ and suitable choices of ε, \mathcal{S} . Then, upon defining $\gamma = \varphi(UU^\top - X^*) \leq \bar{\varphi}$, we have $\|D - \gamma \bar{D}\| \leq 2\bar{\varphi} \|U\| \delta$ for every $D \in \partial f_{\ell_1}(U)$ and $\bar{D} \in \partial \bar{f}_{\ell_1}(U)$. In other words, for sufficiently small δ , $\partial f_{\ell_1}(U)$ and $\partial \bar{f}_{\ell_1}(U)$ are almost aligned under $(r' + r, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP. A caveat of this analysis is that the required parameters of Sign-RIP depend on the search rank r' . One of the important contributions of this work is to relax this dependency by showing that every matrix in the sequence $\{U_t U_t^\top - X^*\}_{t=0}^T$ generated by SubGM is ε -approximate rank- r , for some small $\varepsilon > 0$.

At the first glance, one may speculate that Sign-RIP is extremely restrictive: it requires the uniform concentration of the set-valued function $\mathcal{Q}(X)$ over ε -approximate rank- k matrices. However, we show that, statistically, Sign-RIP is no more restrictive than ℓ_2 - (Recht et al., 2010) and ℓ_1/ℓ_2 -RIP (Li et al., 2020a), and—unlike its classical counterparts—holds under different noise models.

Definition 8 (Outlier Noise Model) *With probability p , each entry of the noise vector \mathbf{s} is independently drawn from a zero mean distribution \mathbb{P} ; otherwise, it is set to zero.*

Notice that our proposed noise model does not impose any assumption on the magnitude of the nonzero elements of \mathbf{s} , or the specific form of their distribution, which makes it particularly suitable for modeling outliers with arbitrary magnitudes.

Definition 9 (Gaussian Noise Model) *Each element of the noise vector \mathbf{s} is independently drawn from a Gaussian distribution with zero mean and variance $\nu_g^2 < \infty$.*

Our next two theorems characterize the sample complexity of Sign-RIP under the outlier and Gaussian noise models.

Theorem 10 (Sign-RIP under Outlier Noise Model) *Assume that the measurement matrices $\{A_i\}_{i=1}^m$ defining the linear operator $\mathcal{A}(\cdot)$ have i.i.d. standard Gaussian entries, and that the noise vector \mathbf{s} follows the outlier noise model with $0 \leq p < 1$ (Definition 8). Then, with probability of at least $1 - C_1 e^{-C_2 m(1-p)^2 \delta^2}$, $(k, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP holds with parameters $k \leq d$, $\delta \leq 1$, $\mathcal{S} = \{X : \zeta \leq \|X\|_F \leq R\}$ for some arbitrary $R \geq \zeta > 0$, $\varepsilon \lesssim \zeta \sqrt{k/m}$, and a scaling function $\varphi(X) = \sqrt{\frac{2}{\pi}} \left(1 - p + p \mathbb{E} \left[e^{-s^2/(2\|X\|_F^2)} \right]\right)$, provided that the number of samples satisfies $m \gtrsim \frac{dk \log^2(m) \log(R/\zeta)}{(1-p)^2 \delta^2}$.*

The proof of the above theorem is provided in Appendix B.2. Theorem 10 shows that, for any fixed R , ζ , p , and δ , Sign-RIP is satisfied with $\tilde{\mathcal{O}}(dk)$ number of Gaussian measurements, which has the same order as ℓ_2 - (Recht et al., 2010) and ℓ_1/ℓ_2 -RIP (Li et al., 2020a) (modulo logarithmic factors). However, unlike ℓ_2 - and ℓ_1/ℓ_2 -RIP, Sign-RIP is *not* oblivious to noise. In particular, our theorem shows that Sign-RIP holds with a number of samples that scales with $(1-p)^{-2}$, ultimately alleviating the issue raised in Subsection 3.1. Moreover, our result does not impose any restriction on p , which improves upon the assumption $p < 1/\sqrt{r^l}$ made by Li et al. (2020a) and Ding et al. (2021).

Theorem 11 (Sign-RIP for Gaussian noise model) *Assume that the measurement matrices $\{A_i\}_{i=1}^m$ defining the linear operator $\mathcal{A}(\cdot)$ have i.i.d. standard Gaussian entries, and that the noise vector \mathbf{s} follows the Gaussian noise model (Definition 9). Then, with probability of at least $1 - C_1 e^{-C_2 m \zeta^2 \delta^2 / \nu_g^2}$, $(k, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP holds with parameters $k \leq m$, $\delta \leq 1$, $\mathcal{S} = \{X : \zeta \leq \|X\|_F \leq R\}$ for some arbitrary $R \geq \zeta > 0$, $\varepsilon \lesssim \zeta \sqrt{k/m}$, and a scaling function $\varphi(X) = \sqrt{\frac{2}{\pi}} \frac{\|X\|_F}{\sqrt{\|X\|_F^2 + \nu_g^2}}$, provided that the number of samples satisfies $m \gtrsim \frac{\nu_g^2 dk \log^2(m) \log(R/\zeta)}{\zeta^2 \delta^2}$.*

The proof of the above theorem is provided in Appendix B.3. Theorem 11 extends Sign-RIP beyond the outlier noise model, showing that it holds even when all measurements are corrupted with Gaussian noise. However, unlike the outlier noise model, the sample complexity of Sign-RIP scales with the noise variance.

3.3 Choice of Step-size

Next, we discuss our choice of the step-size, and its effect on the performance of SubGM. For simplicity, let $\Delta_t = U_t U_t^\top - X^*$ and $\varphi_t = \varphi(U_t U_t - X^*)$. Under Sign-RIP, we have

$D_t \approx (\varphi_t / \|\Delta_t\|_F) \cdot \Delta_t U_t$ for every $D_t \in \partial f_{\ell_1}(U_t)$. Therefore, the iterations of SubGM can be approximated as $U_{t+1} \approx U_t - (\eta_t \varphi_t / \|\Delta_t\|_F) \cdot \Delta_t U_t$. Consequently, with the choice of $\eta_t = \eta \varphi_t^{-1} \|\Delta_t\|_F$, the iterations of SubGM reduce to

$$U_{t+1} = U_t - \eta \cdot \Delta_t U_t + \text{deviation}.$$

Ignoring the deviation term, the above update coincides with the iterations of GD with a constant step-size η , applied to the expected loss function $\bar{f}_{\ell_2}(U) = \|UU^\top - X^*\|_F^2$. By controlling the effect of the deviation term, we show that SubGM on $\bar{f}_{\ell_2}(U)$ behaves similarly to GD with a constant step-size. A caveat of this analysis is that the proposed step-size $\eta_t = \eta \varphi_t^{-1} \|\Delta_t\|_F$ depends on X^* , which is not known *a priori*. In the noiseless scenario, Sign-RIP can be invoked to show that $\varphi_t^{-1} \|\Delta_t\|_F$ can be accurately estimated by $f_{\ell_1}(U_t)$, as shown in the following lemma.

Lemma 12 *Suppose that the measurements are noiseless, and satisfy $(k, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP for some $\delta \leq 1$, $k \leq d$, $\varepsilon \geq 0$, $\mathcal{S} \neq \emptyset$, and uniformly positive and bounded scaling function $\varphi(\cdot)$. Moreover, suppose that Δ_t is ε -approximate rank- k and $\Delta_t \in \mathcal{S}$. Then, we have*

$$(1 - \delta) \varphi_t \|\Delta_t\|_F \leq f_{\ell_1}(U_t) \leq (1 + \delta) \varphi_t \|\Delta_t\|_F. \quad (9)$$

The above lemma is the byproduct of a more general result presented in Appendix B.4. It implies that, for small δ , the step-size $\eta_t = \eta f_{\ell_1}(U_t)$ satisfies $\eta_t \approx \eta \varphi_t \|\Delta_t\|_F$, and hence, $U_{t+1} \approx U_t - \eta \varphi_t^2 \Delta_t U_t$, which again resembles the iterations of GD on $\bar{f}_{\ell_2}(U)$ with the “effective” step-size $\eta \varphi_t^2$.

In the noisy setting, the value of $\varphi_t^{-1} \|\Delta_t\|_F$ *cannot* be estimated merely based on $f_{\ell_1}(U_t)$, since $f_{\ell_1}(U_t)$ is highly sensitive to the magnitude of the noise. To alleviate this issue, we propose an exponentially decaying step-size that circumvents the need for an accurate estimate of $\|\Delta_t\|_F$. In particular, consider the following choice of step-size

$$\eta_t = \frac{\eta}{\|Q_t\|} \cdot \rho^t, \quad \text{where } Q_t \in \mathcal{Q}(\Delta_t), \quad (10)$$

for appropriate values of η and $0 < \rho < 1$. We note that the set $\mathcal{Q}(\Delta_t)$ can be explicitly characterized without any prior knowledge on Δ_t :

$$\mathcal{Q}(\Delta_t) = \frac{1}{m} \sum_{i=1}^m \text{Sign}(\langle A_i, \Delta_t \rangle - s_i) A_i = \frac{1}{m} \sum_{i=1}^m \text{Sign}(\langle A_i, U_t U_t^\top \rangle - y_i) A_i.$$

Our next lemma shows that the above choice of step-size is well-defined (i.e., $Q_t \neq 0$), so long as Δ_t is not too small and the measurements satisfy $(k, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP.

Lemma 13 *Suppose that the measurements satisfy $(k, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP with $\delta < 2/(1 + 5\sqrt{k})$, $k \leq d$, $\varepsilon > 0$, $\mathcal{S} \neq \emptyset$, and a uniformly positive and bounded scaling function $\varphi(\cdot)$. Moreover, suppose that Δ_t is ε -approximate rank- k , $\|\Delta_t\| \geq 4\varepsilon$, and $\Delta_t \in \mathcal{S}$. Then, we have*

$$\left(1 - \left(\frac{1 + 5\sqrt{k}}{2}\right) \delta\right) \frac{\eta \rho^t \|\Delta_t\|_F}{\varphi_t \|\Delta_t\|} \leq \eta_t \leq \left(1 + \left(\frac{1 + 5\sqrt{k}}{2}\right) \delta\right) \frac{\eta \rho^t \|\Delta_t\|_F}{\varphi_t \|\Delta_t\|}. \quad (11)$$

The proof of the above lemma can be found in Appendix B.5. Lemma 13 implies that the chosen step-size remains close to $(\eta\rho^t/\varphi(\Delta_t))(\|\Delta_t\|_F/\|\Delta_t\|)$, as long as the error is not close to zero. Due to Lemma 13, the iterations of SubGM with exponentially-decaying step-size can be approximated as

$$U_{t+1} = U_t - \left(\frac{\eta\rho^t}{\|\Delta_t\|} \right) \Delta_t U_t + \text{deviation}. \quad (12)$$

In other words, SubGM selects an approximately correct direction of descent, while the exponentially decaying step-size ensures convergence to the ground truth.

3.4 Effect of Over-parameterization

At every iteration of SubGM, the rank of the error matrix $\Delta_t = U_t U_t^\top - X^*$ may be as large as $r + r'$. Therefore, in order to guarantee the direction-preserving property of the sub-differentials, a sufficient condition is to satisfy Sign-RIP for every rank- $(r + r')$ matrix. Such crude analysis implies that the performance of SubGM may depend on the search rank r' . In particular, with Gaussian measurements, this would increase the required number of samples to $\tilde{\mathcal{O}}\left(\frac{dr'}{(1-p)^2\delta^2}\right)$, which scales linearly with the over-parameterized rank. To address this issue, we provide a finer analysis of the iterations. Consider the eigen-decomposition of X^* , given as

$$X^* = [V \quad V_\perp] \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V^\top \\ V_\perp^\top \end{bmatrix}^\top = V \Sigma V^\top,$$

where $V \in \mathbb{R}^{d \times r}$ and $V_\perp \in \mathbb{R}^{d \times (d-r)}$ are (column) orthonormal matrices satisfying $V^\top V_\perp = 0$, and $\Sigma \in \mathbb{R}^{r \times r}$ is a diagonal matrix collecting the nonzero eigenvalues of X^* . We assume that the diagonal entries of Σ are in decreasing order, i.e., $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$. Moreover, without loss of generality, we assume that $\sigma_1 \geq 1 \geq \sigma_r$. Based on this eigen-decomposition, we introduce the *signal-residual decomposition* of U_t as follows:

$$U_t = V S_t + V_\perp \underbrace{(F_t + G_t)}_{E_t}, \quad \text{where} \quad S_t = V^\top U_t, \quad E_t = V_\perp^\top U_t, \quad F_t = E_t P_{S_t}, \quad G_t = E_t P_{S_t}^\perp. \quad (13)$$

In the above expression, P_{S_t} is the orthogonal projection onto the row space of S_t , and $P_{S_t}^\perp$ is its orthogonal complement. It is easy to see that $U_t U_t^\top = X^*$ if and only if $S_t S_t^\top = \Sigma$ and $E_t E_t^\top = 0$. Therefore, our goal is to show that $S_t S_t^\top$ and $E_t E_t^\top$ converge to Σ and 0, respectively. Based on the above signal-residual decomposition, one can write

$$\begin{aligned} \Delta_t &= U_t U_t^\top - X^* \\ &= \underbrace{V \left(S_t S_t^\top - \Sigma \right) V^\top + V S_t E_t^\top V_\perp^\top + V_\perp E_t S_t^\top V^\top + V_\perp F_t F_t^\top V_\perp^\top}_{\text{rank-}4r} + \underbrace{V_\perp G_t G_t^\top V_\perp^\top}_{\text{small norm}}. \end{aligned}$$

An important implication of the above equation is that Δ_t can be treated as an ε -approximate rank- $4r$ matrix, where $\varepsilon = \|V_\perp G_t G_t^\top V_\perp^\top\|_F$. We show that $\|V_\perp G_t G_t^\top V_\perp^\top\|_F = \mathcal{O}(\sqrt{d}\alpha)$ throughout the solution trajectory, and hence, Δ_t is approximately rank- $4r$, provided that

the initialization scale α is small enough. To this goal, we first characterize the generalization error $\|\Delta_t\|$ in terms of the *signal term* $\|S_t S_t^\top - X^*\|$, *cross term* $\|S_t E_t^\top\|$, and the *residual term* $\|E_t E_t^\top\|$.

Lemma 14 *We have*

$$\|\Delta_t\| \leq \left\| \Sigma - S_t S_t^\top \right\| + 2 \left\| S_t E_t^\top \right\| + \left\| E_t E_t^\top \right\|. \quad (14)$$

The proof of the above lemma follows directly from the signal-residual decomposition (13), and is omitted for brevity. Motivated by the above lemma, we will study the dynamics of the signal, cross, and residual terms under different settings.

4. Expected Loss

In this section, we consider a special scenario, where the measurements are noiseless, the measurement matrices $\{A_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries, and the number of measurements m approaches infinity. Evidently, these assumptions do not hold in practice. Nonetheless, our analysis of this ideal scenario will be the building block for our subsequent analysis. Since the number of measurements approaches infinity, the uniform law of large numbers implies that $f_{\ell_1}(U)$ converges to its expectation $\mathbb{E}[f_{\ell_1}(U)]$ almost surely, over any compact set of U (Wainwright, 2019). The next lemma provides the explicit form of $\mathbb{E}[f_{\ell_1}(U)]$.

Lemma 15 *Suppose that the measurements are noiseless and the measurement matrices $\{A_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries. Then, we have*

$$\mathbb{E}[f_{\ell_1}(U)] = \sqrt{\frac{2}{\pi}} \left\| UU^\top - X^* \right\|_F. \quad (15)$$

Proof Due to the i.i.d. nature of $\{A_i\}_{i=1}^m$ and the absence of noise, one can write $\mathbb{E}[f_{\ell_1}(U)] = \mathbb{E}[|\langle A, UU^\top - X^* \rangle|]$, where A is random matrix with i.i.d. standard Gaussian entries. It is easy to see that $\langle A, UU^\top - X^* \rangle$ is a Gaussian random variable with variance $\|UU^\top - X^*\|_F^2$. The proof is completed by noting that, for a zero-mean Gaussian random variable X with variance σ^2 , we have $\mathbb{E}[|X|] = \sqrt{\frac{2}{\pi}}\sigma$. \blacksquare

Next, we study the performance of SubGM with small initialization for $\bar{f}_{\ell_1}(U) = \sqrt{\pi/2}\mathbb{E}[f_{\ell_1}(U)]$. First, it is easy to see that $\partial\bar{f}_{\ell_1}(U) = \frac{2(UU^\top - X^*)U}{\|UU^\top - X^*\|_F}$ for $UU^\top \neq X^*$. Moreover, $\partial\bar{f}_{\ell_1}(U)$ is nonempty and bounded for every U that satisfies $UU^\top = X^*$. Therefore, upon choosing the step-size as $\eta_t = (\eta/2)\|UU^\top - X^*\|_F$, the update rule for SubGM reduces to $U_{t+1} = U_t - \eta_t D_t = \eta(U_t U_t^\top - X^*)U_t$, for any $D_t \in \partial\bar{f}_{\ell_1}(U_t)$. In other words, the iterations of SubGM with step-size $\eta_t = (\eta/2)\|UU^\top - X^*\|_F$ on $\bar{f}_{\ell_1}(U)$ are equivalent to the iterations of GD with constant step-size η on the expected ℓ_2 -loss function $\bar{f}_{\ell_2}(U) = (1/4)\|UU^\top - X^*\|_F^2$.

Due to this equivalence, we instead study the behavior of GD on $\bar{f}_{\ell_2}(U)$. Based on the decomposition of the generalization error in Lemma 14, we show that the iterations of SubGM on the expected loss undergo three phases:

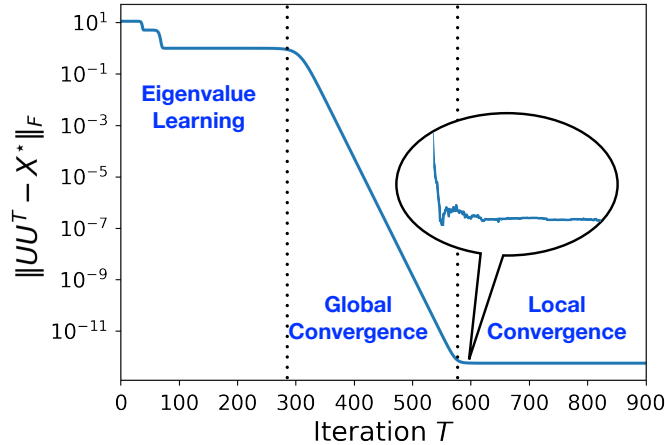


Figure 5: The iterations of SubGM for the expected loss (15) undergo three phases: eigenvalue learning phase, where the eigenvalues of $S_t S_t^\top$ converge to those of X^* ; global convergence phase, where the generalization error decays linearly; and local convergence phase, where the generalization error decays sub-linearly.

- *Eigenvalue learning*: Due to small initialization, the signal, residual, and cross terms are small at the initial point. Therefore, the generalization error is dominated by the signal term $\|S_t S_t^\top - \Sigma\| \approx \|\Sigma\|$. We show that, in the first phase, SubGM improves the generalization error by *learning the eigenvalues of X^** , i.e., by reducing $\|S_t S_t^\top - \Sigma\|$. During this phase, the residual term $\|E_t E_t^\top\|$ will decrease at a sublinear rate.
- *Global convergence*: Once the eigenvalues are learned to a certain accuracy, both signal and cross terms $\|S_t S_t^\top - \Sigma\|$ and $\|S_t E_t^\top\|$ start to decay at a linear rate, while the residual term maintains its sublinear decay rate.
- *Local convergence*: The discrepancy between the decay rates of the signal and cross terms, and that of the residual term implies that, at some point, the residual term becomes the dominant term, and hence, the generalization error starts to decay at a sublinear rate.

Figure 5 illustrates the three phases of SubGM on the expected loss $\bar{f}_{\ell_1}(U)$ with a rank-3 ground truth X^* . Here, we assume that the problem is fully over-parameterized, i.e., $r' = d = 20$. A closer look at the first phase of the algorithm reveals that SubGM learns the eigenvalues of X^* at different rates: the larger eigenvalues are learned faster than the small ones (Figure 6a). A similar observation has been made for gradient flow applied to low-rank matrix factorization (Li et al., 2020b), and is referred to as *incremental learning* (Gidel et al., 2019). Finally, Figure 6b illustrates the dynamics of the signal, cross, and residual terms.

Proposition 16 (Minimum eigenvalue dynamic) *Consider the iterations of SubGM for the expected loss $\bar{f}_{\ell_1}(U)$ with the step-size $\eta_t = (\eta/2)\bar{f}_{\ell_1}(U_t)$. Suppose that $\eta \lesssim 1/\sigma_1$,*

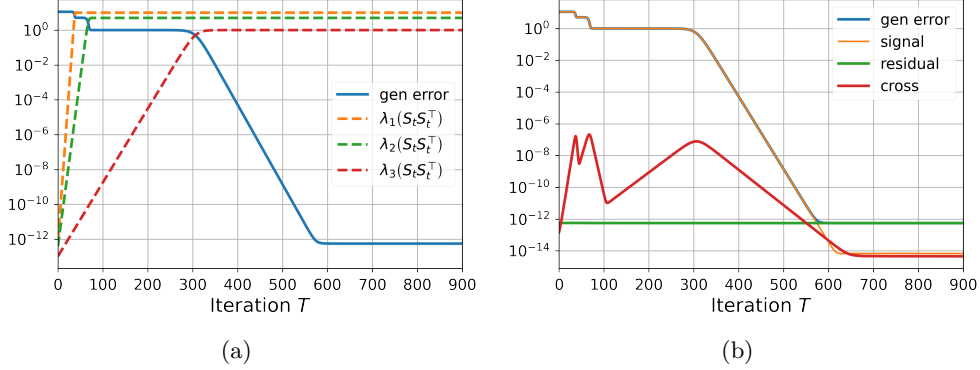


Figure 6: (a) The eigenvalues of X^* are learned at different rates. (b) During the eigenvalue learning phase, the generalization error is dominated by the signal term. In the global convergence phase, both signal and cross terms decay linearly. Finally, the residual term becomes the dominant term in the local convergence phase, and it governs the generalization error.

$S_t S_t^\top \succ 0$, $\|E_t E_t^\top\| \leq \sigma_1$, and $\|S_t S_t\| \leq 2\sigma_1$. Then, we have

$$\lambda_{\min}(S_{t+1} S_{t+1}^\top) \geq \left((1 + \eta\sigma_r)^2 - 2\eta \|E_t E_t^\top\| \right) \lambda_{\min}(S_t S_t^\top) - 2\eta (1 + \eta\sigma_r) \lambda_{\min}(S_t S_t^\top)^2. \quad (16)$$

The proof of Proposition 16 can be found in Appendix C.1. The above proposition shows that the minimum eigenvalue of $S_t S_t^\top$ grows exponentially fast at a rate of $1 + \Theta(\eta\sigma_r)$, provided that η and $\|E_t E_t^\top\|$ are small. This implies that the minimum eigenvalue satisfies $\lambda_{\min}(S_t S_t^\top) \gtrsim \sigma_r$ after $\mathcal{O}(\log(1/\lambda_{\min}(S_0 S_0^\top))/(\eta\sigma_r))$ iterations.

Proposition 17 (Signal, cross, and residual dynamics) Consider the iterations of SubGM for the expected loss $\bar{f}_{\ell_1}(U)$, and with the step-size $\eta_t = (\eta/2)\bar{f}_{\ell_1}(U_t)$. Suppose that $\eta \lesssim 1/\sigma_1$, $\|S_t S_t^\top\| \leq 1.01\sigma_1$ and $\|E_t E_t^\top\| \leq \sigma_1$. Then, we have

$$\|\Sigma - S_{t+1} S_{t+1}^\top\| \leq \left(1 - \eta\lambda_{\min}(S_t S_t^\top)\right) \|\Sigma - S_t S_t^\top\| + 5\eta \|S_t E_t^\top\|^2, \quad (17)$$

$$\|S_{t+1} E_{t+1}^\top\| \leq \left(1 - \eta\lambda_{\min}(S_t S_t^\top) + 2\eta \|\Sigma - S_t S_t^\top\| + 2\eta \|E_t E_t^\top\|\right) \|S_t E_t^\top\|, \quad (18)$$

$$\|E_{t+1} E_{t+1}^\top\| \leq \|E_t E_t^\top\| - \eta \|E_t E_t^\top\|^2, \quad (19)$$

$$\|S_{t+1} S_{t+1}^\top\| \leq 1.01\sigma_1. \quad (20)$$

The proof of Proposition 17 can be found in Appendix C.2. The above proposition shows that, once the minimum eigenvalue of $S_t S_t^\top$ approaches σ_r , the iterations enter the second phase, in which the signal and cross terms start to decay exponentially fast at the rate of $1 - \Theta(\eta\sigma_r)$. Moreover, it shows that the residual term is independent of $\lambda_{\min}(S_t S_t^\top)$, and decreases sublinearly throughout the entire solution path. Given these dynamics, we present our main result.

Theorem 18 (Global convergence of SubGM for expected loss) *Consider the iterations of SubGM for the expected loss $\bar{f}_{\ell_1}(U)$ with the step-size $\eta_t = (\eta/2)\bar{f}_{\ell_1}(U_t)$. Suppose that $\eta \lesssim 1/\sigma_1$ and the initial point is selected such that $\|U_0U_0^\top - 2\alpha^2X^*\| \leq \alpha^2\sigma_r$ for some $\alpha \lesssim \sqrt{\sigma_r}$. Then, the following statements hold:*

- **Linear convergence:** After $\bar{T} \lesssim \frac{\log(\sigma_1/\alpha)}{\eta\sigma_r}$ iterations, we have

$$\|U_{\bar{T}}U_{\bar{T}}^\top - X^*\| \lesssim \alpha^2.$$

- **Sub-linear convergence:** For every $t \geq \bar{T}$, we have

$$\|U_tU_t^\top - X^*\| \lesssim \frac{\alpha^2}{\eta\alpha^2t + 1}.$$

The detailed proof of Theorem 18 is presented in Section A.1. According to the above theorem, for any accuracy $\varepsilon > 0$, one can guarantee $\|U_{\bar{T}}U_{\bar{T}}^\top - X^*\| \leq \varepsilon$ after $\mathcal{O}(\log(\sigma_1/\varepsilon)/(\eta\sigma_r))$ iterations, provided that $\alpha \lesssim \sqrt{\varepsilon} \wedge \sqrt{\sigma_r}$. In Section 5, we will show how to obtain an initial point that satisfies the conditions of Theorem 18.

5. Empirical Loss with Noiseless Measurements

A key difference between the behavior of SubGM for the empirical loss $f_{\ell_1}(U)$ and its expected counterpart $\bar{f}_{\ell_1}(U)$ is the fact that the residual term $\|E_tE_t^\top\|$ no longer enjoys a monotonically decreasing behavior. In particular, Figure 7a shows that, even with an infinitesimal initialization scale α , the residual term grows to a non-negligible value, before decaying linearly to a small level. In order to analyze this behavior, we further decompose E_t as

$$E_t = F_t + G_t, \quad \text{where} \quad F_t = E_tP_{S_t}, \quad \text{and} \quad G_t = E_tP_{S_t}^\perp.$$

Based on the above decomposition and Lemma 14, the generalization error can be written as:

$$\|U_tU_t^\top - X^*\| \leq \|S_tS_t^\top - \Sigma\| + 2\|S_tE_t^\top\| + \|F_tF_t^\top\| + \|G_tG_t^\top\|. \quad (21)$$

This decomposition plays a key role in characterizing the behavior of the residual term: we show that the increasing nature of $\|E_tE_t^\top\|$ in the initial stage of the algorithm can be attributed to the dynamic of $\|F_tF_t^\top\|$. During this phase, the term $\|G_tG_t^\top\|$ also increases, but at a much slower rate. In particular, we show that $\|G_tG_t^\top\|$ remains in the order of α^γ for some $0 < \gamma \leq 2$ throughout the entire solution path. In the second phase, $\|G_tG_t^\top\|$ remains roughly in the same order, while $\|F_tF_t^\top\|$ decays linearly until it is dominated by $\|G_tG_t^\top\|$. At the end of this phase, the overall error will be in the order of $\mathcal{O}(\alpha^\gamma)$. Figure 7b illustrates the behavior of $\|F_tF_t^\top\|$ and $\|G_tG_t^\top\|$, together with $\|E_tE_t^\top\|$.

Similar to our analysis for the expected loss, our first step towards analyzing the behavior of SubGM is to characterize the dynamic of the minimum eigenvalue of $S_tS_t^\top$. For simplicity of notation, we define $\bar{\eta}_t = \eta\varphi(\Delta_t)^2$ in the sequel. Recall that, due to our assumption on $\varphi(\Delta_t)^2$, we have $\underline{\varphi}^2\eta \leq \bar{\eta}_t \leq \bar{\varphi}^2\eta$, provided that $\Delta_t \in \mathcal{S}$.

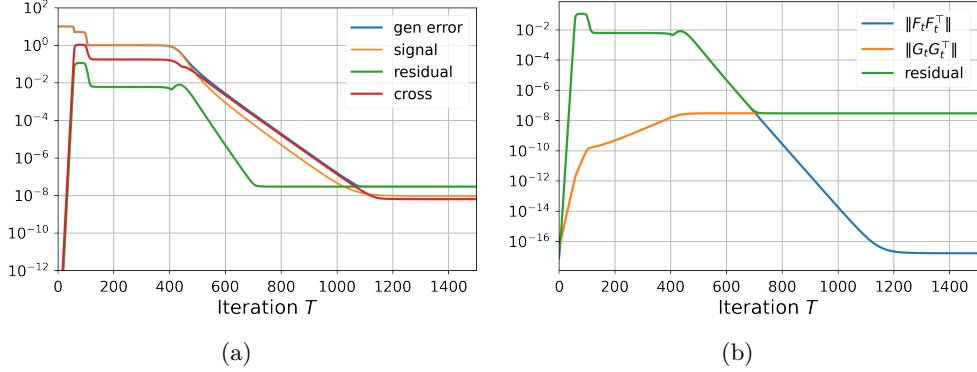


Figure 7: (a) The dynamics of the signal, cross, and residual terms for the empirical loss. Unlike the expected loss, the residual term for the empirical loss has a non-monotonic behavior. (b) The dynamics of $\|F_t F_t^\top\|$ and $\|G_t G_t^\top\|$. The non-monotonic behavior of the residual term can be attributed to the dynamic of $\|F_t F_t^\top\|$.

Proposition 19 (Minimum eigenvalue dynamic) *Consider the iterations of SubGM for the empirical loss $f_{\ell_1}(U)$ with the step-size $\eta_t = (\eta/2)f_{\ell_1}(U_t)$. Suppose that the measurements are noiseless and satisfy $(4r, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP with $\delta \lesssim 1/\sqrt{r}$, $\varepsilon = \sqrt{d}\|G_t\|^2$, and $\mathcal{S} = \{X : \zeta \leq \|X\|_F \leq R\}$ for $\zeta = \varepsilon(1/\delta \vee \sqrt{d})$ and $R = 5\sqrt{r}\sigma_1$. Moreover, suppose that $\eta \lesssim 1/(\bar{\varphi}^2\sigma_1)$, $S_t S_t^\top \succ 0$, $\|E_t E_t^\top\| \leq \sigma_1$, $\|S_t S_t^\top\| \leq 2\sigma_1$, $\Delta_t \in \mathcal{S}$ is ε -approximate rank- $4r$, and $\|E_t S_t^\top (S_t S_t^\top)^{-1}\| \leq 1/3$. Then, we have*

$$\begin{aligned} \lambda_{\min}(S_{t+1} S_{t+1}^\top) &\geq \left((1 + \bar{\eta}_t \sigma_r)^2 - 2\bar{\eta}_t \|E_t E_t^\top\| - 72\bar{\eta}_t \delta \|\Delta_t\|_F \right) \lambda_{\min}(S_t S_t^\top) \\ &\quad - 2\bar{\eta}_t (1 + \bar{\eta}_t \sigma_r) \lambda_{\min}(S_t S_t^\top)^2. \end{aligned}$$

The proof of Proposition 19 can be found in Appendix D.2. Later, we will show that the conditions of Proposition 19 are satisfied with a sufficiently small initial point. The above proposition shows that, in the first phase of the algorithm, $\lambda_{\min}(S_t S_t^\top)$ grows exponentially with a rate of at least $1 + \Omega(\eta \underline{\varphi}^2 \sigma_r)$. Comparing this result with Proposition 16 reveals that $\lambda_{\min}(S_{t+1} S_{t+1}^\top)$ for the empirical loss behaves almost the same as its expected counterpart. This will play an important role in establishing the linear convergence of SubGM for the empirical loss. Finally, note that Sign-RIP must be satisfied for every ε -approximate rank- $4r$ matrix, where $\varepsilon = \sqrt{d}\|G_t\|^2$. Later, we will show that, with small initialization, the value of $\sqrt{d}\|G_t\|^2$ scales with α , and hence, can be kept small throughout the iterations. Our next proposition characterizes the behavior of the signal and cross terms for the empirical loss.

Proposition 20 (Signal and cross dynamics) *Consider the iterations of SubGM for the empirical loss $f_{\ell_1}(U)$ with the step-size $\eta_t = (\eta/2)f_{\ell_1}(U_t)$. Suppose that the measurements are noiseless and satisfy $(4r, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP with $\delta \lesssim 1/\sqrt{r}$, $\varepsilon = \sqrt{d}\|G_t\|^2$, and $\mathcal{S} = \{X : \zeta \leq \|X\|_F \leq R\}$ for $\zeta = \varepsilon(1/\delta \vee \sqrt{d})$ and $R = 5\sqrt{r}\sigma_1$. Moreover, suppose that*

Algorithm 2 Initialization Scheme

Input: initialization scale α , measurement matrices $\{A_i\}_{i=1}^m$, measurement vector $\mathbf{y} = [y_1, \dots, y_m]^\top$, and the search rank r' ;
Output: An initialization matrix $U_0 \in \mathbb{R}^{d \times r'}$;
 Obtain $C \in \frac{1}{m} \sum_{i=1}^m \text{Sign}(y_i) \frac{A_i + A_i^\top}{2}$;
 Compute the eigenvalue decomposition $C = \Lambda D \Lambda^\top$;
 Define $D_+^{r'}$ as the top $r' \times r'$ sub-matrix of D corresponding to the r' largest eigenvalues of C , whose negative values are replaced by 0;
 Set $U_0 = \alpha \Lambda \left(D_+^{r'}\right)^{1/2}$.

$\eta \lesssim 1/(\bar{\varphi}^2 \sigma_1)$, $\|S_t S_t^\top\| \leq 1.01 \sigma_1$, $\|E_t E_t^\top\| \leq \sigma_1$, $\|E_t E_t^\top\|_F \leq \sqrt{r} \sigma_1$, and $\Delta_t \in \mathcal{S}$ is ε -approximate rank- $4r$. Then, we have

$$\|\Sigma - S_{t+1} S_{t+1}^\top\| \leq \left(1 - \bar{\eta}_t \lambda_{\min}(S_t S_t^\top)\right) \|\Sigma - S_t S_t^\top\| + 5\bar{\eta}_t \|S_t E_t^\top\|^2 + 37\bar{\eta}_t \delta \sigma_1 \|\Delta_t\|_F, \quad (22)$$

$$\begin{aligned} \|S_{t+1} E_{t+1}^\top\| &\leq \left(1 - \bar{\eta}_t \lambda_{\min}(S_t S_t^\top) + 2\bar{\eta}_t \|\Sigma - S_t S_t^\top\| + 2\bar{\eta}_t \|E_t E_t^\top\|\right) \|S_t E_t^\top\| \\ &\quad + 22\bar{\eta}_t \delta \sigma_1 \|\Delta_t\|_F, \end{aligned} \quad (23)$$

$$\|S_{t+1} S_{t+1}^\top\| \leq 1.01 \sigma_1. \quad (24)$$

The proof of this proposition is presented in Appendix D.3. Proposition 20 shows that, under Sign-RIP, the one-step dynamics of the signal and cross terms behave almost the same as their expected counterparts, provided that δ is sufficiently small.

Finally, we provide the one-step dynamic of the residual term. To this goal, we will separately analyze F_t and G_t , i.e., the projection of E_t onto the row space of S_t and its orthogonal complement. This together with $E_t = F_t + G_t$ characterizes the dynamic of the residual term.

Proposition 21 (Residual dynamic) Consider the iterations of SubGM for the empirical loss $f_{\ell_1}(U)$ with the step-size $\eta_t = (\eta/2)f_{\ell_1}(U_t)$. Suppose that the measurements are noiseless and satisfy $(4r, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP with $\delta \lesssim 1/\sqrt{r}$, $\varepsilon = \sqrt{d} \|G_t\|^2$, and $\mathcal{S} = \{X : \zeta \leq \|X\|_F \leq R\}$ for $\zeta = \varepsilon(1/\delta \vee \sqrt{d})$ and $R = 5\sqrt{r} \sigma_1$. Moreover, suppose that $\|S_t S_t^\top\| \leq 1.01 \sigma_1$, $\|E_t E_t^\top\| \leq \sigma_1$, $\Delta_t \in \mathcal{S}$ is ε -approximate rank- $4r$, and $\|E_t S_t (S_t S_t)^{-1}\| \leq 1/3$. Then, the following statements hold:

- If $\bar{\eta}_t \lesssim 1/\|\Delta_t\|$, we have

$$\|G_{t+1}\| \leq \left(1 + \bar{\eta}_t^2 \left(2 \|E_t S_t^\top\|^2 + \|E_t\|^4 + 2 \|\Delta_t\| \|E_t S_t^\top\|\right) + 7\bar{\eta}_t \delta \|\Delta_t\|_F\right) \|G_t\|.$$

- If $\eta \lesssim 1/(\bar{\varphi}^2 \sigma_1)$, we have

$$\|F_{t+1}\| \leq \left(1 - \bar{\eta}_t \lambda_{\min}(S_t S_t^\top) + 3\bar{\eta}_t \delta \|\Delta_t\|_F\right) \|F_t\| + 3\bar{\eta}_t \delta \|\Delta_t\|_F \|S_t\| + 6\bar{\eta}_t \|\Delta_t\| \|G_t\|.$$

The proof of the above proposition can be found in Appendix D.4. Note that the condition $\eta \lesssim 1/(\bar{\varphi}^2\sigma_1)$ for the dynamic of $\|F_t\|$ readily implies $\bar{\eta}_t \lesssim 1/\|\Delta_t\|$. Therefore, the one-step dynamic of $\|G_t\|$ holds under a milder condition on the step-size. Moreover, unlike $\|F_t\|$, the dynamic of $\|G_t\|$ is independent of $\lambda_{\min}(S_t S_t^\top)$. At the early stages of the algorithm, the term $\bar{\eta}_t^2 \left(2\|E_t S_t^\top\|^2 + \|E_t\|^4 + 2\|\Delta_t\| \|E_t S_t^\top\| \right) = \mathcal{O}(\alpha^2)$ is dominated by $\bar{\eta}_t \delta \|\Delta_t\|_F \approx \bar{\eta}_t \delta \|X^*\|_F$. Therefore, $\|G_t\|$ grows at a slow rate of $1 + \mathcal{O}(1)\eta\delta\bar{\varphi}^2 \|X^*\|_F$. As the algorithm makes progress, $\|\Delta_t\|_F$ decreases, leading to an even slower growth rate for $\|G_t\|$. This is in line with the behavior of $\|G_t\|$ in Figure 7b: the growth rate of $\|G_t\|$ decreases as SubGM makes progress towards the ground truth, and it eventually “flattens out” at a level proportional to the initialization scale. However, unlike $\|G_t\|$, the term $\|F_t\|$ does not have a monotonic behavior. In particular, according to Proposition 21, $\|F_t\|$ may increase at the early stages of the algorithm, where $\lambda_{\min}(S_t S_t^\top)$ is negligible compared to $\|\Delta_t\|_F$. However, $\|F_t\|$ will start decaying as soon as $\lambda_{\min}(S_t S_t^\top) \gtrsim \delta \|\Delta_t\|_F$, which, according to Proposition 19, is guaranteed to happen after certain number of iterations. The non-monotonic behavior of $\|F_t\|$ is also observed in practice (see Figure 7b).

Before presenting the main result, we provide our proposed initialization scheme in Algorithm 2. The presented initialization method is analogous to the classical spectral initialization in the noiseless matrix recovery problems, with a key difference that we scale down the norm of the initial point by a factor of α^2 . As will be shown later, the scaling of the initial point is crucial for establishing the linear convergence of SubGM; without such scaling, both GD and SubGM suffer from sublinear convergence rates, as evidenced by the recent works (Zhuo et al., 2021; Ding et al., 2021).

Theorem 22 (Global Convergence of SubGM with Noiseless Measurements)

Consider the iterations of SubGM for the empirical loss $f_{\ell_1}(U)$ with the step-size $\eta_t = (\eta/2)f_{\ell_1}(U_t)$. Suppose that the initial point U_0 is obtained from Algorithm 2 with an initialization scale that satisfies $\alpha \lesssim 1/(\bar{\varphi}\sqrt{d}) \wedge 1/\kappa$. Suppose that the measurements are noiseless and satisfy $(4r, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP with $\delta \lesssim 1/(\sqrt{r}\kappa^2\bar{\varphi}^4 \log^2(1/\alpha))$, $\varepsilon \asymp \sqrt{d}\alpha^{2-\mathcal{O}(\sqrt{r}\kappa^2\delta)}\delta$, and $\mathcal{S} = \{X : \zeta \leq \|X\|_F \leq R\}$ for $\zeta = \varepsilon(1/\delta \vee \sqrt{d})$ and $R = 5\sqrt{r}\sigma_1$. Finally, suppose that $\eta \lesssim 1/(\bar{\varphi}^2\sigma_1)$. Then, after $t = T_{\text{end}} \lesssim \log(1/\alpha)/(\eta\sigma_r\bar{\varphi}^2)$ iterations, we have

$$\left\| U_t U_t^\top - X^* \right\|_F \lesssim d\alpha^{2-\mathcal{O}(\sqrt{r}\kappa^2\delta)}.$$

The proof of Theorem 22 is presented in Subsection A.2, and follows the same structure as the proof of Theorem 18. However, unlike the expected loss, the final error will be in the order of $\alpha^{\beta(\delta)}$, for some $0 < \beta(\delta) \leq 2$ that is a decreasing function of δ . Indeed, smaller δ will improve the dependency of the final generalization error on α . Moreover, for an arbitrarily small $\varepsilon > 0$, one can guarantee $\|U_T U_T^\top - X^*\|_F \leq \varepsilon$ within $T \lesssim \log(d/\varepsilon)/(\beta(\delta)\eta\sigma_r\bar{\varphi}^2)$ iterations, provided that the initialization scale satisfies $\alpha \lesssim (\varepsilon/d)^{1/\beta(\delta)}$.

Finally, we characterize the sample complexity of SubGM with noiseless, Gaussian measurements.

Corollary 23 (Gaussian Measurements) *Suppose that the measurement matrices $\{A_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries. Consider the iterations of SubGM for*

the empirical loss $f_{\ell_1}(U)$, with the step-size $\eta_t = (\eta/2)f_{\ell_1}(U_t)$ and $\eta \lesssim 1/\sigma_1$. Suppose that the initial point U_0 is obtained from Algorithm 2 with an initialization scale that satisfies $\alpha \lesssim 1/\sqrt{d} \wedge 1/\kappa$. Finally, suppose that the number of measurements satisfies $m \gtrsim \kappa^4 d r^2 \log^5(1/\alpha^2) \log^2(m)$. Then, after $t = T_{\text{end}} \lesssim \log(1/\alpha)/(\eta\sigma_r)$ iterations, and with an overwhelming probability, we have

$$\left\| U_T U_T^\top - X^\star \right\|_F \lesssim d\alpha^{2-\mathcal{O}\left(\sqrt{\frac{\kappa^4 d r^2 \log(1/\alpha) \log^2(m)}{m}}\right)}.$$

The above corollary is a direct consequence of Theorem 10 after setting the corruption probability p to zero. To the best of our knowledge, Corollary 23 is the first result showing that, with Gaussian measurements, the sample complexity of SubGM is *independent* of the search rank, provided that the initial point is sufficiently close to the origin.

6. Empirical Loss with Noisy Measurements

In this section, we establish the convergence of SubGM with small initialization and noisy measurements. A key difference compared to our previous analysis is in the choice of the step-size: in the presence of noise, the value of $\eta_t = (\eta/2)f_{\ell_1}(U_t)$ can be arbitrarily far from the error $\eta\varphi_t^{-1} \|\Delta_t\|$. To circumvent this issue, we instead propose to use the following geometric step-size:

$$\eta_t = \eta \cdot \frac{\rho^t}{\|Q_t\|}, \quad \text{where } Q_t \in \mathcal{Q}(\Delta_t). \quad (25)$$

Our first goal is to show that, under a similar Sign-RIP condition, our previous guarantees on SubGM extend to geometric step-size. Then, we show how our general result can be readily tailored to specific noise models. Our next result characterizes the dynamic of $\lambda_{\min}(S_t S_t^\top)$ with the above choice of step-size.

Proposition 24 (Minimum eigenvalue dynamic) *Consider the iterations of SubGM on $f_{\ell_1}(U)$ with the step-size defined as (25). Suppose that the measurements satisfy $(4r, \varepsilon, \delta, \mathcal{S})$ -Sign-RIP with $\delta \lesssim 1/\sqrt{r}$, $\varepsilon = \sqrt{d} \|G_t\|^2$, and $\mathcal{S} = \{X : \zeta \leq \|X\|_F \leq R\}$ for $\zeta = \varepsilon (1/\delta \vee \sqrt{d})$ and $R = 5\sqrt{r}\sigma_1$. Moreover, suppose that $S_t S_t^\top \succ 0$, $\|E_t E_t^\top\| \leq \sigma_1$, $\|S_t S_t^\top\| \leq 2\sigma_1$, $\|E_t S_t^\top (S_t S_t^\top)^{-1}\| \leq 1/3$, $\frac{\eta\rho^t}{\|\Delta_t\|} \lesssim \frac{1}{\sigma_1}$, and $\Delta_t \in \mathcal{S}$ is ε -approximate rank- $4r$. Then, we have*

$$\begin{aligned} \lambda_{\min}(S_{t+1} S_{t+1}^\top) &\geq \left(\left(1 + \frac{\eta\rho^t}{\|\Delta_t\|} \sigma_r \right)^2 - \frac{2\eta\rho^t}{\|\Delta_t\|} \|E_t E_t^\top\| - 384\sqrt{r}\eta\rho^t\delta \right) \lambda_{\min}(S_t S_t^\top) \\ &\quad - 2 \frac{\eta\rho^t}{\|\Delta_t\|} \left(1 + \frac{\eta\rho^t}{\|\Delta_t\|} \sigma_r \right) \lambda_{\min}(S_t S_t^\top)^2. \end{aligned}$$

The proof of the above proposition is presented in Appendix E.2. Recalling our discussion in Section 3.3, SubGM with geometric step-size moves towards a direction close to $\Delta_t/\|\Delta_t\|$ with an “effective” step-size of $\eta\rho^t/\|\Delta_t\|$. In light of this, the above proposition is analogous to Proposition 19, with an additional assumption that the effective step-size

is upper bounded by $1/\sigma_1$. Proposition 24 can be used to show the exponential growth of $\lambda_{\min}(S_{t+1}S_{t+1}^\top)$ in the first phase of the algorithm. To see this, note that, due to small initialization, we have $\|\Delta_t\| \approx \|X^*\| = \sigma_1$, $\lambda_{\min}(S_t S_t^\top)^2 \ll \lambda_{\min}(S_t S_t^\top)$, and $\|E_t E_t^\top\| \approx 0$ at the early stages of the algorithm. This implies that the minimum eigenvalue dynamic can be accurately approximated as $\lambda_{\min}(S_{t+1}S_{t+1}^\top) \geq (1 + \Omega(\eta/\kappa)) \lambda_{\min}(S_t S_t^\top)$, which grows exponentially fast. We next characterize the dynamics of the signal and cross terms.

Proposition 25 (Signal and cross dynamics) *Consider the iterations of SubGM on $f_{\ell_1}(U)$ with the step-size defined as (25). Suppose that the measurements satisfy $(4r, \varepsilon, \delta, \mathcal{S})$ -Sign-RIP with $\delta \lesssim 1/\sqrt{r}$, $\varepsilon = \sqrt{d}\|G_t\|^2$, and $\mathcal{S} = \{X : \zeta \leq \|X\|_F \leq R\}$ for $\zeta = \varepsilon(1/\delta \vee \sqrt{d})$ and $R = 5\sqrt{r}\sigma_1$. Moreover, suppose that $S_t S_t^\top \succ 0$, $\|E_t E_t^\top\| \leq \sigma_1$, $\|S_t S_t^\top\| \leq 1.01\sigma_1$, $\|E_t S_t^\top (S_t S_t^\top)^{-1}\| \leq 1/3$, $\frac{\eta\rho^t}{\|\Delta_t\|} \lesssim \frac{1}{\sigma_1}$, and $\Delta_t \in \mathcal{S}$ is ε -approximate rank- $4r$. Then, we have*

$$\|\Sigma - S_{t+1}S_{t+1}^\top\| \leq \left(1 - \frac{\eta\rho^t}{\|\Delta_t\|} \lambda_{\min}(S_t S_t^\top)\right) \|\Sigma - S_t S_t^\top\| + 5\frac{\eta\rho^t}{\|\Delta_t\|} \|S_t E_t^\top\|^2 + 193\sqrt{r}\eta\rho^t \delta \sigma_1, \quad (26)$$

$$\|S_{t+1}E_{t+1}^\top\| \leq \left(1 - \frac{\eta\rho^t}{\|\Delta_t\|} \left(\lambda_{\min}(S_t S_t^\top) - 2\|\Sigma - S_t S_t^\top\| - 2\|E_t E_t^\top\|\right)\right) \|S_t E_t^\top\| + 113\sqrt{r}\eta\rho^t \delta \sigma_1, \quad (27)$$

$$\|S_{t+1}S_{t+1}^\top\| \leq 1.01\sigma_1. \quad (28)$$

The proof of Proposition 25 is analogous to Proposition 20, and can be found in Appendix E.3. Assuming $\|\Delta_t\| \asymp \rho^t$, the above proposition shows that both signal and cross terms behave similarly to their expected counterparts in Proposition 17, and their deviation diminishes exponentially fast.

Proposition 26 *Consider the iterations of SubGM on $f_{\ell_1}(U)$ with the step-size defined as (25). Suppose that the measurements satisfy $(4r, \varepsilon, \delta, \mathcal{S})$ -Sign-RIP with $\delta \lesssim 1/\sqrt{r}$, $\varepsilon = \sqrt{d}\|G_t\|^2$, and $\mathcal{S} = \{X : \zeta \leq \|X\|_F \leq R\}$ for $\zeta = \varepsilon(1/\delta \vee \sqrt{d})$ and $R = 5\sqrt{r}\sigma_1$. Moreover, suppose that $S_t S_t^\top \succ 0$, $\|E_t E_t^\top\| \leq \sigma_1$, $\|S_t S_t^\top\| \leq 1.1\sigma_1$, $\|E_t S_t^\top (S_t S_t^\top)^{-1}\| \leq 1/3$, and $\Delta_t \in \mathcal{S}$ is ε -approximate rank- $4r$. Then, the following statements hold:*

- If $\eta \lesssim \frac{1}{\|\Delta_t\|}$, we have

$$\|G_{t+1}\| \leq \left(1 + \frac{\eta^2 \rho^{2t}}{\|\Delta_t\|^2} \left(2\|E_t S_t^\top\|^2 + \|E_t\|^4 + 2\|\Delta_t\| \|E_t S_t^\top\|\right)\right) \|G_t\| + 49\sqrt{r}\eta_0 \rho^t \delta \|G_t\|,$$

which can be further simplified as

$$\|G_{t+1}\| \leq (1 + 5\eta^2 \rho^{2t} + 49\sqrt{r}\eta_0 \rho^t \delta) \|G_t\|. \quad (29)$$

- If $\frac{\eta\rho^t}{\|\Delta_t\|} \lesssim \frac{1}{\sigma_1}$, we have

$$\|F_{t+1}\| \leq \left(1 - \frac{\eta\rho^t}{\|\Delta_t\|} \lambda_{\min}(S_t S_t^\top) + 16\sqrt{r}\eta\rho^t \delta\right) \|F_t\| + 16\sqrt{r}\eta\rho^t \delta \|S_t\| + 6\eta\rho^t \|G_t\|.$$

The proof of Proposition 26 follows that of Proposition 21, and can be found in Appendix E.4. Inequality (29) implies that the growth rate of $\|G_t\|$ diminishes with t . We will use this property to show that $\|G_t\|$ remains proportional to the initialization scale α throughout the solution trajectory, which will be used to control the final generalization error. Moreover, unlike the dynamic of $\|F_t\|$, (29) holds even when $\|\Delta_t\|$ decays faster than $\eta\sigma_1\rho^t$; this will play a key role in the proof of our next theorem.

Theorem 27 (Global Convergence of SubGM with Noisy Measurements)

Consider the iterations of SubGM on $f_{\ell_1}(U)$ with the step-size defined as (25), and parameters $\eta \lesssim 1/(\kappa \log(1/\alpha))$ and $\rho = 1 - \Theta(\eta/(\kappa \log(1/\alpha)))$. Suppose that the initial point U_0 is obtained from Algorithm 2 with an initialization scale that satisfies $\alpha \lesssim 1/(\sqrt{d}) \wedge 1/\kappa \wedge \underline{\varphi}$. Suppose that the measurements satisfy $(4r, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP with $\delta \lesssim 1/(\sqrt{r}\kappa^2\bar{\varphi}^4 \log^2(1/\alpha))$, $\varepsilon \asymp \sqrt{d}\alpha^{2-\mathcal{O}(\sqrt{r}\kappa\delta)}\delta$, and $\mathcal{S} = \{X : \zeta \leq \|X\|_F \leq R\}$ for $\zeta \geq \varepsilon(1/\delta \vee \sqrt{d})$ and $R = 5\sqrt{r}\sigma_1$. Then, for any $t \geq T_{end} = \mathcal{O}((\kappa/\eta) \log^2(1/\alpha))$, we have

$$\left\| U_t U_t^\top - X^\star \right\|_F \lesssim d\alpha^{2-\mathcal{O}(\sqrt{r}\kappa\delta)} \vee \zeta.$$

The proof of the above theorem can be found in Section A.3. Upon defining $\beta(\delta) = 2 - \mathcal{O}(\sqrt{r}\kappa\delta)$, the above result implies that, for any arbitrary accuracy $\varepsilon \geq \zeta$, SubGM converges to a solution that satisfies $\|U_t U_t^\top - X^\star\|_F \leq \varepsilon$ within $\mathcal{O}(\kappa \log^2(d/\varepsilon)/(\beta(\delta)\eta))$ iterations, provided that $\alpha \lesssim (\varepsilon/d)^{1/\beta(\delta)}$. Compared to the noiseless setting, the final error in Theorem (27) has an additional term ζ . This is due to the fact that we only require a lower bound on the choice of ζ ; as will be explained later, this additional freedom will be used to show the convergence of SubGM under the Gaussian noise model. Moreover, compared to the noiseless setting, the iteration complexity of SubGM in the noisy regime is higher by a factor of $\log(d/\varepsilon)$, and its step-size must be chosen more conservatively. The higher iteration complexity is due to the lack of a prior estimate of $\|\Delta_t\|_F$; to alleviate this issue, we proposed a geometric step-size, which inevitably leads to a slightly higher iteration complexity.

In the proof of the above theorem, we show that SubGM already achieves a small error after T_{end} iterations, and any additional iterations may in fact worsen the final error. In fact, this is the main reason behind the early stopping of gradient descent with constant step-size, as shown by Li et al. (2018); Stöger and Soltanolkotabi (2021). However, we no longer require an early stopping for SubGM. This is due to the fact that, because of their exponentially-decaying nature, the step-sizes become so small that SubGM stops making any meaningful progress after T_{end} iterations. More precisely, we control the dynamics of SubGM in the last $t - T_{end}$ iterations, showing that it does not deviate too much from the solution at iteration T_{end} .

Finally, according to Theorem 27, ρ must satisfy $1 - \rho = \Theta(\eta/(\kappa \log(1/\alpha)))$, which requires a prior knowledge of the condition number κ . However, our analysis in fact guarantees the convergence of SubGM for any $1 - \Theta(\eta/(\kappa \log(1/\alpha))) \leq \rho < 1$. Therefore, if κ is not known (or only a conservative upper bound of κ is known), ρ can be conservatively close to one at the expense of a slower convergence rate.

Equipped with the above theorem and Theorems 10 and 11, we next characterize the behavior of SubGM under both outlier and Gaussian noise regimes.

Corollary 28 (Outlier Noise Model) *Suppose that the measurement matrices $\{A_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries, and the noise vector \mathbf{s} follows an outlier noise model with a corruption probability $0 \leq p < 1$ (Definition 8). Consider the iterations of SubGM on $f_{\ell_1}(U)$ with the step-size defined as (10), and parameters $\eta \lesssim 1/(\kappa \log(1/\alpha))$ and $\rho = 1 - \Theta(\eta/(\kappa \log(1/\alpha)))$. Suppose that the initial point U_0 is obtained from Algorithm 2 with an initialization scale that satisfies $\alpha \lesssim 1/\sqrt{d} \wedge 1/\kappa \wedge (1-p)$. Suppose that the number of measurements satisfies $m \gtrsim \kappa^4 dr^2 \log^5(1/\alpha^2) \log^2(m)/(1-p)^2$. Then, for any $t \geq T_{\text{end}} = \mathcal{O}((\kappa/\eta) \log^2(1/\alpha))$ and with an overwhelming probability, we have*

$$\|U_t U_t^\top - X^*\|_F \lesssim d\alpha^{2-\mathcal{O}\left(\sqrt{\frac{\kappa^2 dr^2 \log(1/\alpha) \log^2(m)}{(1-p)^2 m}}\right)}.$$

Corollary 29 (Gaussian Noise Model) *Suppose that the measurement matrices $\{A_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries, and the noise vector \mathbf{s} follows a Gaussian noise model with a variance $\nu_g < \infty$ (Definition 9). Consider the iterations of SubGM on $f_{\ell_1}(U)$ with the step-size defined as (10), and parameters $\eta \lesssim 1/(\kappa \log(1/\alpha))$ and $\rho = 1 - \Theta(\eta/(\kappa \log(1/\alpha)))$. Suppose that the initial point U_0 is obtained from Algorithm 2 with an initialization scale that satisfies $\alpha \lesssim 1/\sqrt{d} \wedge \sqrt{dr}/m \wedge 1/\kappa$. Then, for any $t \geq T_{\text{end}} = \mathcal{O}((\kappa/\eta) \log^2(1/\alpha))$ and with an overwhelming probability, we have*

$$\|U_t U_t^\top - X^*\|_F = \mathcal{O}\left(\sqrt{\frac{\nu_g^2 \kappa^4 dr^2 \log^5(1/\alpha) \log^2(m)}{m}}\right).$$

The proof of Corollary 29 follows directly from Theorems 11 and 27 after choosing

$$\zeta = C \sqrt{\nu_g^2 \kappa^4 dr^2 \log^5(1/\alpha) \log^2(m)/m},$$

for sufficiently large constant C . The details are omitted for brevity.

Remark 30 *Our result can be readily extended to settings where the measurements are corrupted with both outlier and Gaussian noise values. Consider measurements of the form $y_i = \langle A_i, X^* \rangle + s_i^{(1)} + s_i^{(2)}$, where $s_i^{(1)}$ and $s_i^{(2)}$ follow the outlier and Gaussian noise models delineated in Definitions 8 and 9. In this setting, Corollaries 28 and 29 can be combined to show that, with $m = \tilde{\Omega}(\nu_g^2 \kappa^4 dr^2 / (1-p)^2)$ samples, SubGM with small initialization and geometric step-size achieves the error $\|U_t U_t^\top - X^*\|_F^2 = \tilde{\mathcal{O}}(\nu_g^2 \kappa^4 dr^2 / ((1-p)^2 m))$ (modulo logarithmic factors).*

Remark 31 *We highlight that our required measurement number for the convergence of SubGM must scale poly-logarithmically with the initialization scale. According to Theorems 22 and 27, the final error of SubGM depends on the initialization scale, which in turn leads to the dependency of the final error on the measurement number. This dependency is inevitable in the Gaussian noise regime (Corollary 29) due to the information-theoretical lower bounds. Moreover, for the noiseless (Corollary 23) and outlier noise settings (Corollary 28), the dependency of the final error on the measurement number is of the form $\exp(-\Omega(\text{poly}(m)))$. As such, to obtain an ϵ -accurate solution, the measurement number must scale as $\text{poly} \log(1/\epsilon)$. While we note that this dependency is mild, as a future research direction, we will investigate the possibility of removing such a dependency.*

7. Concluding Remarks

In this work, we study the performance of sub-gradient method (SubGM) on a nonconvex and nonsmooth formulation of the robust matrix recovery with noisy measurements, where the rank of the true solution r is unknown and over-estimated instead with $r' \geq r$. We prove that the over-estimation of the rank has no effect on the performance of SubGM, provided that the initial point is sufficiently close to the origin. Moreover, we prove that SubGM is robust against outlier and Gaussian noise values. In particular, we show that SubGM provably converges to the ground truth, even if the globally optimal solutions of the problem are “spurious”, i.e., they do not correspond to the ground truth. At the heart of our method lies a new notion of restricted isometry property, called Sign-RIP, which guarantees a direction-preserving property for the sub-differentials of the ℓ_1 -loss. We show that, while the classical notions of restricted isometry property face major breakdowns in the face of noise, Sign-RIP can handle a wide range of noisy measurements, and hence, is better-suited for analyzing the robust variants of low-rank matrix recovery. A few remarks are in order next:

Spectral vs. random initialization: In our work, we assume that the initial point is obtained via a special form of the spectral method, followed by a norm reduction. A natural question thus arises as to whether the spectral method can be replaced by small random initialization. Here we briefly mention the potential technical difficulty in extending our theoretical result to the random initialization. According to the proofs of Theorems 22 and 27, we require the minimum eigenvalue of the signal term to dominate the maximum eigenvalues of the cross and residual terms at the initial point. If the initial point is chosen randomly, it is likely that such an assumption is violated. As a consequence, it is unclear how to ensure the initial exponential growth of the signal term under our current error perturbation analysis. Despite this theoretical hurdle, we observed in our simulations that SubGM with small random initialization behaves almost the same as SubGM with spectral initialization. Therefore, we conjecture that a small random initialization followed by a few iterations of SubGM is in fact equivalent to spectral initialization; a similar result has been recently proven by Stöger and Soltanolkotabi (2021) for gradient descent on ℓ_2 -loss. We consider a rigorous verification of this conjecture as an enticing challenge for future research.

Beyond Sign-RIP: Another natural question pertains to the performance of SubGM on problems that *do not* satisfy Sign-RIP. An important and relevant example is *over-parameterized matrix completion*, where the linear measurement operator is an element-wise projector that reveals partial and potentially noisy observations of a low-rank matrix. Indeed, the performance SubGM on problems of this type requires a more refined analysis, which is left as future work.

Acknowledgments

This research is supported by NSF Award DMS-2152776 and ONR Award N00014-22-1-2127. The authors would like to thank Richard Y. Zhang and Cédric Jozs for fruitful discussions on earlier versions of this manuscript. The authors would also like to thank an

anonymous reviewer for pointing out the connection between the exponential decay in the step-size and the final error, which led the authors to remove the early-stopping condition for SubGM.

References

- Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016.
- Thierry Bouwmans and El Hadi Zahzah. Robust pca via principal component pursuit: A review for a comparative evaluation in video surveillance. *Computer Vision and Image Understanding*, 122:22–34, 2014.
- Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.
- Emmanuel J Candes 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.
- Yan Mei Chen, Xiao Shan Chen, and Wen Li. On perturbation bounds for orthogonal projections. *Numerical Algorithms*, 73(2):433–444, 2016.
- Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.
- Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.
- Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.
- Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.
- Lijun Ding, Liwei Jiang, Yudong Chen, Qing Qu, and Zhihui Zhu. Rank overspecified robust matrix recovery: Subgradient method and exact recovery. *arXiv preprint arXiv:2109.11154*, 2021.
- Stanley C Eisenstat and Ilse CF Ipsen. Relative perturbation techniques for singular value problems. *SIAM Journal on Numerical Analysis*, 32(6):1972–1988, 1995.
- Stanley C Eisenstat and Ilse CF Ipsen. Relative perturbation results for eigenvalues and eigenvectors of diagonalisable matrices. *BIT Numerical Mathematics*, 38(3):502–509, 1998.

- Salar Fattahi and Somayeh Sojoudi. Exact guarantees on the absence of spurious local minima for non-negative rank-1 robust principal component analysis. *Journal of machine learning research*, 2020.
- Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.
- Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.
- Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *arXiv preprint arXiv:1904.13262*, 2019.
- Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.
- Cedric Jozs, Yi Ouyang, Richard Y Zhang, Javad Lavaei, and Somayeh Sojoudi. A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization. *arXiv preprint arXiv:1805.08204*, 2018.
- Kenji Kawaguchi. Deep learning without poor local minima. *arXiv preprint arXiv:1605.07110*, 2016.
- Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. *SIAM Journal on Optimization*, 30(1):660–686, 2020a.
- Yuanxin Li, Cong Ma, Yuxin Chen, and Yuejie Chi. Nonconvex matrix factorization from rank-one measurements. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1496–1505. PMLR, 2019.
- Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.
- Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *arXiv preprint arXiv:2012.09839*, 2020b.
- Xiao Luan, Bin Fang, Linghui Liu, Weibin Yang, and Jiye Qian. Extracting sparse error of robust pca for face recognition in the presence of varying illumination and occlusion. *Pattern Recognition*, 47(2):495–508, 2014.
- Xin Luo, Mengchu Zhou, Yunni Xia, and Qingsheng Zhu. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Transactions on Industrial Informatics*, 10(2):1273–1284, 2014.
- Jianhao Ma and Salar Fattahi. Implicit regularization of sub-gradient method in robust matrix recovery: Don’t be afraid of outliers. *arXiv preprint arXiv:2102.02969*, 2021a.

- Jianhao Ma and Salar Fattahi. Sign-rip: A robust restricted isometry property for low-rank matrix recovery. *arXiv preprint arXiv:2102.02969*, 2021b.
- Jianhao Ma and Salar Fattahi. Blessing of depth in linear regression: Deeper models have flatter landscape around the true solution. In *Advances in Neural Information Processing Systems*, 2022.
- Jianhao Ma, Lingjun Guo, and Salar Fattahi. Behind the scenes of gradient descent: A trajectory analysis via basis function decomposition. *arXiv preprint arXiv:2210.00346*, 2022.
- Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.
- Qing Qu, Yuexiang Zhai, Xiao Li, Yuqian Zhang, and Zhihui Zhu. Analysis of the optimization landscapes for overcomplete representation learning. *arXiv preprint arXiv:1912.02427*, 2019.
- 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.
- Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *arXiv preprint arXiv:2106.15013*, 2021.
- Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016.
- Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021.
- Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.
- Ramon Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- Gavin Zhang, Salar Fattahi, and Richard Y Zhang. Preconditioned gradient descent for overparameterized nonconvex burer–monteiro factorization with global optimality certification. *arXiv preprint arXiv:2206.03345*, 2022.
- Jialun Zhang, Salar Fattahi, and Richard Zhang. Preconditioned gradient descent for overparameterized nonconvex matrix factorization. *Advances in Neural Information Processing Systems*, 34, 2021.

Richard Y Zhang. Sharp global guarantees for nonconvex low-rank matrix recovery in the overparameterized regime. *arXiv preprint arXiv:2104.10790*, 2021.

Richard Y Zhang, Somayeh Sojoudi, and Javad Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *J. Mach. Learn. Res.*, 20(114):1–34, 2019.

Yuqian Zhang, Qing Qu, and John Wright. From symmetry to geometry: Tractable non-convex problems. *arXiv preprint arXiv:2007.06753*, 2020.

Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. *arXiv preprint arXiv:1506.06081*, 2015.

Jiacheng Zhuo, Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the computational and statistical complexity of over-parameterized matrix sensing. *arXiv preprint arXiv:2102.02756*, 2021.

Appendix A. Proofs of the Main Theorems

A.1 Proof of Theorem 18

Before delving into the details, we first present the general overview of our proof technique for Theorem 18. First, we prove that the conditions of Propositions 16 and 17 hold for every $0 \leq t < \infty$. Then, we use the minimum eigenvalue dynamic in Proposition 17 to show that $\lambda_{\min}(S_t S_t^\top) \geq 0.98\sigma_r$ after $\mathcal{O}(\log(1/\alpha^2)/(\eta\sigma_r))$ iterations. In the second phase, we leverage the lower bound $\lambda_{\min}(S_t S_t^\top) \geq 0.98\sigma_r$ to further simplify the one-step dynamics in Proposition 16, and show that both signal and cross terms decay linearly, while the residual term remains in the order of α^2 . This phase lasts for $\mathcal{O}(\log(\sigma_1/\alpha^2)/(\eta\sigma_r))$ iterations, and the generalization error can be upper bounded by α^2 at the end of this phase. Finally, in the third phase, we show that the residual term will dominate the signal and cross terms, and the generalization error will decay at a sublinear rate.

Lemma 32 *The conditions of Propositions 16 and 17 are satisfied for every $0 \leq t < \infty$. In particular, for any $0 \leq t < \infty$, we have*

$$\|E_t E_t^\top\| \leq \frac{\alpha^2}{\eta\alpha^2 t + 1}, \quad (30)$$

$$\|S_t S_t^\top\| \leq 1.01\sigma_1, \quad (31)$$

$$S_t S_t^\top \succ 0. \quad (32)$$

The proof of the above lemma can be found in Appendix C.3. Given Lemma 32, we proceed to prove Theorem 18.

Phase 1: Eigenvalue Learning. Due to Proposition 16 and Lemma 32, we have

$$\begin{aligned} \lambda_{\min}(S_{t+1} S_{t+1}^\top) &\geq \left(1 + 2\eta\sigma_r - 2\eta\|E_t E_t^\top\|\right) \lambda_{\min}(S_t S_t^\top) - 2.01\eta\lambda_{\min}^2(S_t S_t^\top) \\ &\geq (1 + 1.99\eta\sigma_r) \lambda_{\min}(S_t S_t^\top) - 2.01\eta\lambda_{\min}^2(S_t S_t^\top), \end{aligned} \quad (33)$$

where we used the assumption $2\|E_t E_t^\top\| \leq 2\alpha^2 \leq 0.01\sigma_r$ due to our choice of α . Now, we consider two cases:

- Suppose that T_1 is the largest iteration such that $\lambda_{\min}(S_t S_t^\top) \leq \sigma_r/2.01$ for every $t \leq T_1$. According to (33), we have

$$\lambda_{\min}(S_t S_t^\top) \geq (1 + 0.99\eta\sigma_r)^t \lambda_{\min}(S_0 S_0^\top) \geq (1 + 0.99\eta\sigma_r)^t \alpha^2 \sigma_r.$$

This implies that, after $\mathcal{O}(\log(1/\alpha^2)/(\eta\sigma_r))$ iterations, we have $\lambda_{\min}(S_t S_t^\top) > \sigma_r/2.01$, and hence, $T_1 = \mathcal{O}(\log(1/\alpha^2)/(\eta\sigma_r))$.

- For $t > T_1$, let $x_t = \sigma_r - \lambda_{\min}(S_t S_t^\top)$. Then, according to (33), we have

$$\begin{aligned} x_{t+1} &\leq (1 - 2.03\eta\sigma_r) x_t + 2.01\eta x_t^2 + 0.02\eta\sigma_r^2 \\ &\leq (1 - 1.02\eta\sigma_r) x_t + 0.02\eta\sigma_r^2, \end{aligned} \quad (34)$$

where in the second inequality, we used the fact that $x_t = \sigma_r - \lambda_{\min}(S_t S_t^\top) \leq 1.01\sigma_r/2.01$. The above inequality implies

$$\begin{aligned} x_{t+1} - 0.0196\sigma_r &\leq (1 - 1.02\eta\sigma_r) (x_t - 0.0196\sigma_r) \\ \implies x_{t+1} - 0.0196\sigma_r &\leq (1 - 1.02\eta\sigma_r)^{t-T_1+1} (x_{T_1} - 0.0196\sigma_r). \end{aligned}$$

Hence, we have $x_t \leq 0.02\sigma_r$ after $T_3 = T_1 + T_2$ iterations, where $T_2 = \mathcal{O}(1/\eta\sigma_r)$, which in turn shows that $\lambda_{\min}(S_t S_t^\top) \geq 0.98\sigma_r$.

The above analysis shows that $\lambda_{\min}(S_t S_t^\top) \geq 0.98\sigma_r$ for every $t \geq T_3 = T_1 + T_2 = \mathcal{O}(\log(1/\alpha^2)/(\eta\sigma_r))$.

Phase 2: Global Convergence. We have $0.98\sigma_r \leq \lambda_{\min}(S_t S_t^\top) \leq 1.01\sigma_1$ for every $t \geq T_3$. This, combined with the one-step signal dynamic (17), implies that

$$\left\| \Sigma - S_{t+1} S_{t+1}^\top \right\| \leq (1 - 0.98\eta\sigma_r) \left\| \Sigma - S_t S_t^\top \right\| + 5\eta \left\| S_t E_t^\top \right\|^2.$$

On the other hand, due to Lemma 32, we have

$$\left\| S_t E_t^\top \right\|^2 \leq \left\| S_t S_t^\top \right\| \left\| E_t E_t^\top \right\| \leq (1.01\sigma_1)\alpha^2.$$

This implies that

$$\begin{aligned} \left\| \Sigma - S_{t+1} S_{t+1}^\top \right\| &\leq (1 - 0.98\eta\sigma_r) \left\| \Sigma - S_t S_t^\top \right\| + 6\eta\sigma_1\alpha^2 \\ \implies \left\| \Sigma - S_{t+1} S_{t+1}^\top \right\| - \frac{6\sigma_1\alpha^2}{0.98\sigma_r} &\leq (1 - 0.98\eta\sigma_r) \left(\left\| \Sigma - S_t S_t^\top \right\| - \frac{6\sigma_1\alpha^2}{0.98\sigma_r} \right) \\ \implies \left\| \Sigma - S_{t+1} S_{t+1}^\top \right\| - \frac{6\sigma_1\alpha^2}{0.98\sigma_r} &\leq (1 - 0.98\eta\sigma_r)^{t-T_3+1} \left(\left\| \Sigma - S_{T_3} S_{T_3}^\top \right\| - \frac{6\sigma_1\alpha^2}{0.98\sigma_r} \right). \end{aligned} \quad (35)$$

Therefore,

$$\left\| \Sigma - S_t S_t^\top \right\| \leq 7\kappa\alpha^2 \quad \text{for } t \geq T_5 = T_3 + T_4, \quad \text{where } T_4 = \mathcal{O}\left(\frac{\log\left(\frac{\sigma_1}{\kappa\alpha^2}\right)}{\eta\sigma_r}\right).$$

Here, we use the inequality $\|\Sigma - S_{T_3} S_{T_3}^\top\| \leq \|\Sigma\| + \|S_{T_3} S_{T_3}^\top\| \leq 2.01\sigma_1$. On the other hand, the one-step dynamics for the cross term (18) implies that

$$\begin{aligned} \left\| S_{t+1} E_{t+1}^\top \right\| &\leq \left(1 - \eta\lambda_{\min}\left(S_t S_t^\top\right) + 2\eta\left\|\Sigma - S_t S_t^\top\right\| + 2\eta\left\|E_t E_t^\top\right\|\right) \left\|S_t E_t^\top\right\| \\ &\leq (1 - 0.5\eta\sigma_r) \left\|S_t E_t^\top\right\| \\ \implies \left\| S_{t+1} E_{t+1}^\top \right\| &\leq (1 - 0.5\eta\sigma_r)^{t-T_5+1} \left\|S_{T_5} E_{T_5}^\top\right\| \\ &\leq (1 - 0.5\eta\sigma_r)^{t-T_5+1} (1.01\alpha\sqrt{\sigma_1}), \end{aligned} \tag{36}$$

where the second inequality follows from the proven upper bound $\|\Sigma - S_t S_t^\top\| \leq 7\kappa\alpha^2$ and Lemma 32. Moreover, the last inequality is due to the fact that

$$\left\| S_{T_5} E_{T_5}^\top \right\| \leq \|S_{T_5}\| \|E_{T_5}\| \leq 1.01\alpha\sqrt{\sigma_1}.$$

The inequality (36) results in

$$\left\| S_t E_t^\top \right\| \leq \alpha^2 \quad \text{for } t \geq T_7 = T_6 + T_5 \quad \text{where } T_6 = \mathcal{O}\left(\frac{\log\left(\frac{\sqrt{\sigma_1}}{\alpha}\right)}{\eta\sigma_r}\right).$$

This upper bound can in turn be used in (35) further to strengthen the upper bound on the signal term as follows

$$\left\| \Sigma - S_t S_t^\top \right\| \leq \alpha^2 \quad \text{for } t \geq T_8 = T_7 + T_6 = \mathcal{O}\left(\frac{\log\left(\frac{\sigma_1}{\alpha^2}\right)}{\eta\sigma_r}\right).$$

Finally, invoking the signal-residual decomposition in Lemma 14, we have

$$\left\| U_t U_t^\top - X^* \right\| \leq \left\| S_t S_t^\top - \Sigma \right\| + 2\left\| S_t E_t^\top \right\| + \left\| E_t E_t^\top \right\| \lesssim \alpha^2$$

for $t \geq T_8 = \mathcal{O}\left(\frac{\log\left(\frac{\sigma_1}{\alpha^2}\right)}{\eta\sigma_r}\right)$.

Phase 3: Sublinear convergence. Once both signal and cross terms are in the order of α^2 , the residual term becomes the dominant term, while both signal and cross terms maintain their linear decay rates. Therefore, we have

$$\left\| U_t U_t^\top - X^* \right\| \lesssim \left\| E_t E_t^\top \right\| \lesssim \frac{\alpha^2}{\eta\alpha^2 t + 1}.$$

This completes the proof. \square

A.2 Proof of Theorem 22

The proof of Theorem 22 follows the same structure as the proof of Theorem 18: first, we use Proposition 19 to show that $\lambda_{\min}(S_t S_t^\top)$ reaches $0.98\sigma_r$ after $T_1 \lesssim \log(1/\alpha)/(\eta\sigma_r\varphi^2)$ iterations. Given this inequality and equipped with the one-step dynamics of the signal, cross, and residual terms (Propositions 20 and 21), we then establish the linear convergence of SubGM to the ground truth. As a first step, we show an important property of the proposed initialization scheme.

Lemma 33 *Suppose that the measurements are noiseless and satisfy $(4r, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP where $\varepsilon = 0$, and $X^* \in \mathcal{S}$. Then, the initial point $U_0 = VS_0 + V_\perp E_0$ generated from Algorithm 2 satisfies*

$$\left\| U_0 U_0^\top - \alpha^2 \varphi(X^*) \frac{X^* X^{*\top}}{\|X^*\|_F} \right\| \leq 2\alpha^2 \varphi(X^*) \delta,$$

The proof can be found in Appendix D.5. An immediate consequence of the above lemma is the following inequality:

$$\left\| S_0 S_0^\top - \alpha^2 \varphi(X^*) \frac{\Sigma}{\|X^*\|_F} \right\| \vee \left\| S_0 E_0^\top \right\| \vee \left\| E_0 E_0^\top \right\| \leq 2\alpha^2 \varphi(X^*) \delta.$$

Given this property of the proposed initialization scheme, we next show that the conditions of Propositions 19, 20, and 21 are satisfied throughout solution path.

Lemma 34 *We either have $\|\Delta_t\|_F \lesssim d\alpha^{2-\mathcal{O}(\sqrt{r}\kappa^2\delta)}$ for some $0 \leq t \leq T_{\text{end}}$, or the conditions of Propositions 19, 20, and 21 are satisfied for every $0 \leq t \leq T_{\text{end}}$. In particular, for any $0 \leq t \leq T_{\text{end}}$, we have*

$$\|F_t\| \leq \eta\bar{\varphi}^2 \left(100\sqrt{r}\sigma_1^{1.5}\delta + 30\sigma_1\sqrt{\alpha\bar{\varphi}\delta} \right) (t+1), \quad (37)$$

$$\|G_t\| \leq \alpha^{1-\mathcal{O}(\sqrt{r}\kappa^2\delta)} \sqrt{\bar{\varphi}\delta}, \quad (38)$$

$$\|\Delta_t\| \leq 5\sigma_1, \quad (39)$$

$$\left\| S_t S_t^\top \right\| \leq 1.01\sigma_1, \quad (40)$$

$$S_t S_t^\top \succ 0, \quad (41)$$

$$\left\| E_t S_t^\top \left(S_t S_t^\top \right)^{-1} \right\| \leq 1/3. \quad (42)$$

The proof of Lemma 34 is provided in Appendix D.6. Note that the inequality $\|\Delta_t\|_F \lesssim d\alpha^{2-\mathcal{O}(\sqrt{r}\kappa^2\delta)}$ for some $0 \leq t \leq T_{\text{end}}$ readily implies the final result. On the other hand, if $\|\Delta_t\|_F \gtrsim d\alpha^{2-\mathcal{O}(\sqrt{r}\kappa^2\delta)}$, Lemma 34 implies that Propositions 19, 20, and 21 hold for every $0 \leq t \leq T_{\text{end}}$.

Phase 1: Eigenvalue Learning. Based on Lemma 34, the conditions of Proposition 19 are satisfied for $T_1 \lesssim \log(1/\alpha)/(\eta\sigma_r\varphi^2)$, and we have

$$\begin{aligned} \lambda_{\min} \left(S_{t+1} S_{t+1}^\top \right) &\geq \left((1 + 2\bar{\eta}_t \sigma_r) - 2\bar{\eta}_t \left\| E_t E_t^\top \right\| - 72\bar{\eta}_t \delta \|\Delta_t\|_F \right) \lambda_{\min} \left(S_t S_t^\top \right) \\ &\quad - 2\bar{\eta}_t (1 + \bar{\eta}_t \sigma_r) \lambda_{\min} \left(S_t S_t^\top \right)^2. \end{aligned} \quad (43)$$

Due to (37) and (38), we have $\|E_t E_t^\top\| \leq \|F_t\|^2 + \|G_t\|^2 \leq 0.005\sigma_r$ for every $t \lesssim \log(1/\alpha)/(\eta\sigma_r\varphi^2)$, where we used Lemma 34 and the assumed upper bounds on δ and α . Moreover, $72\delta\|\Delta_t\|_F \leq 150\sigma_1\delta \leq 0.005\sigma_r$, where again we used the assumed upper bound on δ and Lemma 34. These two inequalities together with (43) lead to

$$\lambda_{\min}\left(S_{t+1}S_{t+1}^\top\right) \geq (1 + 1.99\bar{\eta}_t\sigma_r)\lambda_{\min}\left(S_tS_t^\top\right) - 2.01\bar{\eta}_t\lambda_{\min}^2\left(S_tS_t^\top\right).$$

The above inequality is identical to (33), after noticing that $\eta\underline{\varphi}^2 \leq \bar{\eta}_t \leq \eta\bar{\varphi}^2$. On the other hand, Lemma 33 shows that $\lambda_{\min}(S_0S_0^\top) \geq \alpha^2\varphi(X^*)\left(\frac{\sigma_r}{\|X^*\|_F} - 2\delta\right) \geq \alpha^2\underline{\varphi}\frac{1}{2\sqrt{r\kappa}}$. Therefore, using an argument analogous to the proof of Theorem 18, we have $\lambda_{\min}(S_tS_t^\top) \geq 0.98\sigma_r$ for $t \geq T_1 = \mathcal{O}(\log(1/\alpha)/(\eta\sigma_r\varphi^2))$. The details are omitted for brevity.

Phase 2: Global convergence. Recall the signal-residual decomposition

$$\left\|U_tU_t^\top - X^*\right\| \leq \left\|\Sigma - S_tS_t^\top\right\| + 2\left\|S_tE_t^\top\right\| + \left\|F_tF_t^\top\right\| + \left\|G_tG_t^\top\right\|.$$

In what follows, we show that once $\lambda_{\min}(S_tS_t^\top) \geq 0.98\sigma_r$, all terms in the above inequality decay at a linear rate, except for $\|G_tG_t^\top\|$. To this goal, first note that Propositions 20 and 21 together with $0.98\sigma_r \leq \lambda_{\min}(S_tS_t^\top) \leq 1.01\sigma_1$ lead to the following one-step dynamics:

$$\left\|\Sigma - S_{t+1}S_{t+1}^\top\right\| \leq (1 - 0.98\bar{\eta}_t\sigma_r)\left\|\Sigma - S_tS_t^\top\right\| + 5\bar{\eta}_t\left\|S_tE_t^\top\right\|^2 + 37\bar{\eta}_t\delta\bar{\varphi}^2\sigma_1\|\Delta_t\|_F, \quad (44)$$

$$\left\|S_{t+1}E_{t+1}^\top\right\| \leq \left(1 - 0.98\bar{\eta}_t\sigma_r + 2\bar{\eta}_t\left\|\Sigma - S_tS_t^\top\right\| + 2\bar{\eta}_t\|E_tE_t\|\right)\left\|S_tE_t^\top\right\| + 22\bar{\eta}_t\delta\sigma_1\|\Delta_t\|_F, \quad (45)$$

$$\|F_{t+1}\| \leq (1 - 0.98\bar{\eta}_t\sigma_r)\|F_t\| + 5\sqrt{\sigma_1}\bar{\eta}_t\delta\|\Delta_t\|_F + 6\bar{\eta}_t\|\Delta_t\|\|G_t\|. \quad (46)$$

Note that, unlike $\|\Sigma - S_tS_t^\top\|$ and $\|F_t\|$, the cross term $\|S_tE_t^\top\|$ enjoys linear decay only under the condition $\|\Sigma - S_tS_t^\top\| < 0.98\sigma_r$, which is not necessarily satisfied in the eigenvalue learning phase. Our next lemma shows that this condition is satisfied shortly after the eigenvalue learning phase.

Lemma 35 *We have $\|\Sigma - S_tS_t^\top\| < 0.03\sigma_r$, for every $T_{\text{end}} \geq t \geq T_1 + T_2$, where $T_2 = \mathcal{O}(\log(\kappa)/(\eta\sigma_r\varphi^2))$.*

Proof It is easy to see that

$$5\bar{\eta}_t\left\|S_tE_t^\top\right\|^2 \leq 5.05\sigma_1\bar{\eta}_t\|E_t\|^2 \leq 5.05\bar{\eta}_t\sigma_1(\|F_t\| + \|G_t\|)^2 \leq 0.01\bar{\eta}_t\sigma_r^2, \quad (47)$$

where the last inequality is due to our choice of δ and Lemma 34. Similarly, we can show that $37\bar{\eta}_t\delta\sigma_1\|\Delta_t\|_F \leq 0.01\bar{\eta}_t\sigma_r^2$. These two inequalities combined with (44) lead to

$$\left\|\Sigma - S_{t+1}S_{t+1}^\top\right\| \leq (1 - 0.98\bar{\eta}_t\sigma_r)\left\|\Sigma - S_tS_t^\top\right\| + 0.02\bar{\eta}_t\sigma_r^2.$$

This implies that $\|\Sigma - S_tS_t^\top\| \leq 0.03\sigma_r$ after $\mathcal{O}(\log(\kappa)/(\eta\sigma_r\varphi^2))$ iterations. \blacksquare

The above lemma shows that the one-step dynamic of the cross term can be simplified as

$$\left\| S_{t+1} E_{t+1}^\top \right\| \leq (1 - 0.49\bar{\eta}_t \sigma_r) \left\| S_t E_t^\top \right\| + 22\bar{\eta}_t \delta^2 \sigma_1 \|\Delta_t\|_F. \quad (48)$$

Moreover, recall that

$$\begin{aligned} \|G_{t+1}\| &\leq \left(1 + \bar{\eta}_t^2 \left(2 \left\| E_t S_t^\top \right\|^2 + \|E_t\|^4 + 2 \|\Delta_t\| \left\| E_t S_t^\top \right\| \right) + 7\bar{\eta}_t \delta \|\Delta_t\|_F \right) \|G_t\| \\ &\leq \left(1 + 5\bar{\eta}_t^2 \|\Delta_t\|^2 + 7\bar{\eta}_t \delta \|\Delta_t\|_F \right) \|G_t\|. \end{aligned} \quad (49)$$

Here we use the fact that $\|E_t S_t^\top\| \vee \|E_t\|^2 \leq \|\Delta_t\|$. Now, let us define $\gamma_t := \|\Sigma - S_t S_t^\top\| + 2\|S_t E_t^\top\| + \|F_t\|^2 + \|G_t\|^2$ as an upper bound for the generalization error $\|U_t U_t^\top - X^*\|$. Combining (44), (46), (48), and (49), we have

$$\begin{aligned} \gamma_{t+1} &= \left\| \Sigma - S_{t+1} S_{t+1}^\top \right\| + 2 \left\| S_{t+1} E_{t+1}^\top \right\| + \|F_{t+1}\|^2 + \|G_{t+1}\|^2 \\ &\leq (1 - C_1 \bar{\eta}_t \sigma_r) \gamma_t + C_2 \bar{\eta}_t \delta \sigma_1 \|\Delta_t\|_F + C_3 \bar{\eta}_t \sigma_r \|G_t\|^2 \\ &\leq (1 - C_1 \bar{\eta}_t \sigma_r + C_2 \bar{\eta}_t \sqrt{r} \sigma_1 \delta) \gamma_t + C_4 \bar{\eta}_t \sigma_r \sqrt{d} \|G_t\|^2 \\ &\leq (1 - C_5 \bar{\eta}_t \sigma_r) \gamma_t + C_4 \bar{\eta}_t \sigma_r \sqrt{d} \|G_t\|^2, \end{aligned}$$

for some universal constants $C_1, C_2, C_3, C_4, C_5 > 0$. This implies that

$$\begin{aligned} \gamma_{t+1} - C_6 \sqrt{d} \|G_t\|^2 &\leq (1 - C_5 \bar{\eta}_t \sigma_r) \left(\gamma_t - C_6 \sqrt{d} \|G_t\|^2 \right) \\ \implies \gamma_{t+1} - C_6 \sqrt{d} \|G_t\|^2 &\leq (1 - C_5 \bar{\eta}_t \sigma_r)^{t - (T_1 + T_2)} (\gamma_{T_1 + T_2} - C_6 \sqrt{d} \|G_t\|^2). \end{aligned}$$

Note that $\gamma_{T_1 + T_2} \leq 0.1\sigma_r$, according to Lemma 35, inequality (47), and the upper bounds on $\|F_t\|^2$ and $\|G_t\|^2$. On the other hand, $\|G_t\| \leq \alpha^{1 - \mathcal{O}(\sqrt{r}\kappa^2\delta)} \sqrt{\bar{\varphi}\delta}$, according to Lemma 34. Therefore, after $T_1 + T_2 + T_3$ iterations with $T_3 = \mathcal{O}(\log(1/\alpha)/(\eta\sigma_r\varphi^2))$, we have $\|\Delta_t\|_F \leq \sqrt{d}\gamma_t \lesssim d\alpha^{2 - \mathcal{O}(\sqrt{r}\kappa^2\delta)} \bar{\varphi}\delta$. This completes the proof. \square

A.3 Proof of Theorem 27

Without loss of generality, we assume that $\|\Delta_t\|_F \geq \zeta$ for every $0 \leq t \leq T_{end}$; otherwise, the final bound for the generalization error holds, and the proof is complete. Before delving into the details, we first provide a general overview of our approach. The proof of Theorem 27 is similar to that of Theorem 22 with a key difference that we divide our analysis into two parts depending on the value of $\|\Delta_t\|$: if $\eta\rho^t/\|\Delta_t\| \lesssim 1/\sigma_1$, then $\|\Delta_t\|$ decays slower than $\sigma_1\eta\rho^t$. Under this assumption, we will use the one-step dynamics of the signal, cross, and residual terms in Propositions 24, 25, and 26 to prove that $\|\Delta_t\|$ decays exponentially fast. Alternatively, $\eta\rho^t/\|\Delta_t\| \gtrsim 1/\sigma_1$ implies that $\|\Delta_t\| \lesssim \sigma_1\eta\rho^t$, which readily establishes the exponential decay of the generalization error. Indeed, the above two cases may occur alternatively, which requires a more delicate analysis.

Similar to the proof of Theorem 22, we show that SubGM undergoes two phases: (1) eigenvalue learning phase, and (2) global convergence phase. In the first phase, we show

that $\lambda_{\min}(S_t S_t^\top)$ and $\|\Delta_t\|$ converge to $0.98\sigma_r$ and $0.02\sigma_r$, respectively. The main difference between the proofs of Theorems 22 and 27 is the fact that the one-step dynamics of the signal, cross, and residual terms may not hold during the entire solution trajectory. However, our next lemma shows that they indeed hold in the first phase of our analysis.

Lemma 36 *Suppose that $\|\Delta_t\| \geq 0.02\sigma_r$ for every $0 \leq t \leq \bar{T} \leq T_{\text{end}}$. Then, the assumptions of Propositions 24, 25, and 26 are satisfied for every $0 \leq t \leq \bar{T}$. In particular, for any $0 \leq t \leq \bar{T}$, we have*

$$\|F_t\| \leq \left(15\sqrt{r}\eta\rho^t\sqrt{\sigma_1}\delta + 12\sqrt{2}\eta\sqrt{\bar{\varphi}\alpha\delta}\right)(t+1), \quad (50)$$

$$\|G_t\| \leq 2\sqrt{2}\alpha^{1-\mathcal{O}(\sqrt{r}\kappa\delta)}\sqrt{\bar{\varphi}\delta}, \quad (51)$$

$$\|\Delta_t\| \leq 5\sigma_1, \quad (52)$$

$$\left\|S_t S_t^\top\right\| \leq 1.1\sigma_1, \quad (53)$$

$$S_t S_t^\top \succ 0, \quad (54)$$

$$\left\|E_t S_t^\top \left(S_t S_t^\top\right)^{-1}\right\| \leq 1/3. \quad (55)$$

Moreover, for any $t \geq \bar{T}$, we also have

$$\|G_t\| \leq 2\sqrt{2}\alpha^{1-\mathcal{O}(\sqrt{r}\kappa\delta)}\sqrt{\bar{\varphi}\delta}. \quad (56)$$

The proof of Lemma 36 is analogous to that of Lemma 34, and can be found in Appendix E.5. Equipped with this lemma, we provide proof for the eigenvalue learning phase.

Phase 1: Eigenvalue Learning. We will show that $\lambda_{\min}(S_t S_t^\top) \geq 0.98\sigma_r$ within $t \leq T_1 = \mathcal{O}((\kappa/\eta) \log(1/\alpha))$ iterations. After that, we prove that we only need $T_2 = \mathcal{O}((\kappa/\eta) \log(\kappa))$ additional iterations to ensure that $\|\Delta_t\| \leq 0.02\sigma_r$; this marks the end of Phase 1. Without loss of generality, we may assume that $\|\Delta_t\| \geq 0.02\sigma_r$ for every $t \leq T_1$. To see this, suppose that $\|\Delta_t\| \leq 0.02\sigma_r$ for some $t \leq T_1$. This implies that $\|\Sigma - S_t S_t^\top\| \leq \|\Delta_t\| \leq 0.02\sigma_r$, which in turn leads to $\lambda_{\min}(S_t S_t^\top) \geq 0.98\sigma_r$. On the other hand, $\|\Delta_t\| \geq 0.02\sigma_r$ together with Lemma 36 implies that the one-step dynamic in Proposition 24 holds and we have

$$\begin{aligned} & \lambda_{\min}\left(S_{t+1} S_{t+1}^\top\right) \\ & \geq \left(\left(1 + \frac{\eta\rho^t}{\|\Delta_t\|}\sigma_r\right)^2 - \frac{2\eta\rho^t}{\|\Delta_t\|}\|E_t E_t^\top\| - 384\sqrt{r}\eta\rho^t\delta\right)\lambda_{\min}\left(S_t S_t^\top\right) \\ & \quad - 2\frac{\eta\rho^t}{\|\Delta_t\|}\left(1 + \frac{\eta\rho^t}{\|\Delta_t\|}\sigma_r\right)\lambda_{\min}\left(S_t S_t^\top\right)^2 \\ & \geq \left(1 + \frac{2\eta\rho^t}{\|\Delta_t\|}(\sigma_r - \mathcal{O}(\sqrt{r}\sigma_1\delta))\right)\lambda_{\min}\left(S_t S_t^\top\right) - \frac{2.01\eta\rho^t}{\|\Delta_t\|}\lambda_{\min}\left(S_t S_t^\top\right)^2 \\ & \geq \left(1 + \frac{1.99\eta\rho^t\sigma_r}{\|\Delta_t\|}\right)\lambda_{\min}\left(S_t S_t^\top\right) - \frac{2.01\eta\rho^t}{\|\Delta_t\|}\lambda_{\min}\left(S_t S_t^\top\right)^2 \end{aligned} \quad (57)$$

where the second inequality follows from $\|\Delta_t\| \geq 0.02\sigma_r$ and $\eta \lesssim 1$, and the last inequality follows from $\delta \lesssim 1/(\sqrt{r}\sigma_1)$. The rest of the proof for the eigenvalue learning phase is

similar to the arguments made after (33) in the proof of Theorem 18. Suppose that T'_1 is the largest iteration such that $\lambda_{\min}(S_t S_t^\top) \leq \sigma_r/2.01$ for every $t \leq T'_1$. We show that $T'_1 = \mathcal{O}((\kappa/\eta) \log(1/\alpha))$. To see this, note that for every $t \leq T'_1$, the above inequality can be simplified as

$$\lambda_{\min}(S_t S_t^\top) \geq \left(1 + \frac{0.99\eta\rho^t\sigma_r}{\|\Delta_t\|}\right) \lambda_{\min}(S_t S_t^\top) \geq \lambda_{\min}(S_0 S_0^\top) \prod_{s=0}^t \left(1 + \frac{\eta\rho^s}{6\kappa}\right), \quad (58)$$

where we used the assumption $\delta \lesssim 1/(\sqrt{r}\kappa)$ and $\|\Delta_t\| \leq 5\sigma_1$. To proceed with the proof, we need the following technical lemma.

Lemma 37 *For any $\alpha > 0, 0 < \rho < 1$ and $T \in \mathbb{N}_+$, we have*

$$\exp\left(\frac{\alpha T \rho^T}{1 + \alpha}\right) \leq \prod_{t=0}^T (1 + \alpha\rho^t) \leq \exp\left(\frac{\alpha}{1 - \rho}\right).$$

The proof of Lemma 37 can be found in Appendix F.5. Lemma 37 together with Lemma 33 and (58) leads to

$$\begin{aligned} \lambda_{\min}(S_t S_t^\top) &\geq \exp\left(\frac{\eta t \rho^t}{6\kappa + \eta}\right) \alpha^2 \varphi(X^*) \left(\frac{\sigma_r}{\|X^*\|_F} - \delta\right) \\ &\geq \exp\left(\frac{\eta t \rho^t}{6\kappa + \eta}\right) \alpha^2 \varphi(X^*) \left(\frac{\sigma_r}{\sqrt{r}\sigma_1} - \delta\right) \\ &\geq \exp\left(\frac{\eta t \rho^t}{6\kappa + \eta}\right) \frac{\alpha^2 \varphi(X^*)}{2\sqrt{r}\kappa}. \end{aligned}$$

Due to our assumption $\rho = 1 - \Theta(\eta/(\kappa \log(1/\alpha)))$, we have $\rho^t \geq 1 - \mathcal{O}(\eta t/(\kappa \log(1/\alpha)))$. This together with the assumption $t \leq \mathcal{O}((\kappa/\eta) \log(1/\alpha))$ leads to $\rho^t \geq 1 - \mathcal{O}(1) \geq \Omega(1)$. Therefore, one can write

$$\lambda_{\min}(S_t S_t^\top) \geq \exp\left(\Omega(1) \frac{\eta t}{\kappa}\right) \frac{\alpha^2 \varphi(X^*)}{2\sqrt{r}\kappa}.$$

Given the above equation, it is easy to verify that after $\mathcal{O}((\kappa/\eta) \log(1/\alpha))$ iterations, we have $\lambda_{\min}(S_t S_t^\top) \geq \sigma_r/2.01$. This implies that $T'_1 = \mathcal{O}((\kappa/\eta) \log(1/\alpha))$.

For $t > T'_1$, define $x_t = \sigma_r - \lambda_{\min}(S_t S_t^\top)$. Then, arguments analogous to the proof of Theorem 18 can be used to write

$$\begin{aligned} x_{t+1} - 0.0196\sigma_r &\leq \left(1 - 1.02 \frac{\eta\rho^t\sigma_r}{\|\Delta_t\|}\right) (x_t - 0.0196\sigma_r) \\ &\leq \left(1 - \frac{\eta\rho^t}{5\kappa}\right) (x_t - 0.0196\sigma_r) \\ &\leq (x_{T'_1} - 0.0196\sigma_r) \prod_{s=T'_1}^t \left(1 - \frac{\eta\rho^s}{5\kappa}\right). \end{aligned}$$

Recall that $\rho^t = \Omega(1)$ for every $t \leq \mathcal{O}((\kappa/\eta) \log(1/\alpha))$. Therefore, for every $T_1' < t \leq \mathcal{O}((\kappa/\eta) \log(1/\alpha))$, we have

$$x_{t+1} - 0.0196\sigma_r \leq (x_{T_1'} - 0.0196\sigma_r) \left(1 - \Omega(1) \frac{\eta\sigma_r}{\kappa}\right)^{t-T_1'+1}.$$

This in turn implies that after additional $T_1'' = \mathcal{O}(\kappa/\eta)$ iterations, we have $x_{t+1} \leq 0.02\sigma_r$. Therefore, we have $\lambda_{\min}(S_t S_t^\top) \geq 0.98\sigma_r$ after $T_1 = T_1' + T_1'' = \mathcal{O}((\kappa/\eta) \log(1/\alpha))$ iterations. Now, it suffices to show that, after additional $T_2 = \mathcal{O}((\kappa/\eta) \log(\kappa))$ iterations, we have $\|\Delta_t\| \leq 0.02\sigma_r$. To this goal, suppose that $\|\Delta_t\| \geq 0.02\sigma_r$ for every $T_1 \leq t \leq T_1 + T_2$. Recalling the signal-residual decomposition (21), one can write

$$\begin{aligned} \|\Delta_t\| &\leq \left\| \Sigma - S_t S_t^\top \right\| + 2 \left\| S_t E_t^\top \right\| + \|F_t\|^2 + \|G_t\|^2 \\ &\leq \left\| \Sigma - S_t S_t^\top \right\| + 0.01\sigma_r, \end{aligned}$$

where the second inequality follows from Lemma 36. Given the above inequality, $\|\Delta_t\| \geq 0.02\sigma_r$ implies that $\left\| \Sigma - S_t S_t^\top \right\| \geq 0.01\sigma_r$ for every $T_1 \leq t \leq T_1 + T_2$. Combined with the one-step dynamic for the signal term, we have

$$\begin{aligned} \left\| \Sigma - S_{t+1} S_{t+1}^\top \right\| &\leq \left(1 - \frac{\eta\rho^t}{\|\Delta_t\|} \lambda_{\min}(S_t S_t^\top)\right) \left\| \Sigma - S_t S_t^\top \right\| + 5 \frac{\eta\rho^t}{\|\Delta_t\|} \left\| S_t E_t^\top \right\|^2 \\ &\quad + 193\sqrt{r}\eta\rho^t\delta\sigma_1 \\ &\stackrel{(a)}{\leq} \left(1 - \frac{0.98\eta\rho^t}{\|\Delta_t\|} \sigma_r\right) \left\| \Sigma - S_t S_t^\top \right\| + \mathcal{O}(\sqrt{r}\eta\rho^t\delta\sigma_1) \\ &\leq \left\| \Sigma - S_t S_t^\top \right\| - 0.98\eta\rho^t\sigma_r \frac{\left\| \Sigma - S_t S_t^\top \right\|}{\left\| \Sigma - S_t S_t^\top \right\| + 0.01\sigma_r} \\ &\quad + \mathcal{O}(\sqrt{r}\eta\rho^t\delta\sigma_1) \\ &\leq \left\| \Sigma - S_t S_t^\top \right\| - 0.49\eta\rho^t\sigma_r + \mathcal{O}(\sqrt{r}\eta\rho^t\delta\sigma_1) \\ &\leq \left\| \Sigma - S_t S_t^\top \right\| - \Omega(\eta\rho^t\sigma_r), \end{aligned} \tag{59}$$

where in (a) we used the fact that $\|S_t E_t^\top\| \lesssim \sqrt{\sigma_1} \|F_t\| \lesssim \sigma_1 \sqrt{r} \eta \rho^t \delta t$, $\delta \lesssim 1/(\sqrt{r} \kappa^2 \bar{\varphi}^4 \log^2(1/\alpha))$, and $t \lesssim (\kappa/\eta) \log(1/\alpha)$. The above inequality leads to

$$\left\| \Sigma - S_t S_t^\top \right\| \leq \left\| \Sigma - S_{T_1} S_{T_1}^\top \right\| - \Omega \left(\sum_{s=T_1}^{t-1} \eta\rho^s\sigma_r \right).$$

Therefore, after $T_3 = T_1 + \mathcal{O}((\kappa/\eta) \log \kappa)$ iterations, we have $\left\| \Sigma - S_t S_t^\top \right\| \leq 0.01\sigma_r$. This completes the proof of the first phase.

Phase 2: Global Convergence. In the second phase, we show that, once $\|\Delta_t\| \leq 0.02\sigma_r$, $\|\Delta_t\|$ starts to decay linearly until it is dominated by $\sqrt{d} \|G_t\|^2$. Similar to before, we define $\gamma_t = \left\| \Sigma - S_t S_t^\top \right\| + 2 \left\| S_t E_t^\top \right\| + \|F_t\|^2 + \|G_t\|^2$. Our next lemma plays a central role in our subsequent arguments:

Lemma 38 *Suppose that $T_3 \leq t \leq T_{end}$ is chosen such that $\gamma_s \leq 0.1\sigma_r$ and $\|\Delta_s\| \geq \sqrt{d}\|G_s\|^2$, for every $T_3 \leq s \leq t$. Then, we have*

$$- \gamma_{t+1} \leq 0.1\sigma_r.$$

Moreover, at least one of the following statements is satisfied:

$$\begin{aligned} - \|\Delta_{t+1}\| &\geq \sqrt{d}\|G_{t+1}\|^2, \\ - \|\Delta_{t+1}\| &\lesssim \sqrt{d}\alpha^{2-\mathcal{O}(\sqrt{r\kappa\delta})}. \end{aligned}$$

To streamline the presentation, we defer the proof of the above lemma to Appendix F.6. According to the above lemma, we may assume that $\gamma_t \leq 0.1\sigma_r$ and $\|\Delta_t\| \geq \sqrt{d}\|G_t\|^2$ for all iterations $T_3 \leq t \leq T_{end}$; otherwise, we have $\|\Delta_t\| \lesssim \sqrt{d}\alpha^{2-\mathcal{O}(\sqrt{r\kappa\delta})}$ for some $T_3 \leq t \leq T_{end}$, which readily completes the proof. On the other hand, the assumptions $\gamma_t \leq 0.1\sigma_r$ and $\|\Delta_t\| \geq \sqrt{d}\|G_t\|^2$ lead to

$$\lambda_{\min}(S_t S_t^\top) \geq 0.9\sigma_r, \quad \|S_t S_t^\top\| \leq 1.1\sigma_1,$$

and

$$\|E_t E_t^\top\| \leq 0.1\sigma_r, \quad \left\| E_t S_t^\top (S_t S_t^\top)^{-1} \right\| \leq 0.2. \quad (60)$$

Together with our analysis in Phase 1, this implies that the one-step dynamic of G_t holds for every $0 \leq t \leq T_{end}$, and we have $\|G_t\| \leq 2\sqrt{2}\alpha^{1-\mathcal{O}(\sqrt{r\kappa\delta})}\sqrt{\bar{\varphi}\delta}$, for every $0 \leq t \leq T_{end}$.

Under the assumption $\|\Delta_t\| \geq \sqrt{d}\|G_t\|^2$, our next lemma shows that, if $\|\Delta_{t+T_3}\| \geq 0.02\sigma_r\rho^t$ for some $t \geq 0$, then there exists t' satisfying $t' - t = \mathcal{O}(1/\eta)$ such that $\|\Delta_{t'+T_3}\| \leq 0.02\sigma_r\rho^{t'}$. This in turn ensures that the generalization error decays by a constant factor every $\mathcal{O}(1/\eta)$ iterations until it reaches the same order as $\sqrt{d}\|G_t\|^2$. In particular, we have the following lemma:

Lemma 39 *Suppose that $\|\Delta_t\| \geq \sqrt{d}\|G_t\|^2$ for every $T_3 \leq t \leq T_{end}$. Suppose that $t_0 \geq 1$ satisfies $\|\Delta_{t_0+T_3-1}\| \leq 0.02\sigma_r\rho^{t_0-1}$ and $\|\Delta_{t_0+T_3}\| > 0.02\sigma_r\rho^{t_0}$. Then, after at most $\Delta t = \mathcal{O}(1/\eta)$ iterations, we have $\|\Delta_{t_0+\Delta t+T_3}\| \leq 0.02\sigma_r\rho^{t_0+\Delta t}$.*

The proof of the above lemma is presented in Appendix F.7. We show how Lemma 39 can be used to finish the proof of Theorem 27. Recall that $\rho = 1 - \Theta(\eta/(\kappa \log(1/\alpha)))$. Let us pick $T_4 = \mathcal{O}((\kappa/\eta) \log^2(1/\alpha))$. Simple calculation reveals that $0.02\sigma_r\rho^{T_4} \lesssim \alpha^2$. According to Lemma 39, if $\|\Delta_{T_4+T_3}\| > 0.02\sigma_r\rho^{T_4}$, then there exists $\Delta t = \mathcal{O}(1/\eta)$ such that $\|\Delta_{T_4+T_3+\Delta t}\| \leq 0.02\sigma_r\rho^{T_4+\Delta t} \lesssim \alpha^2$. Combined with Lemma 38, we have $\|\Delta_t\|_F \leq \sqrt{d}\|\Delta_t\| \lesssim d\|G_t\|^2\vee\zeta \lesssim d\alpha^{2-\mathcal{O}(\sqrt{r\kappa\delta})}\vee\zeta$ after at most $T_{end} = T_3+T_4+\Delta t = \mathcal{O}((\kappa/\eta) \log^2(1/\alpha))$ iterations. Last, note that $\|G_t\| \leq 2\sqrt{2}\alpha^{1-\mathcal{O}(\sqrt{r\kappa\delta})}\sqrt{\bar{\varphi}\delta}, \forall t \geq 0$ as shown in Lemma 36. We immediately derive $\|\Delta_t\|_F \leq \sqrt{d}\|\Delta_t\| \lesssim d\|G_t\|^2\vee\zeta \lesssim d\alpha^{2-\mathcal{O}(\sqrt{r\kappa\delta})}\vee\zeta$. This completes the proof of Theorem 27. \square

Appendix B. Proofs of Sign-RIP

B.1 Preliminary

We first provide the preliminary probability tools for proving Theorems 10 and 11.

Definition 40 (Sub-Gaussian random variable) We say a random variable $X \in \mathbb{R}$ with expectation $\mathbb{E}[X] = \mu$ is σ^2 -sub-Gaussian if for all $\lambda \in \mathbb{R}$, we have $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\lambda^2 \sigma^2}{2}}$. Moreover, the sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} := \sup_{p \in \mathbb{N}_+} \{p^{-1/2}(\mathbb{E}[|X|^p])^{1/p}\}$.

According to (Wainwright, 2019), the following statements are equivalent:

- X is σ^2 -sub-Gaussian.
- (Tail bound) For any $t > 0$, we have $\mathbb{P}(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$.
- (Moment bound) We have $\|X\|_{\psi_2} \lesssim \sigma$.

Next, we provide the definitions of the sub-Gaussian process, ξ -net, and the covering number.

Definition 41 (Sub-Gaussian process) A zero mean stochastic process $\{\mathcal{X}_\theta, \theta \in \mathbb{T}\}$ is a σ^2 -sub-Gaussian process with respect to a metric d on a set \mathbb{T} , if for every $\theta, \theta' \in \mathbb{T}$, the random variable $\mathcal{X}_\theta - \mathcal{X}_{\theta'}$ is $(\sigma d(\theta, \theta'))^2$ -sub-Gaussian.

Definition 42 (ξ -net and covering number) A set \mathcal{N} is called an ξ -net on (\mathbb{T}, d) if for every $t \in \mathbb{T}$, there exists $\pi(t) \in \mathcal{N}$ such that $d(t, \pi(t)) \leq \xi$. The covering number $N(\mathbb{T}, d, \xi)$ is defined as the smallest cardinality of an ξ -net for (\mathbb{T}, d) :

$$N(\mathbb{T}, d, \xi) := \inf\{|\mathcal{N}| : \mathcal{N} \text{ is an } \xi\text{-net for } (\mathbb{T}, d)\}.$$

Next, we introduce some additional notations which will be used throughout our arguments. Define the rank- k and ε -approximate rank- k unit balls as:

$$\begin{aligned} \mathbb{S}_k &= \{X \in \mathbb{R}^{d \times d} : \text{rank}(X) \leq k, \|X\|_F = 1\}, \\ \mathbb{S}_{k,\varepsilon} &= \{X \in \mathbb{R}^{d \times d} : \|X\|_F = 1 \text{ and } \exists X' \text{ s.t. } \text{rank}(X') \leq k, \|X - X'\|_F \leq \varepsilon\}. \end{aligned}$$

For simplicity of notation, we use $N_{k,\xi}$ to denote $N(\mathbb{S}_k, \|\cdot\|_F, \xi)$. Moreover, we define $\mathcal{S}_{k,\varepsilon}$ as the restriction of the set \mathcal{S} to the set of ε -approximate rank- k matrices, i.e., $\mathcal{S}_{k,\varepsilon} = \{X : X \in \mathcal{S}, X \text{ is } \varepsilon\text{-approximate rank-}k\}$. The following lemma characterizes the covering number of the set of low-rank matrices with the unit norm.

Lemma 43 (Li et al. (2020a)) We have $N_{k,\xi} = N(\mathbb{S}_k, \|\cdot\|_F, \xi) \leq \left(\frac{9}{\xi}\right)^{(2d+1)k}$.

The following well-known result characterizes a concentration bound on the supremum of a sub-Gaussian process.

Theorem 44 (Corollary 5.25 and Theorem 5.29 in (Van Handel, 2014)) Let $\{X_t\}_{t \in \mathbb{T}}$ be a separable sub-Gaussian process on (\mathbb{T}, d) . Then, the following statements hold:

- We have

$$\mathbb{E} \left[\sup_{t \in \mathbb{T}} X_t \right] \leq 12 \int_0^\infty \sqrt{\log N(\mathbb{T}, d, \xi)} d\xi.$$

- For all $t_0 \in \mathbb{T}$ and $x \geq 0$, we have

$$\mathbb{P} \left(\sup_{t \in \mathbb{T}} \{X_t - X_{t_0}\} \geq C \int_0^\infty \sqrt{\log N(\mathbb{T}, d, \xi)} d\xi + x \right) \leq C e^{-x^2/C \text{diam}(\mathbb{T})^2},$$

where $C < \infty$ is a universal constant, and $\text{diam}(\mathbb{T}) = \sup_{t, t' \in \mathbb{T}} \{d(t, t')\}$ is the diameter of \mathbb{T} .

Equipped with these preliminary results, we proceed with the proof of Theorem 10.

B.2 Proof of Theorem 10

To provide the proof of Theorem 10, we first define the following stochastic process:

$$\mathcal{H}_{X,Y} = \sup \left\{ \frac{1}{m} \sum_{i=1}^m \text{Sign}(\langle A_i, X \rangle - s_i) \langle A_i, Y \rangle - \varphi(X) \left\langle \frac{X}{\|X\|_F}, Y \right\rangle \right\},$$

where $\varphi(X) := \sqrt{\frac{2}{\pi}} \left(1 - p + p \mathbb{E} \left[e^{-s^2/(2\|X\|_F^2)} \right] \right)$, and the supremum is taken over the set-valued function $\text{Sign}(\cdot)$. Moreover, to streamline the presentation and whenever there is no ambiguity, we drop the supremum when it is taken with respect to the set-valued function $\text{Sign}(\cdot)$. Our next lemma provides a sufficient condition for Sign-RIP.

Lemma 45 *Sign-RIP holds with parameters delineated in Theorem 10, if*

$$\sup_{X \in \mathcal{S}_{k,\varepsilon}, Y \in \mathbb{S}_{k,\varepsilon/\zeta}} \mathcal{H}_{X,Y} \lesssim \inf_{X \in \mathcal{S}_{k,\varepsilon}} \varphi(X) \delta. \quad (61)$$

Proof According to the definition, Sign-RIP is satisfied if, for every $X, Y \in \mathcal{S}_{k,\varepsilon}$ and $Q \in \mathcal{Q}(X)$, we have

$$\left\langle Q - \varphi(X) \frac{X}{\|X\|_F}, \frac{Y}{\|Y\|_F} \right\rangle \leq \varphi(X) \delta. \quad (62)$$

Recall that $\mathcal{S}_{k,\varepsilon} = \{X : \zeta \leq \|X\|_F \leq R, X \text{ is } \varepsilon\text{-approximate rank-}k\}$. This implies that $Y \in \mathcal{S}_{k,\varepsilon}$ if $Y/\|Y\|_F \in \mathbb{S}_{k,\varepsilon/\zeta}$. Hence, it suffices to restrict $Y \in \mathbb{S}_{k,\varepsilon/\zeta}$. Therefore, Sign-RIP is satisfied if

$$\mathcal{H}_{X,Y} \leq \varphi(X) \delta, \quad \forall X \in \mathcal{S}_{k,\varepsilon}, Y \in \mathbb{S}_{k,\varepsilon/\zeta}.$$

Hence, to guarantee Sign-RIP, it suffices to have

$$\sup_{X \in \mathcal{S}_{k,\varepsilon}, Y \in \mathbb{S}_{k,\varepsilon/\zeta}} \mathcal{H}_{X,Y} \leq \inf_{X \in \mathcal{S}_{k,\varepsilon}} \varphi(X) \delta. \quad \blacksquare$$

Relying on the above lemma, we instead focus on analyzing the stochastic process $\{\mathcal{H}_{X,Y}\}_{X \in \mathcal{S}_{k,\varepsilon}, Y \in \mathbb{S}_{k,\varepsilon/\zeta}}$. As a first step towards this goal, we show that the scaling function can be used to characterize $\mathbb{E}[\langle Q, Y \rangle]$, for every $Q \in \mathcal{Q}(X)$.

Lemma 46 *Suppose that the matrix A has i.i.d. standard Gaussian entries and the noise satisfies Assumption 8. Then, for every $Q \in \mathcal{Q}(X)$, we have*

$$\mathbb{E}[\langle Q, Y \rangle] = \sqrt{\frac{2}{\pi}} \left(1 - p + p \mathbb{E} \left[e^{-s^2/(2\|X\|_F)} \right] \right) \left\langle \frac{X}{\|X\|_F}, Y \right\rangle.$$

Proof Without loss of generality, we assume $\|X\|_F = \|Y\|_F = 1$. Let us denote $u := \langle A, X \rangle$, $v := \langle A, Y \rangle$, $\rho := \text{Cov}(u, v) = \langle X, Y \rangle$. Then, we have

$$\begin{aligned} & \mathbb{E}[\text{Sign}(\langle A, X \rangle - s) \langle A, Y \rangle | s \neq 0] \\ &= \mathbb{E}[\text{Sign}(u - s) v | s \neq 0] \\ &\stackrel{(a)}{=} \rho \mathbb{E}[\text{Sign}(u - s) u | s \neq 0] \\ &= \rho \mathbb{E}_{s \sim \mathbb{P}} \left[\int_s^\infty u \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du - \int_{-\infty}^s u \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \right] \\ &= \rho \mathbb{E}_{s \sim \mathbb{P}} \left[\int_s^\infty u \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du + \int_{-s}^\infty u \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \right] \\ &= 2\rho \mathbb{E}_{s \sim \mathbb{P}} \left[\int_{|s|}^\infty u \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \right] \\ &= \sqrt{\frac{2}{\pi}} \mathbb{E}_{s \sim \mathbb{P}} \left[\int_{|s|}^\infty d(-e^{-u^2/2}) \right] \langle X, Y \rangle \\ &= \sqrt{\frac{2}{\pi}} \mathbb{E}_{s \sim \mathbb{P}} \left[e^{-s^2/2} \right] \langle X, Y \rangle, \end{aligned}$$

where, in (a), we used the fact that $v|u, s \sim \mathcal{N}(\rho u, 1 - \rho^2)$ since s is independent of u, v . Similarly, one can show that $\mathbb{E}[\text{Sign}(\langle A, X \rangle) \langle A, Y \rangle] = \sqrt{\frac{2}{\pi}} \langle X, Y \rangle$. The proof is completed by noting that

$$\begin{aligned} & \mathbb{E}[\text{Sign}(s + \langle A, X \rangle) \langle A, Y \rangle] \\ &= p \mathbb{E}[\text{Sign}(\langle A, X \rangle - s) \langle A, Y \rangle | s \neq 0] + (1 - p) \mathbb{E}[\text{Sign}(\langle A, X \rangle) \langle A, Y \rangle]. \end{aligned}$$

This completes the proof. ■

Now, we provide an overview of our proof technique for Theorem 10. Let $\mathcal{G}_Y = \sup_{X \in \mathbb{S}_{k, \varepsilon}} \mathcal{H}_{X, Y}$ and $\bar{\mathcal{G}}_Y = \mathcal{G}_Y - \mathbb{E}[\mathcal{G}_Y]$ be stochastic processes indexed by Y . According to Lemma 45, it suffices to control $\sup_{Y \in \mathbb{S}_{k, \varepsilon/\zeta}} \mathcal{G}_Y$. We consider the following decomposition:

$$\begin{aligned} \sup_{Y \in \mathbb{S}_{k, \varepsilon/\zeta}} \mathcal{G}_Y &\leq \sup_{Y \in \mathbb{S}_{k, \varepsilon/\zeta}} \bar{\mathcal{G}}_Y + \sup_{Y \in \mathbb{S}_{k, \varepsilon/\zeta}} \mathbb{E}[\mathcal{G}_Y] \\ &\leq \sup_{Y \in \mathbb{S}_k} \bar{\mathcal{G}}_Y + \frac{\varepsilon}{\zeta} \sup_{Y \in \mathbb{S}} \bar{\mathcal{G}}_Y + \sup_{Y \in \mathbb{S}_k} \mathbb{E}[\mathcal{G}_Y] + \frac{\varepsilon}{\zeta} \sup_{Y \in \mathbb{S}} \mathbb{E}[\mathcal{G}_Y], \end{aligned} \quad (63)$$

where the second inequality follows from $\mathbb{S}_{k, \varepsilon/\zeta} \subset \mathbb{S}_k + \frac{\varepsilon}{\zeta} \mathbb{S}$, and the fact that both $\bar{\mathcal{G}}_Y$ and $\mathbb{E}[\mathcal{G}_Y]$ are linear functions of Y . We control the individual terms in the above decomposition

separately. To provide an upper bound for $\sup_{Y \in \mathbb{S}_k} \bar{\mathcal{G}}_Y$ and $\sup_{Y \in \mathbb{S}} \bar{\mathcal{G}}_Y$, we rely on the following key lemma.

Lemma 47 *The stochastic processes $\{\bar{\mathcal{G}}_Y\}_{Y \in \mathbb{S}_k}$ and $\{\bar{\mathcal{G}}_Y\}_{Y \in \mathbb{S}}$ are $\mathcal{O}(1/m)$ -sub-Gaussian processes.*

Proof Since $\mathbb{S}_k \subset \mathbb{S}$, it suffices to show that $\{\bar{\mathcal{G}}_Y\}_{Y \in \mathbb{S}}$ is $\mathcal{O}(1/m)$ -sub-Gaussian. According to Definition 41, the stochastic process $\{\bar{\mathcal{G}}_Y\}_{Y \in \mathbb{S}}$ is sub-Gaussian if for any arbitrary $Y, Y' \in \mathbb{S}$, $\bar{\mathcal{G}}_Y - \bar{\mathcal{G}}_{Y'}$ is $\mathcal{O}(\|Y - Y'\|_F^2/m)$ -sub-Gaussian. Note that $\bar{\mathcal{G}}_Y - \bar{\mathcal{G}}_{Y'}$ is sub-Gaussian if and only if $\mathcal{G}_Y - \mathcal{G}_{Y'}$ is sub-Gaussian with the same parameter. The latter will be proven by checking the moment bound condition in Definition 40. For arbitrary $Y, Y' \in \mathbb{S}$, denote $\Delta Y = Y - Y'$. Then, for any $p \in \mathbb{N}_+$, we have

$$\begin{aligned}
 & \mathbb{E} [|\mathcal{G}_Y - \mathcal{G}_{Y'}|^{2p}] \\
 & \leq \mathbb{E} \left[\left| \sup_{X \in \mathcal{S}_{k,\varepsilon}} \frac{1}{m} \sum_{i=1}^m \text{Sign}(\langle A_i, X \rangle - s_i) \langle A_i, \Delta Y \rangle - \varphi(X) \left\langle \frac{X}{\|X\|_F}, \Delta Y \right\rangle \right|^{2p} \right] \\
 & \leq \mathbb{E} \left[\left(\sup_{X \in \mathcal{S}_{k,\varepsilon}} \left| \frac{1}{m} \sum_{i=1}^m \text{Sign}(\langle A_i, X \rangle - s_i) \langle A_i, \Delta Y \rangle \right| + \sup_{X \in \mathcal{S}_{k,\varepsilon}} \left| \varphi(X) \left\langle \frac{X}{\|X\|_F}, \Delta Y \right\rangle \right| \right)^{2p} \right] \\
 & \leq \mathbb{E} \left[\left(\frac{1}{m} \sum_{i=1}^m |\langle A_i, \Delta Y \rangle| + \sqrt{\frac{2}{\pi}} \|\Delta Y\|_F \right)^{2p} \right] \\
 & \leq 2^{2p} \left(\mathbb{E} \left[\left(\frac{1}{m} \sum_{i=1}^m |\langle A_i, \Delta Y \rangle| \right)^{2p} \right] + \left(\frac{2}{\pi} \right)^p \|\Delta Y\|_F^{2p} \right),
 \end{aligned}$$

where in the third inequality, we used $\varphi(X) \leq \sqrt{2/\pi}$, which holds for every $X \in \mathbb{R}^{d \times d}$. According to (Li et al., 2020a, Appendix A.2), the random variable $(1/m) \sum_{i=1}^m |\langle A_i, \Delta Y \rangle|$ is $\mathcal{O}(\|\Delta Y\|_F^2/m)$ -sub-Gaussian with mean $\sqrt{2/\pi} \|\Delta Y\|_F$. Therefore, we have

$$\begin{aligned}
 & \mathbb{E} \left[\left(\frac{1}{m} \sum_{i=1}^m |\langle A_i, \Delta Y \rangle| \right)^{2p} \right] \\
 & = \mathbb{E} \left[\left(\frac{1}{m} \sum_{i=1}^m |\langle A_i, \Delta Y \rangle| - \sqrt{\frac{2}{\pi}} \|\Delta Y\|_F + \sqrt{\frac{2}{\pi}} \|\Delta Y\|_F \right)^{2p} \right] \\
 & \leq 2^{2p} \left(\mathbb{E} \left[\left(\frac{1}{m} \sum_{i=1}^m |\langle A_i, \Delta Y \rangle| - \sqrt{\frac{2}{\pi}} \|\Delta Y\|_F \right)^{2p} \right] + \left(\frac{2}{\pi} \right)^p \|\Delta Y\|_F^{2p} \right) \\
 & \leq 2^{2p} \left((2p-1)!! \frac{1}{m^p} \|\Delta Y\|_F^{2p} + \left(\frac{2}{\pi} \right)^p \|\Delta Y\|_F^{2p} \right).
 \end{aligned}$$

Combining the above inequalities leads to

$$\mathbb{E} [|\mathcal{G}_Y - \mathcal{G}_{Y'}|^{2p}]^{1/2p} \leq 4 \left((2p-1)!! \frac{1}{m^p} \|\Delta Y\|_F^{2p} + \left(\frac{2}{\pi}\right)^p \|\Delta Y\|_F^{2p} \right)^{1/2p} \lesssim \sqrt{\frac{p}{m}} \|\Delta Y\|_F,$$

given that $p > m$. Therefore, $\{\mathcal{G}_Y\}_{Y \in \mathbb{S}}$ is a $\mathcal{O}(1/m)$ -sub-Gaussian process, which implies that $\{\mathcal{G}_Y\}_{Y \in \mathbb{S}_k}$ is also a $\mathcal{O}(1/m)$ -sub-Gaussian process. \blacksquare

Given that both $\{\bar{\mathcal{G}}_Y\}_{Y \in \mathbb{S}_k}$ and $\{\bar{\mathcal{G}}_Y\}_{Y \in \mathbb{S}}$ are sub-Gaussian processes, we can readily obtain sharp concentration bounds on their suprema.

Lemma 48 *The following statements hold:*

$$\mathbb{E} \left[\sup_{Y \in \mathbb{S}_k} \bar{\mathcal{G}}_Y \right] \lesssim \sqrt{\frac{dk}{m}}, \quad \mathbb{P} \left(\sup_{Y \in \mathbb{S}_k} \bar{\mathcal{G}}_Y \geq \mathbb{E} \left[\sup_{Y \in \mathbb{S}_k} \bar{\mathcal{G}}_Y \right] + \gamma \right) \lesssim e^{-cm\gamma^2}; \quad (64)$$

$$\mathbb{E} \left[\sup_{Y \in \mathbb{S}} \bar{\mathcal{G}}_Y \right] \lesssim \sqrt{\frac{d^2}{m}}, \quad \mathbb{P} \left(\sup_{Y \in \mathbb{S}} \bar{\mathcal{G}}_Y \geq \mathbb{E} \left[\sup_{Y \in \mathbb{S}} \bar{\mathcal{G}}_Y \right] + \gamma \right) \lesssim e^{-cm\gamma^2}. \quad (65)$$

Proof The proof follows directly from Theorem 44. The details are omitted for brevity. \blacksquare

Equipped with the above lemma, we provide a concentration bound on $\sup_{Y \in \mathbb{S}_{k,\varepsilon/\zeta}} \bar{\mathcal{G}}_Y$.

Lemma 49 *Assume that $m \gtrsim \sqrt{dk/\gamma^2}$ and $\varepsilon/\zeta \lesssim \sqrt{k/d}$. Then, the following inequality holds with probability of at least $1 - Ce^{-cm\gamma^2}$:*

$$\sup_{Y \in \mathbb{S}_{k,\varepsilon/\zeta}} \bar{\mathcal{G}}_Y \leq \left(3 + \sqrt{\frac{k}{d}} \right) \gamma.$$

Proof Based on (64), we have, with probability of at least $1 - Ce^{-cm\gamma^2}$

$$\sup_{Y \in \mathbb{S}_k} \bar{\mathcal{G}}_Y \leq \mathbb{E} \left[\sup_{Y \in \mathbb{S}_k} \bar{\mathcal{G}}_Y \right] + \gamma \lesssim \sqrt{\frac{dk}{m}} + \gamma \leq 2\gamma,$$

where the last inequality follows from the assumption $m \gtrsim \sqrt{dk/\gamma^2}$. Similarly, based on (65), the following inequalities hold with a probability of at least $1 - Ce^{-cm\gamma^2}$:

$$\frac{\varepsilon}{\zeta} \sup_{Y \in \mathbb{S}} \bar{\mathcal{G}}_Y \leq \frac{\varepsilon}{\zeta} \left(\mathbb{E} \left[\sup_{Y \in \mathbb{S}} \bar{\mathcal{G}}_Y \right] + \gamma \right) \lesssim \sqrt{\frac{\varepsilon^2 d^2}{\zeta^2 m}} + \frac{\varepsilon \gamma}{\zeta} \leq \left(1 + \sqrt{\frac{k}{d}} \right) \gamma,$$

where the last inequality follows from $m \gtrsim \sqrt{dk/\gamma^2}$ and $\varepsilon/\zeta \lesssim \sqrt{k/d}$. Finally, a simple union bound implies that

$$\sup_{Y \in \mathbb{S}_{k,\varepsilon/\zeta}} \bar{\mathcal{G}}_Y \leq \sup_{Y \in \mathbb{S}_k} \bar{\mathcal{G}}_Y + \frac{\varepsilon}{\zeta} \sup_{Y \in \mathbb{S}} \bar{\mathcal{G}}_Y \leq \left(3 + \sqrt{\frac{k}{d}} \right) \gamma,$$

with probability of at least $1 - Ce^{-cm\gamma^2}$, thereby completing the proof. \blacksquare

Recalling (63), it remains to control the terms $\sup_{Y \in \mathbb{S}_k} \mathbb{E}[\mathcal{G}_Y]$ and $\sup_{Y \in \mathbb{S}} \mathbb{E}[\mathcal{G}_Y]$.

Lemma 50 *The following inequality holds:*

$$\sup_{Y \in \mathbb{S}_k} \mathbb{E}[\mathcal{G}_Y] \leq \sup_{Y \in \mathbb{S}} \mathbb{E}[\mathcal{G}_Y] \lesssim \sqrt{\frac{dk}{m} \log^2(m) \log\left(\frac{R}{\zeta}\right)}. \quad (66)$$

Due to its length, we defer the proof of Lemma 50 to Appendix F.1. Equipped with Lemmas 49 and 50, we are ready to present the proof of Theorem 10.

Proof of Theorem 10. The inequality (63) combined with Lemmas 49 and 50 implies that the following inequality holds with probability of at least $1 - Ce^{-cm\gamma^2}$:

$$\sup_{Y \in \mathbb{S}_{k,\varepsilon/\zeta}} \mathcal{G}_Y \lesssim \sqrt{\frac{dk}{m} \log^2(m) \log\left(\frac{R}{\zeta}\right)} + \gamma.$$

On the other hand, one can write

$$\inf_{X \in \mathcal{S}_{k,\varepsilon}} \varphi(X) = \inf_{X \in \mathcal{S}_{k,\varepsilon}} \left\{ \sqrt{\frac{2}{\pi}} \left(1 - p + p \mathbb{E} \left[e^{-s^2/(2\|X\|_F^2)} \right] \right) \right\} \geq \sqrt{\frac{2}{\pi}}(1-p). \quad (67)$$

Therefore, upon choosing

$$m \gtrsim \frac{dk \log^2(m) \log(R/\zeta)}{(1-p)^2 \delta^2}, \quad \text{and} \quad \gamma \lesssim (1-p)\delta,$$

we have

$$\sup_{Y \in \mathbb{S}_{k,\varepsilon/\zeta}} \mathcal{G}_Y = \sup_{\substack{X \in \mathcal{S}_{k,\varepsilon} \\ Y \in \mathbb{S}_{k,\varepsilon/\zeta}}} \mathcal{H}_{X,Y} \leq \inf_{X \in \mathcal{S}_{k,\varepsilon}} \varphi(X) \delta$$

with probability of at least $1 - Ce^{-cm(1-p)^2\delta^2}$. This completes the proof of Theorem 10. \square

B.3 Proof of Theorem 11

Recall that, with the outlier noise model, the scaling function takes the form $\varphi(X) = \sqrt{\frac{2}{\pi}} \left(1 - p + p \mathbb{E} \left[e^{-s^2/(2\|X\|_F^2)} \right] \right)$. Setting $p = 1$, and $s_i \sim \mathcal{N}(0, \nu_g^2)$ immediately implies $\varphi(X) = \mathbb{E}_{s \sim \mathcal{N}(0, \nu_g^2)} \left[e^{-s^2/(2\|X\|_F^2)} \right] = \sqrt{\frac{2}{\pi}} \frac{\|X\|_F}{\sqrt{\|X\|_F^2 + \nu_g^2}}$. On the other hand, using the same method in the proof of Theorem 10, we can show that the following inequality holds with probability at least $1 - Ce^{-cm\gamma^2}$:

$$\sup_{Y \in \mathbb{S}_{k,\varepsilon/\zeta}} \mathcal{G}_Y \lesssim \sqrt{\frac{dk}{m} \log^2(m) \log\left(\frac{R}{\zeta}\right)} + \gamma,$$

where \mathcal{G}_Y is defined in the proof of Theorem 10. It remains to bound $\inf_{X \in \mathcal{S}_{k,\varepsilon}} \varphi(X)$. To this goal, note that $\|X\|_F \geq \zeta, \forall X \in \mathcal{S}_{k,\varepsilon}$. Hence, we have the following lower bound for the scaling function

$$\inf_{X \in \mathcal{S}_{k,\varepsilon}} \varphi(X) = \inf_{X \in \mathcal{S}_{k,\varepsilon}} \sqrt{\frac{2}{\pi}} \frac{\|X\|_F}{\sqrt{\|X\|_F^2 + \nu_g^2}} \geq \sqrt{\frac{2}{\pi}} \frac{\zeta}{\sqrt{\zeta^2 + \nu_g^2}} \gtrsim \frac{\zeta}{\nu_g}, \quad (68)$$

provided that $\zeta \lesssim \nu_g$. Therefore, upon choosing $m \gtrsim \frac{\nu_g^2 dk \log^2(m) \log(R/\zeta)}{\zeta^2 \delta^2}$ and $\gamma \lesssim \frac{\zeta}{\nu_g} \delta$, we have

$$\sup_{Y \in \mathcal{S}_{k,\varepsilon/\zeta}} \mathcal{G}_Y = \sup_{\substack{X \in \mathcal{S}_{k,\varepsilon} \\ Y \in \mathcal{S}_{k,\varepsilon/\zeta}}} \mathcal{H}_{X,Y} \leq \inf_{X \in \mathcal{S}_{k,\varepsilon}} \varphi(X) \delta, \quad (69)$$

with probability at least $1 - C_1 e^{-C_2 m \zeta^2 \delta^2 / \nu_g^2}$, which completes the proof. \square

B.4 Proof of Lemma 12

Lemma 51 is a direct consequence of the following more general result:

Lemma 51 *Suppose that the measurements are noiseless, and satisfy $(k, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP for any $\varepsilon \geq 0$, \mathcal{S} , and scaling function $\varphi(\cdot)$. Then, for any ε -approximate rank- k matrix $X \in \mathcal{S}$, we have*

$$\left| \frac{1}{m} \sum_{i=1}^m |\langle A_i, X \rangle| - \varphi(X) \|X\|_F \right| \leq \delta \varphi(X) \|X\|_F.$$

Proof According to the definition of Sign-RIP, for every ε -approximate rank- k matrices $X, Y \in \mathcal{S}$, and every $Q \in \mathcal{Q}(X)$, we have $\left| \left\langle Q - \varphi(X) \frac{X}{\|X\|_F}, \frac{Y}{\|Y\|_F} \right\rangle \right| \leq \varphi(X) \delta$. In particular, upon choosing $Y = X$, we have

$$\left| \left\langle Q - \varphi(X) \frac{X}{\|X\|_F}, X \right\rangle \right| = \left| \frac{1}{m} \sum_{i=1}^m |\langle A_i, X \rangle| - \varphi(X) \|X\|_F \right| \leq \varphi(X) \delta \|X\|_F.$$

This completes the proof. \blacksquare

Evidently, the result of Lemma 51 can be readily recovered from the above lemma after substituting X with Δ_t . Moreover, it is worth noting that the above lemma recovers the so-called ℓ_1/ℓ_2 -RIP introduced in (Li et al., 2020a) with $\varepsilon = 0$ and $\varphi(X) = \sqrt{2/\pi}$.

B.5 Proof of Lemma 13

Due to Sign-RIP, we have

$$\begin{aligned} \|Q_t\| &\leq \varphi(\Delta_t) \left(\frac{\|\Delta_t\|}{\|\Delta_t\|_F} + \delta \right) \\ \implies \frac{1}{\|Q_t\|} &\geq \frac{1}{\varphi(\Delta_t)} \frac{1}{\delta + \frac{\|\Delta_t\|}{\|\Delta_t\|_F}} \\ \implies \frac{1}{\|Q_t\|} &\geq \frac{1}{\varphi(\Delta_t)} \frac{\|\Delta_t\|_F}{\|\Delta_t\|} \left(1 - \delta \frac{\|\Delta_t\|_F}{\|\Delta_t\|} \right). \end{aligned} \quad (70)$$

By our assumption, Δ_t is ε -approximate rank- k . This implies that, there exists a rank- k matrix X' such that $\|\Delta_t - X'\|_F \leq \varepsilon$. This in turn implies that

$$\begin{aligned} \|\Delta_t\|_F &\leq \|\Delta_t - X'\|_F + \|X'\|_F \\ &\leq \varepsilon + \|X'\|_F \\ &\leq \varepsilon + \sqrt{k} \|X'\| \\ &\leq (1 + \sqrt{k}) \varepsilon + \sqrt{k} \|\Delta_t\| \end{aligned}$$

and

$$\begin{aligned}
 \|\Delta_t\| &\geq \|X'\| - \|\Delta_t - X'\| \\
 &\geq \|X'\| - \varepsilon \\
 &\geq \|\Delta_t\| - 2\varepsilon \\
 &\geq \|\Delta_t\| / 2,
 \end{aligned}$$

where in the last inequality we used the assumption $\|\Delta_t\| \geq 4\varepsilon$. Combining the above inequalities implies that

$$\frac{\|\Delta_t\|_F}{\|\Delta_t\|} \leq \frac{2(1 + \sqrt{k})\varepsilon}{\|\Delta_t\|} + 2\sqrt{k} \leq \frac{1 + 5\sqrt{k}}{2}.$$

Combining the above inequality with (70) leads to

$$\eta_t = \frac{\eta\rho^t}{\|Q_t\|} \geq \left(1 - \left(\frac{1 + 5\sqrt{k}}{2}\right)\delta\right) \frac{\eta\rho^t}{\varphi(\Delta_t)} \frac{\|\Delta_t\|_F}{\|\Delta_t\|}.$$

The upper bound can be derived in a similar fashion. □

Appendix C. Proofs of the Expected Loss

C.1 Proof of Proposition 16

According to the update rule $U_{t+1} = U_t - \eta(U_t U_t^\top - X^*)U_t$, we have

$$S_{t+1} = V^\top U_{t+1} = S_t - \eta \left((S_t S_t^\top - \Sigma) S_t + S_t E_t^\top E_t \right).$$

Define an auxiliary matrix M as

$$M := (I - \Xi) \left(S_t - \eta (S_t S_t^\top - \Sigma) S_t \right) \left(S_t^\top - \eta S_t^\top (S_t S_t^\top - \Sigma) \right) (I - \Xi^\top), \quad (71)$$

where $\Xi = \eta S_t E_t^\top E_t S_t^\top (S_t S_t^\top)^{-1} (I - \eta (S_t S_t^\top - \Sigma))^{-1}$. Note that, due to our assumptions and the choice of η , we have $S_t S_t^\top \succ 0$ and $\eta \|S_t S_t^\top - \Sigma\| < 1$. Therefore, both matrices $S_t S_t^\top$ and $I - \eta (S_t S_t^\top - \Sigma)$ are invertible and the matrix Ξ is well-defined. Our goal is to first show that $\lambda_{\min}(S_{t+1} S_{t+1}^\top) \geq \lambda_{\min}(M)$, and then derive a lower bound for $\lambda_{\min}(M)$. Based on the definition of Ξ , it is easy to verify that $\Xi (S_t - \eta (S_t S_t^\top - \Sigma) S_t) = S_t E_t^\top E_t P_{S_t}$. Combining this with (71) reveals that

$$S_{t+1} S_{t+1}^\top - M = \eta^2 S_t E_t^\top E_t P_{S_t}^\perp E_t^\top E_t S_t^\top \succeq 0.$$

Therefore, instead of obtaining a lower bound for $\lambda_{\min}(S_{t+1}S_{t+1}^\top)$, it suffices to obtain a lower bound for $\lambda_{\min}(M)$. One can write

$$\begin{aligned}
 \|\Xi\| &\leq \eta \left\| S_t E_t^\top E_t S_t^\top \left(S_t S_t^\top \right)^{-1} \right\| \left\| \left(I - \eta \left(S_t S_t^\top - \Sigma \right) \right)^{-1} \right\| \\
 &\lesssim \eta \left\| S_t E_t^\top E_t S_t^\top \left(S_t S_t^\top \right)^{-1} \right\| \\
 &= \eta \left\| \left(S_t S_t^\top \right)^{-1/2} S_t E_t^\top E_t S_t^\top \left(S_t S_t^\top \right)^{-1/2} \right\| \\
 &\leq \eta \left\| E_t E_t^\top \right\| \left\| \left(S_t S_t^\top \right)^{-1/2} S_t S_t^\top \left(S_t S_t^\top \right)^{-1/2} \right\| \\
 &= \eta \left\| E_t E_t^\top \right\| \\
 &< 1,
 \end{aligned} \tag{72}$$

where the last inequality is due to the assumption $\|E_t E_t^\top\| \leq \sigma_1$ and the choice of η . Therefore, we have

$$\begin{aligned}
 \lambda_{\min} \left(S_{t+1} S_{t+1}^\top \right) &\geq \lambda_{\min} (M) \\
 &\geq (1 - \|\Xi\|)^2 \lambda_{\min} \left(\left(S_t - \eta \left(S_t S_t^\top - \Sigma \right) S_t \right) \left(S_t^\top - \eta S_t^\top \left(S_t S_t^\top - \Sigma \right) \right) \right).
 \end{aligned} \tag{73}$$

Now it suffices to bound $\lambda_{\min} \left(\left(S_t - \eta \left(S_t S_t^\top - \Sigma \right) S_t \right) \left(S_t^\top - \eta S_t^\top \left(S_t S_t^\top - \Sigma \right) \right) \right)$. First, note that

$$S_t - \eta \left(S_t S_t^\top - \Sigma \right) S_t = \left(I + \eta \Sigma \left(I - \eta S_t S_t^\top \right)^{-1} \right) \left(S_t - \eta S_t S_t^\top S_t \right).$$

Based on the above equality, one can write

$$\begin{aligned}
 & \lambda_{\min} \left(\left(S_t - \eta \left(S_t S_t^\top - \Sigma \right) S_t \right) \left(S_t^\top - \eta S_t^\top \left(S_t S_t^\top - \Sigma \right) \right) \right) \\
 &= \lambda_{\min} \left(\left(I + \eta \Sigma \left(I - \eta S_t S_t^\top \right)^{-1} \right) \left(S_t - \eta S_t S_t^\top S_t \right) \right. \\
 &\quad \left. \left(S_t^\top - \eta S_t^\top S_t S_t^\top \right) \left(I + \eta \Sigma \left(I - \eta S_t S_t^\top \right)^{-1} \right) \right) \\
 &\geq \lambda_{\min} \left(\left(I + \eta \Sigma \left(I - \eta S_t S_t^\top \right)^{-1} \right) \right)^2 \lambda_{\min} \left(\left(S_t - \eta S_t S_t^\top S_t \right) \left(S_t^\top - \eta S_t^\top S_t S_t^\top \right) \right) \\
 &\stackrel{(a)}{\geq} \left(1 + \eta \sigma_r \left(1 - \eta \lambda_{\min} \left(S_t S_t^\top \right) \right)^{-1} \right)^2 \lambda_{\min} \left(S_t S_t^\top - 2\eta \left(S_t S_t^\top \right)^2 + \eta^2 \left(S_t S_t^\top \right)^3 \right) \quad (74) \\
 &\stackrel{(b)}{=} \left(1 + \eta \sigma_r \left(1 - \eta \lambda_{\min} \left(S_t S_t^\top \right) \right)^{-1} \right)^2 \\
 &\quad \times \left(\lambda_{\min} \left(S_t S_t^\top \right) - 2\eta \lambda_{\min} \left(S_t S_t^\top \right)^2 + \eta^2 \lambda_{\min} \left(S_t S_t^\top \right)^3 \right) \\
 &\geq \left(1 + \eta \sigma_r \left(1 - \eta \lambda_{\min} \left(S_t S_t^\top \right) \right)^{-1} \right)^2 \left(\lambda_{\min} \left(S_t S_t^\top \right) - 2\eta \lambda_{\min} \left(S_t S_t^\top \right)^2 \right) \\
 &= \left(1 + \eta \sigma_r \right)^2 \lambda_{\min} \left(S_t S_t^\top \right) - 2\eta \left(1 + \eta \sigma_r \right) \lambda_{\min} \left(S_t S_t^\top \right)^2,
 \end{aligned}$$

where in (a), we used the fact that

$$\begin{aligned}
 \lambda_{\min} \left(I + \eta \Sigma \left(I - \eta S_t S_t^\top \right)^{-1} \right) &= 1 + \lambda_{\min} \left(\eta \Sigma \left(I - \eta S_t S_t^\top \right)^{-1} \right) \\
 &= 1 + \eta \lambda_{\min} \left(\Sigma^{1/2} \left(I - \eta S_t S_t^\top \right)^{-1} \Sigma^{1/2} \right) \\
 &\geq 1 + \eta \sigma_r \lambda_{\min} \left(\left(I - \eta S_t S_t^\top \right)^{-1} \right) \\
 &= 1 + \eta \sigma_r \left(1 - \eta \lambda_{\min} \left(S_t S_t^\top \right) \right)^{-1},
 \end{aligned}$$

and in (b) we used the fact that the matrices $S_t S_t^\top$, $(S_t S_t^\top)^2$, and $(S_t S_t^\top)^3$ share the same eigenvectors. Combining (74) with (72) and (73) completes the proof. \square

C.2 Proof of Proposition 17

Before delving into the details, we provide the update rule for S_t and E_t , which will be used frequently throughout our proof. Applying the signal-residual decomposition to $U_{t+1} = U_t - \eta (U_t U_t^\top - X^*) U_t$ leads to

$$S_{t+1} = V^\top U_{t+1} = S_t - \eta \left(\left(S_t S_t^\top - \Sigma \right) S_t + S_t E_t^\top E_t \right), \quad (75)$$

$$E_{t+1} = V_\perp^\top U_{t+1} = E_t - \eta E_t \left(S_t^\top S_t + E_t^\top E_t \right). \quad (76)$$

Bounding $\|\Sigma - S_{t+1}S_{t+1}^\top\|$: The update rule for S_{t+1} leads to

$$\begin{aligned}
 & \Sigma - S_{t+1}S_{t+1}^\top \\
 &= \underbrace{\Sigma - S_tS_t^\top + \eta S_tS_t^\top (S_tS_t^\top - \Sigma) + \eta (S_tS_t^\top - \Sigma) S_tS_t^\top - \eta^2 (S_tS_t^\top - \Sigma) S_tS_t^\top (S_tS_t^\top - \Sigma)}_{(A)} \\
 & \quad + \underbrace{2\eta S_tE_t^\top E_tS_t^\top - \eta^2 (S_tS_t^\top - \Sigma) S_tE_t^\top E_tS_t^\top - \eta^2 S_tE_t^\top E_tS_t^\top (S_tS_t^\top - \Sigma)}_{(B)} \\
 & \quad - \underbrace{\eta^2 S_tE_t^\top E_tE_t^\top E_tS_t^\top}_{(C)}.
 \end{aligned} \tag{77}$$

First, we provide an upper bound for (B). Recall that $\eta = \mathcal{O}(1/\sigma_1)$. Moreover, we have $\|S_tS_t^\top - \Sigma\| \leq 2.01\sigma_1$ due to the assumption $\|S_tS_t^\top\| \leq 1.01\sigma_1$. Therefore, one can write

$$\|(B)\| \leq 2\eta \|S_tE_t^\top E_tS_t^\top\| + 2\eta^2 \|S_tS_t^\top - \Sigma\| \|S_tE_t^\top E_tS_t^\top\| \leq 4\eta \|S_tE_t^\top\|^2.$$

Similarly, due to our assumption on $\|E_tE_t^\top\|$ and η , we have $\|E_tE_t^\top\| \leq 1/\eta$, which in turn implies

$$\|(C)\| \leq \eta^2 \|E_tE_t^\top\| \|S_tE_t^\top\|^2 \leq \eta \|S_tE_t^\top\|^2.$$

Finally, we provide an upper bound for (A). First, one can verify that

$$(A) = \underbrace{(\Sigma - S_tS_t^\top)}_{(A_1)} \underbrace{(0.5I - \eta S_tS_t^\top)}_{(A_2)} + \underbrace{(0.5I - \eta S_tS_t^\top + \eta^2 (S_tS_t^\top - \Sigma) S_tS_t^\top)}_{(A_2)} (\Sigma - S_tS_t^\top).$$

For the first term, we have

$$\|(A_1)\| \leq \|0.5I - \eta S_tS_t^\top\| \|\Sigma - S_tS_t^\top\| \leq (0.5 - \eta\lambda_{\min}(S_tS_t^\top)) \|\Sigma - S_tS_t^\top\|.$$

To provide a bound for (A₂), observe that

$$0.5I - \eta S_tS_t^\top + \eta^2 (S_tS_t^\top - \Sigma) S_tS_t^\top = 0.5I - \eta \left(I + \eta (\Sigma - S_tS_t^\top) \right) S_tS_t^\top.$$

The next step in our proof is to show that, with the choice of $\eta \lesssim 1/\sigma_1$, the eigenvalues of $(I + \eta(\Sigma - S_tS_t^\top)) S_tS_t^\top$ are nonnegative. We prove this by showing that the eigenvalues of $(I + \eta(\Sigma - S_tS_t^\top)) S_tS_t^\top$ are close to those of $S_tS_t^\top$. First, note that $I + \eta(\Sigma - S_tS_t^\top)$ is positive definite due to our choice of step-size. Therefore, the matrix $D = (I + \eta(\Sigma - S_tS_t^\top))^{1/2}$ is well-defined and invertable. Hence, for any $1 \leq i \leq r$, we have

$$\lambda_i \left((I + \eta(\Sigma - S_tS_t^\top)) S_tS_t^\top \right) = \lambda_i \left(D^{-1} \left(I + \eta(\Sigma - S_tS_t^\top) \right) S_tS_t^\top D \right) = \lambda_i \left(DS_tS_t^\top D \right).$$

To proceed with our argument, we first present a relative perturbation bound for symmetric matrices.

Lemma 52 (Relative Perturbation Bound, Eisenstat and Ipsen (1995)) *Let $X \in \mathbb{R}^{d \times d}$ be a symmetric matrix with eigenvalues $\lambda_1 \geq \dots \geq \lambda_d$. Moreover, suppose R is a non-singular matrix. Let $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_d$ be the eigenvalues of $Y = R^\top X R$. Then, we have*

$$\left| \lambda_i - \hat{\lambda}_i \right| \leq |\lambda_i| \left\| I - R^\top R \right\|, \quad \text{for all } i.$$

Invoking Lemma 52 with $X = S_t S_t^\top$ and $Y = D S_t S_t^\top D$ results in

$$\begin{aligned} & \left| \lambda_i \left(\left(I + \eta \left(\Sigma - S_t S_t^\top \right) \right) S_t S_t^\top \right) - \lambda_i \left(S_t S_t^\top \right) \right| \\ & \leq \left| \lambda_i \left(S_t S_t^\top \right) \right| \left\| I - D^2 \right\| \\ & = \left| \lambda_i \left(S_t S_t^\top \right) \right| \left\| I - \left(I + \eta \left(\Sigma - S_t S_t^\top \right) \right) \right\| \\ & = \eta \left| \lambda_i \left(S_t S_t^\top \right) \right| \left\| \Sigma - S_t S_t^\top \right\| \\ & \leq 0.5 \left| \lambda_i \left(S_t S_t^\top \right) \right|. \end{aligned}$$

for every $i = 1, 2, \dots, r$. The above inequality implies that $\lambda_{\min} \left(\left(I + \eta \left(\Sigma - S_t S_t^\top \right) \right) S_t S_t^\top \right) \asymp \lambda_{\min} \left(S_t S_t^\top \right)$, and therefore, $\lambda_{\min} \left(\left(I + \eta \left(\Sigma - S_t S_t^\top \right) \right) S_t S_t^\top \right) \geq 0$. This leads to

$$\begin{aligned} & \left\| 0.5I - \eta S_t S_t^\top + \eta^2 \left(S_t S_t^\top - \Sigma \right) S_t S_t^\top \right\| \\ & = \left\| 0.5I - \eta S_t S_t^\top + \eta^2 \left(S_t S_t^\top \right)^{1/2} \left(S_t S_t^\top - \Sigma \right) \left(S_t S_t^\top \right)^{1/2} \right\| \\ & \leq 0.5 - \lambda_{\min} \left(\eta S_t S_t^\top - \eta^2 \left(S_t S_t^\top \right)^{1/2} \left(S_t S_t^\top - \Sigma \right) \left(S_t S_t^\top \right)^{1/2} \right) \\ & = 0.5 - \lambda_{\min} \left(\eta S_t S_t^\top - \eta^2 \left(S_t S_t^\top - \Sigma \right) \left(S_t S_t^\top \right) \right) \\ & = 0.5 - \eta \lambda_{\min} \left(\left(I + \eta \left(\Sigma - S_t S_t^\top \right) \right) S_t S_t^\top \right) \leq 0.5, \end{aligned}$$

where in the last inequality, we used the fact that $\lambda_{\min} \left(\left(I + \eta \left(\Sigma - S_t S_t^\top \right) \right) S_t S_t^\top \right) \geq 0$. Combined with the definition of (A_2) , we have

$$\|(A_2)\| \leq \left\| 0.5I - \eta S_t S_t^\top + \eta^2 \left(S_t S_t^\top - \Sigma \right) S_t S_t^\top \right\| \left\| \Sigma - S_t S_t^\top \right\| \leq 0.5 \left\| \Sigma - S_t S_t^\top \right\|,$$

which in turn implies

$$\|(A)\| \leq \left(1 - \eta \lambda_{\min} \left(S_t S_t^\top \right) \right) \left\| \Sigma - S_t S_t^\top \right\|.$$

Combining the derived upper bounds for (A) , (B) , and (C) completes the proof for the signal dynamics. \square

Bounding $\|S_{t+1}E_{t+1}^\top\|$: Recalling the update rules (75) and (76), we have

$$\begin{aligned}
 S_{t+1}E_{t+1}^\top &= \underbrace{S_tE_t^\top + \eta(\Sigma - S_tS_t^\top)S_tE_t^\top - \eta S_t(S_t^\top S_t + E_t^\top E_t)E_t^\top}_{(A)} \\
 &\quad + \underbrace{\eta^2(S_tS_t^\top - \Sigma)S_t(S_t^\top S_t + E_t^\top E_t)E_t^\top}_{(B)} - \underbrace{\eta S_tE_t^\top E_tE_t^\top}_{(C)} \\
 &\quad + \underbrace{\eta^2 S_tE_t^\top E_t(S_t^\top S_t + E_t^\top E_t)E_t^\top}_{(D)}.
 \end{aligned}$$

We provide separate bounds for the individual terms in the above equality. First, observe that $\|(C)\| \leq \eta \|E_tE_t^\top\| \|S_tE_t^\top\|$. Moreover, one can write

$$\begin{aligned}
 \|(B)\| &\leq \eta^2 \|\Sigma - S_tS_t^\top\| \|S_t(S_t^\top S_t + E_t^\top E_t)E_t^\top\| \\
 &\leq \eta^2 \|\Sigma - S_tS_t^\top\| (\|S_tS_t^\top\| + \|E_tE_t^\top\|) \|S_tE_t^\top\| \\
 &\leq \eta \|\Sigma - S_tS_t^\top\| \|S_tE_t^\top\|,
 \end{aligned}$$

where, in the last inequality, we used the assumption that $\|E_tE_t^\top\| \lesssim \sigma_1$ and $\|S_tS_t^\top\| \lesssim \sigma_1$. Similarly, we have

$$\begin{aligned}
 \|(D)\| &\stackrel{(a)}{\leq} \eta^2 \|S_tE_t^\top\| (\|S_tS_t^\top\| + \|E_tE_t^\top\|) \|E_tE_t^\top\| \\
 &\leq \eta \|E_tE_t^\top\| \|S_tE_t^\top\|,
 \end{aligned}$$

where, in (a), we used $\|S_tS_t^\top\| E_tE_t^\top \succeq E_tS_t^\top S_tE_t^\top \succeq 0$. Finally, we provide an upper bound for (A).

$$\begin{aligned}
 \|(A)\| &\leq \left(\eta \|\Sigma - S_tS_t^\top\| + \|I - \eta S_tS_t^\top\| \right) \|S_tE_t^\top\| \\
 &\leq \left(\eta \|\Sigma - S_tS_t^\top\| + 1 - \eta \lambda_{\min}(S_tS_t^\top) \right) \|S_tE_t^\top\|.
 \end{aligned}$$

Combining the derived bounds for (A), (B), (C), and (D) concludes the proof. \square

Bounding $\|E_{t+1}E_{t+1}^\top\|$: Due to the update rule for E_{t+1} , one can write

$$\begin{aligned}
 E_{t+1}E_{t+1}^\top &= E_tE_t^\top - 2\eta E_t \left(E_t^\top E_t + S_t^\top S_t \right) E_t^\top + \eta^2 E_t \left(E_t^\top E_t + S_t^\top S_t \right)^2 E_t^\top \\
 &= E_t \left(I - 2\eta \left(E_t^\top E_t + S_t^\top S_t \right) + \eta^2 \left(E_t^\top E_t + S_t^\top S_t \right)^2 \right) E_t^\top.
 \end{aligned} \tag{78}$$

On the other hand, $\eta \lesssim 1/\sigma_1$, $\|E_tE_t^\top\| \lesssim \sigma_1$, and $\|S_tS_t^\top\| \lesssim \sigma_1$ imply that

$$I - \eta E_t^\top E_t \succ I - 2\eta \left(E_t^\top E_t + S_t^\top S_t \right) + \eta^2 \left(E_t^\top E_t + S_t^\top S_t \right)^2 \succ 0. \tag{79}$$

Hence, we have

$$\begin{aligned}
 \|E_t E_t^\top\| &= \left\| E_t \left(I - 2\eta \left(E_t^\top E_t + S_t^\top S_t \right) + \eta^2 \left(E_t^\top E_t + S_t^\top S_t \right)^2 \right) E_t^\top \right\| \\
 &\leq \left\| E_t \left(I - \eta E_t^\top E_t \right) E_t^\top \right\| \\
 &= \left\| E_t E_t^\top - \eta E_t E_t^\top E_t E_t^\top \right\| \\
 &= \left(1 - \eta \left\| E_t E_t^\top \right\| \right) \left\| E_t E_t^\top \right\|,
 \end{aligned}$$

where in the last inequality, we used the fact that $\|E_t E_t^\top\| \leq \sigma_1$.

Bounding $\|S_{t+1} S_{t+1}^\top\|$: First, recall that

$$\begin{aligned}
 S_{t+1} S_{t+1}^\top &= S_t S_t^\top - 2\eta S_t S_t^\top S_t S_t^\top + \eta \Sigma S_t S_t^\top + \eta S_t S_t^\top \Sigma - 2\eta S_t E_t E_t^\top S_t^\top \\
 &\quad + \eta^2 \left(S_t S_t^\top - \Sigma \right) S_t S_t^\top \left(S_t S_t^\top - \Sigma \right) + \eta^2 S_t E_t^\top E_t E_t^\top E_t S_t^\top \\
 &\quad + \eta^2 \left(S_t S_t^\top - \Sigma \right) S_t E_t E_t^\top S_t^\top + \eta^2 S_t E_t E_t^\top S_t^\top \left(S_t S_t^\top - \Sigma \right) \\
 &\leq \underbrace{S_t S_t^\top - 2\eta S_t S_t^\top S_t S_t^\top + \eta \Sigma S_t S_t^\top + \eta S_t S_t^\top \Sigma + \eta^2 \left(S_t S_t^\top - \Sigma \right) S_t S_t^\top \left(S_t S_t^\top - \Sigma \right)}_{(A)} \\
 &\quad + \underbrace{\eta^2 S_t E_t^\top E_t E_t^\top E_t S_t^\top + \eta^2 \left(S_t S_t^\top - \Sigma \right) S_t E_t E_t^\top S_t^\top + \eta^2 S_t E_t E_t^\top S_t^\top \left(S_t S_t^\top - \Sigma \right)}_{(B)},
 \end{aligned}$$

where the last inequality follows by noting that $-2\eta S_t E_t E_t^\top S_t^\top \leq 0$. Now, it is easy to see that

$$\begin{aligned}
 \|(B)\| &\leq \eta^2 \left\| E_t E_t^\top \right\|^2 \left\| S_t S_t^\top \right\| + 2\eta^2 \left\| E_t E_t^\top \right\| \left\| S_t S_t^\top \right\| \left(\sigma_1 + \left\| S_t S_t^\top \right\| \right) \\
 &\leq 3\eta^2 \left\| E_t E_t^\top \right\| \left\| S_t S_t^\top \right\| \left(\sigma_1 + \left\| S_t S_t^\top \right\| \right).
 \end{aligned}$$

It remains to provide an upper bound for (A). One can write

$$\begin{aligned}
 (A) &= \underbrace{S_t S_t^\top - 2\eta \left(S_t S_t^\top \right)^2 + \eta^2 \left(\left(S_t S_t^\top \right)^3 + \Sigma S_t S_t^\top \Sigma \right)}_{(A_1)} \\
 &\quad + \underbrace{\eta \Sigma S_t S_t^\top \left(I - \eta S_t S_t^\top \right) \eta S_t S_t^\top \left(I - \eta S_t S_t^\top \right) \Sigma}_{(A_2)}.
 \end{aligned}$$

For (A₁), we have

$$\begin{aligned}
 \|(A_1)\| &\leq \left\| S_t S_t^\top - 2\eta \left(S_t S_t^\top \right)^2 + \eta^2 \left(S_t S_t^\top \right)^3 \right\| + \eta^2 \left\| \Sigma S_t S_t^\top \Sigma \right\| \\
 &\leq \left\| S_t S_t^\top - 2\eta \left(S_t S_t^\top \right)^2 + \eta^2 \left(S_t S_t^\top \right)^3 \right\| + \eta^2 \sigma_1^2 \left\| S_t S_t^\top \right\| \\
 &\stackrel{(a)}{\leq} \left\| S_t S_t^\top \right\| - 2\eta \left\| S_t S_t^\top \right\|^2 + \eta^2 \left\| S_t S_t^\top \right\|^3 + \eta^2 \sigma_1^2 \left\| S_t S_t^\top \right\|,
 \end{aligned}$$

where (a) follows from the fact that $S_t S_t^\top$, $(S_t S_t^\top)^2$, $(S_t S_t^\top)^3$ share the same eigenvectors, and the assumption $\eta \lesssim 1/\sigma_1$. On the other hand, one can easily verify that

$$\|(A_2)\| \leq 2\eta\sigma_1 \left\| S_t S_t^\top \left(I - \eta S_t S_t^\top \right) \right\| \leq 2\eta\sigma_1 \left(\left\| S_t S_t^\top \right\| - \eta \left\| S_t S_t^\top \right\|^2 \right).$$

Combining the upper bounds for (A) and (B) leads to

$$\begin{aligned} \left\| S_{t+1} S_{t+1}^\top \right\| &\leq \left((1 + \eta\sigma_1)^2 + 3\eta^2\sigma_1 \left\| E_t E_t^\top \right\| \right) \left\| S_t S_t^\top \right\| \\ &\quad - 2\eta \left(1 + \eta\sigma_1 - 1.5\eta \left\| E_t E_t^\top \right\| \right) \left\| S_t S_t^\top \right\|^2 + \eta^2 \left\| S_t S_t^\top \right\|^3 \\ &\stackrel{(a)}{\leq} ((1 + 2.001\sigma_1) \left\| S_t S_t^\top \right\| - 2\eta \left\| S_t S_t^\top \right\|^2 + \eta^2 \left\| S_t S_t^\top \right\|^3), \end{aligned}$$

where, in (a), we used the assumption $\left\| E_t E_t^\top \right\| \lesssim \sigma_1$ and $\eta \lesssim 1/\sigma_1$. Now, let us define the function

$$f(x) := \eta^2 x^3 - 2\eta x^2 + (1 + 2.001\sigma_1) x.$$

It is easy to see that $f(x)$ is increasing within the interval $x \leq 1/4\eta$. On the other hand, we have $1.01\sigma_1 \leq 1/4\eta$ due to our choice of η . Therefore, we have

$$\left\| S_{t+1} S_{t+1}^\top \right\| \leq f \left(\left\| S_t S_t^\top \right\| \right) \leq f(1.01\sigma_1).$$

On the other hand, a simple calculation reveals that

$$f(1.01\sigma_1) = \eta^2(1.01\sigma_1)^3 - 2\eta(1.01\sigma_1)^2 + (1 + 2.001\sigma_1)(1.01\sigma_1) \leq 1.01\sigma_1.$$

for $\eta \leq c/\sigma_1$ with sufficiently small constant c . This completes the proof. \square

C.3 Proof of Lemma 32

We prove this lemma by induction on t . First, due to our choice of the initial point in Theorem 18, we have $\left\| E_0 E_0^\top \right\| \leq \alpha^2$, $\left\| S_0 S_0^\top \right\| \leq 1.01\sigma_1$ and $S_0 S_0^\top \succ 0$. Now, suppose that (30), (31), and (32) hold for t . Then, it is easy to see that

$$\left\| E_t E_t^\top \right\| \leq \left\| E_0 E_0^\top \right\| \leq \alpha^2,$$

due to the decreasing nature of $\left\| E_t E_t^\top \right\|$. Therefore, Proposition 17 can be invoked to show that $\left\| S_{t+1} S_{t+1}^\top \right\| \leq 1.01\sigma_1$. Moreover, we have

$$\begin{aligned} \left\| E_{t+1} E_{t+1}^\top \right\| &\leq \left\| E_t E_t^\top \right\| - \eta \left\| E_t E_t^\top \right\|^2 \\ &\leq \frac{1}{\eta t + \frac{1}{\alpha^2}} - \frac{\eta}{\left(\eta t + \frac{1}{\alpha^2} \right)^2} \\ &\leq \frac{1}{\eta(t+1) + \frac{1}{\alpha^2}}. \end{aligned}$$

On the other hand, Proposition 16 can be used to show that

$$\begin{aligned}\lambda_{\min}\left(S_{t+1}S_{t+1}^\top\right) &\geq (1+\eta^2\sigma_r^2)\lambda_{\min}\left(S_tS_t^\top\right)-2\eta(1+\eta\sigma_r)\lambda_{\min}\left(S_tS_t^\top\right)^2 \\ &\geq (1-2\eta\sigma_1(1+\eta\sigma_r)+\eta^2\sigma_r^2)\lambda_{\min}\left(S_tS_t^\top\right) \\ &> 0,\end{aligned}$$

where in the first inequality, we used $\|E_tE_t^\top\|\lesssim\sigma_r$, which follows from $\|E_tE_t^\top\|\leq\alpha^2$ and our choice of α . Moreover, the last inequality follows from our choice of η . \square

Appendix D. Proofs of Empirical Loss with Noiseless Measurements

D.1 Preliminaries

For simplicity, we define $\Delta_t=U_tU_t^\top-X^*$. Before presenting the proofs for the empirical loss, we first introduce a key decomposition. Recall the update rule

$$U_{t+1}=U_t-2\eta_tQ_tU_t:=\tilde{U}_{t+1}+R_tU_t, \quad (80)$$

where $\eta_t=\frac{\eta}{2}\frac{1}{m}\sum_{i=1}^m|\langle A_i,\Delta_t\rangle|$ and $Q_t\in\mathcal{Q}(\Delta_t)$. Moreover, denote $\tilde{U}_{t+1}=U_t-\bar{\eta}_t\Delta_tU_t$, where $\bar{\eta}_t=\eta\varphi(\Delta_t)^2$ is the update rule obtained from the expected loss, and R_tU_t with $R_t=\bar{\eta}_t\Delta_t-2\eta_tQ_t$ is the residual caused by the deviation of the empirical loss from its expectation. Finally, we define $\tilde{S}_t=V^\top\tilde{U}_t$, and $\tilde{E}_t=V_\perp^\top\tilde{U}_t$.

Lemma 53 *Suppose that the measurements satisfy Sign-RIP with parameters $(4r,\delta,\sqrt{d}\|G_t\|^2)$. Then, we have $\|R_t\|\leq 3\bar{\eta}_t\delta\|\Delta_t\|_F$.*

Proof One can write

$$\Delta_t=V\underbrace{\left(S_tS_t^\top-\Sigma\right)}_{\text{rank-}4r}V^\top+V\underbrace{S_tE_t^\top V_\perp^\top+V_\perp E_tS_t^\top V^\top+V_\perp G_tG_t^\top V_\perp^\top}_{\text{small perturbation}}+V_\perp G_tG_t^\top V_\perp^\top.$$

Note that $\|V_\perp G_tG_t^\top V_\perp^\top\|_F\leq\sqrt{d}\|G_t\|^2$. Therefore, Δ_t is an $\sqrt{d}\|G_t\|^2$ -approximate rank- $4r$ matrix. One can write

$$\begin{aligned}R_t &= \eta\varphi(\Delta_t)^2\Delta_t-2\eta_tQ_t \\ &= \left(\frac{\eta}{2}\varphi(\Delta)\|\Delta_t\|_F-\eta_t\right)Q_t+\left(\frac{\eta}{2}\varphi(\Delta)\|\Delta_t\|_F-\eta_t\right)Q_t^\top \\ &\quad +\frac{\eta}{2}\varphi(\Delta_t)\|\Delta_t\|_F\left(\varphi(\Delta_t)\frac{\Delta_t}{\|\Delta_t\|_F}-Q_t\right) \\ &\quad +\frac{\eta}{2}\varphi(\Delta_t)\|\Delta_t\|_F\left(\varphi(\Delta_t)\frac{\Delta_t}{\|\Delta_t\|_F}-Q_t^\top\right).\end{aligned}$$

Due to the above decomposition, one can write

$$\|R_t\|\leq 2\left|\frac{\eta}{2}\varphi(\Delta)\|\Delta_t\|_F-\eta_t\right\|Q_t\|+\eta\varphi(\Delta_t)\|\Delta_t\|_F\left\|\varphi(X)\frac{\Delta_t}{\|\Delta_t\|_F}-Q_t\right\|. \quad (81)$$

First, note that $|(\eta/2)\varphi(\Delta) \|\Delta_t\|_F - \eta_t| \leq \delta(\eta/2)\varphi(\Delta_t) \|\Delta_t\|_F$ due to Lemma 51. Moreover, due to Sign-RIP, we have $\|Q_t\| \leq (1 + \delta)\varphi(\Delta_t)$ and $\left\| \varphi(X) \frac{\Delta_t}{\|\Delta_t\|_F} - Q_t \right\| \leq \delta\varphi(\Delta_t)$. Combining these upper bounds with (81) completes the proof. \blacksquare

D.2 Proof of Proposition 19

Due to the proposed decomposition (80), one can write

$$S_{t+1} = \tilde{S}_{t+1} + V^\top R_t U_t. \quad (82)$$

Our main goal is to show that the minimum eigenvalue of $S_t S_t^\top$ follows that of $\tilde{S}_t \tilde{S}_t^\top$ —which has been characterized in Lemma 16—plus an additional deviation caused by the term $V^\top R_t U_t$. Similar to the proof of Proposition 16, we characterize the growth rate of $\lambda_{\min}(S_{t+1} S_{t+1}^\top)$ by first resorting to a more tractable lower bound. Adopting the notation introduced in Appendix D, consider the following auxiliary matrix

$$\begin{aligned} M &:= (I + \Xi) \tilde{S}_{t+1} \tilde{S}_{t+1}^\top (I + \Xi)^\top \\ &= (I + \Xi) V^\top (I - \bar{\eta}_t \Delta_t) U_t U_t^\top (I - \bar{\eta}_t \Delta_t) (I + \Xi)^\top, \end{aligned}$$

where $\Xi := V^\top R_t U_t \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1}$. Note that, according to Lemma 32, we have $\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \succ 0$ due to our assumption $S_t S_t^\top \succ 0$. Therefore, $\tilde{S}_{t+1} \tilde{S}_{t+1}^\top$ is invertible, and hence, the matrix Ξ is well-defined. On the other hand, it is easy to see that Ξ satisfies

$$V^\top R_t U_t \tilde{S}_{t+1}^\top = \Xi \tilde{S}_{t+1} \tilde{S}_{t+1}^\top.$$

Based on the above equality, one can write

$$\begin{aligned} S_{t+1} S_{t+1}^\top - M &= \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top + V^\top R_t U_t U_t^\top R_t^\top V + \tilde{S}_{t+1} U_t^\top R_t^\top V + V^\top R_t U_t \tilde{S}_{t+1}^\top \right) \\ &\quad - \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top - \Xi \tilde{S}_{t+1} \tilde{S}_{t+1}^\top \Xi^\top + \Xi \tilde{S}_{t+1} \tilde{S}_{t+1}^\top + \tilde{S}_{t+1} \tilde{S}_{t+1}^\top \Xi^\top \right) \\ &= V^\top R_t U_t U_t^\top R_t^\top V - V^\top R_t U_t \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \tilde{S}_{t+1} U_t^\top R_t^\top V \\ &= V^\top R_t U_t \left(I - \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \tilde{S}_{t+1} \right) U_t^\top R_t^\top V \\ &= V^\top R_t U_t \mathbf{P}_{\tilde{S}_{t+1}}^\perp U_t^\top R_t^\top V \\ &\succeq 0. \end{aligned} \quad (83)$$

Therefore, we have $\lambda_{\min}(S_{t+1} S_{t+1}^\top) \geq \lambda_{\min}(M)$. Our next goal is to provide a lower bound for $\lambda_{\min}(M)$. To this goal, we will show that the minimum eigenvalue of M is close to that of $\tilde{S}_{t+1} \tilde{S}_{t+1}^\top$. Combined with the minimum eigenvalue dynamics of $\tilde{S}_{t+1} \tilde{S}_{t+1}^\top$ presented in Proposition 16, this completes the proof.

To show that $\lambda_{\min}(M) \approx \lambda_{\min}(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top)$, we will use the relative perturbation bound presented in Lemma 52. To this goal, first we need to provide an upper bound for $\|\Xi\|$.

Bounding $\|\Xi\|$. One can write

$$\begin{aligned}
 \|\Xi\| &\leq \|R_t\| \left\| U_t \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\| \\
 &= \|R_t\| \left\| \left(I - \bar{\eta}_t \left(U_t U_t^\top - X^* \right) \right)^{-1} \tilde{U}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\| \\
 &\leq \|R_t\| \left\| \left(I - \bar{\eta}_t \left(U_t U_t^\top - X^* \right) \right)^{-1} \right\| \left\| \tilde{U}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\| \\
 &\stackrel{(a)}{\leq} 2 \|R_t\| \left\| \tilde{U}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\| \\
 &\stackrel{(b)}{\leq} 6\bar{\eta}_t \delta \|\Delta_t\|_F \left\| \tilde{U}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\|,
 \end{aligned} \tag{84}$$

where in (a), we used the fact that $\|U_t U_t^\top - X^*\| \lesssim \sigma_1$ due to our assumptions on $\|S_t S_t^\top\|$ and $\|E_t E_t^\top\|$, and our choice of η . Moreover, (b) follows from Lemma 53. Now, note that

$$\begin{aligned}
 \tilde{U}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} &= V \tilde{S}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} + V_\perp \tilde{E}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \\
 &= V + V_\perp \tilde{E}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1}.
 \end{aligned}$$

Therefore, we have

$$\left\| \tilde{U}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\| \leq 1 + \left\| \tilde{E}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\|.$$

In order to provide an upper bound for $\|\Xi\|$, it suffices to bound $\left\| \tilde{E}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\|$. To this goal, first we present the following technical lemma, the proof of which can be found in Appendix F.2.

Lemma 54 *The following statements hold:*

$$\begin{aligned}
 &- \frac{1}{2} \leq \left\| S_t S_t^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\| \leq 2 \text{ and } \frac{1}{3} \leq \left\| S_t S_t^\top \left(S_{t+1} S_{t+1}^\top \right)^{-1} \right\| \leq 3; \\
 &- \left\| \tilde{E}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\| \leq 3 \left\| E_t S_t^\top \left(S_t S_t^\top \right)^{-1} \right\|.
 \end{aligned}$$

The second statement of the above lemma connects $\left\| \tilde{E}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\|$ to $\left\| E_t S_t^\top \left(S_t S_t^\top \right)^{-1} \right\|$ which, according to our assumption, is upper bounded by 1/3. This upper bound, together with (84) implies that

$$\|\Xi\| \leq 12\bar{\eta}_t \delta \|\Delta_t\|_F.$$

On the other hand, applying the relative perturbation bound to M and $\tilde{S}_{t+1}\tilde{S}_{t+1}^\top$ implies that

$$\begin{aligned} \left| \lambda_{\min}(M) - \lambda_{\min}(\tilde{S}_{t+1}\tilde{S}_{t+1}^\top) \right| &\leq \lambda_{\min}(\tilde{S}_{t+1}\tilde{S}_{t+1}^\top) \left\| I - (I + \Xi)(I + \Xi)^\top \right\| \\ &\leq 3 \|\Xi\| \lambda_{\min}(\tilde{S}_{t+1}\tilde{S}_{t+1}^\top) \\ &\leq 36\bar{\eta}_t\delta \|\Delta_t\|_F \lambda_{\min}(\tilde{S}_{t+1}\tilde{S}_{t+1}^\top). \end{aligned}$$

This in turn implies that

$$\begin{aligned} \lambda_{\min}(S_{t+1}S_{t+1}^\top) &\geq \lambda_{\min}(M) \\ &\geq (1 - 36\bar{\eta}_t\delta \|\Delta_t\|_F) \lambda_{\min}(\tilde{S}_{t+1}\tilde{S}_{t+1}^\top) \\ &\geq \left((1 + \bar{\eta}_t\sigma_r)^2 - 2\bar{\eta}_t \left\| E_t E_t^\top \right\| \right) \lambda_{\min}(S_t S_t^\top) \\ &\quad - 72\bar{\eta}_t\delta \left\| U_t U_t^\top - X^* \right\|_F \lambda_{\min}(S_t S_t^\top) \\ &\quad - 2\bar{\eta}_t (1 + \bar{\eta}_t\sigma_r) \lambda_{\min}(S_t S_t^\top)^2, \end{aligned}$$

which completes the proof. \square

D.3 Proof of Proposition 20

Bounding $\|\Sigma - S_{t+1}S_{t+1}^\top\|$: Recall that $S_{t+1} = \tilde{S}_{t+1} + V^\top R_t U_t$. Therefore, we have

$$\Sigma - S_{t+1}S_{t+1}^\top = \Sigma - \tilde{S}_{t+1}\tilde{S}_{t+1}^\top - \underbrace{V^\top R_t U_t \tilde{S}_{t+1}^\top - \tilde{S}_{t+1} U_t^\top R_t^\top V - V^\top R_t U_t U_t^\top R_t^\top V}_{\text{deviation}}. \quad (85)$$

The main idea behind our proof is to first control the norm of the deviation term, and then invoke Proposition 17 for $\Sigma - \tilde{S}_{t+1}\tilde{S}_{t+1}^\top$. One can write

$$\left\| V^\top R_t U_t \tilde{S}_{t+1}^\top \right\| \leq \|R_t\| \left\| \tilde{S}_{t+1} U_t^\top \right\| \leq 3\bar{\eta}_t\delta \|\Delta_t\|_F \left\| \tilde{S}_{t+1} U_t^\top \right\|,$$

where in the last inequality, we used Lemma 53. Moreover, $\tilde{S}_{t+1} U_t^\top$ can be rewritten as

$$\begin{aligned} \tilde{S}_{t+1} U_t^\top &= \left(\left(I - \bar{\eta}_t (S_t S_t^\top - \Sigma) \right) S_t + S_t E_t^\top E_t \right) \left(S^\top V^\top + E^\top V_\perp^\top \right) \\ &= \left(I - \bar{\eta}_t (S_t S_t^\top - \Sigma) \right) S_t S_t^\top V^\top + \left(I - \bar{\eta}_t (S_t S_t^\top - \Sigma) \right) S_t E_t^\top V_\perp^\top \\ &\quad + S_t E_t^\top E_t S_t^\top V^\top + S_t E_t^\top E_t E_t^\top V_\perp^\top. \end{aligned}$$

Note that $\left\| I - \bar{\eta}_t (S_t S_t^\top - \Sigma) \right\| \leq 2$ due to our choice of η and our assumption on $S_t S_t^\top$. Therefore, we have

$$\left\| \tilde{S}_{t+1} U_t^\top \right\| \leq 2 \left\| S_t S_t^\top \right\| + 2 \left\| S_t E_t^\top \right\| + \left\| S_t E_t^\top \right\|^2 + \left\| S_t E_t^\top \right\| \|E_t\|^2 \leq 6\sigma_1, \quad (86)$$

where we used our assumptions on $\|S_t S_t^\top\|$ and $\|E_t E_t^\top\|$. Combining the above two inequalities leads to

$$\left\| \tilde{S}_{t+1} U_t^\top R_t^\top V \right\| = \left\| V^\top R_t U_t \tilde{S}_{t+1}^\top \right\| \leq 18 \bar{\eta}_t \delta \sigma_1 \|\Delta_t\|_F. \quad (87)$$

Using a similar technique, we have

$$\left\| V^\top R_t U_t U_t^\top R_t^\top V \right\| \leq 9 \bar{\eta}_t^2 \delta^2 \sigma_1 \|\Delta_t\|_F^2 \leq \bar{\eta}_t \delta \sigma_1 \|\Delta_t\|_F, \quad (88)$$

where we used the assumed upper bounds on η and δ , and the fact that $\|E_t E_t^\top\| \leq \sigma_1$, $\|S_t S_t^\top\| \leq 1.01 \sigma_1$, and $\|E_t E_t^\top\|_F \leq \sqrt{r} \sigma_1$, which in turn implies $\|\Delta_t\|_F \lesssim \sqrt{r} \sigma_1$ due to Lemma 14. Moreover, we have already shown in Proposition 17 that

$$\left\| \Sigma - \tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right\| \leq \left(1 - \bar{\eta}_t \lambda_{\min} \left(S_t S_t^\top \right) \right) \left\| \Sigma - S_t S_t^\top \right\| + 5 \bar{\eta}_t \left\| S_t E_t^\top \right\|^2.$$

Hence, combining the above inequalities leads to

$$\begin{aligned} \left\| \Sigma - S_{t+1} S_{t+1}^\top \right\| &\leq \left(1 - \bar{\eta}_t \lambda_{\min} \left(S_t S_t^\top \right) \right) \left\| \Sigma - S_t S_t^\top \right\| \\ &\quad + 5 \bar{\eta}_t \left\| S_t E_t^\top \right\|^2 + 37 \bar{\eta}_t \delta \sigma_1 \|\Delta_t\|_F. \end{aligned}$$

Bounding $\|S_{t+1} E_{t+1}^\top\|$: First, note that

$$S_{t+1} E_{t+1}^\top = \tilde{S}_{t+1} \tilde{E}_{t+1}^\top + \underbrace{V^\top R_t U_t \tilde{E}_{t+1}^\top + \tilde{S}_{t+1} U_t^\top R_t^\top V_\perp + V^\top R_t U_t U_t^\top R_t^\top V_\perp}_{\text{deviation}}.$$

Similar to the signal term, the main idea behind our proof is to first control the norm of the deviation term, and then invoke Proposition 17 for $\tilde{S}_{t+1} \tilde{E}_{t+1}^\top$. We first provide an upper bound for $\left\| V^\top R_t U_t \tilde{E}_{t+1}^\top \right\|$

$$\left\| V^\top R_t U_t \tilde{E}_{t+1}^\top \right\| \leq \|R_t\| \left\| U_t \tilde{E}_{t+1}^\top \right\| \leq 3 \bar{\eta}_t \delta \|\Delta_t\|_F \left\| U_t \tilde{E}_{t+1}^\top \right\|.$$

To bound $\left\| U_t \tilde{E}_{t+1}^\top \right\|$, one can write

$$\begin{aligned} \left\| U_t \tilde{E}_{t+1}^\top \right\| &\leq \|U_t\| \left\| E_t - \bar{\eta}_t S_t^\top S_t E_t^\top - \bar{\eta}_t E_t E_t^\top E_t \right\| \\ &\leq 2 (\|S_t\| + \|E_t\|) \|E_t\| \\ &\leq 5 \sigma_1, \end{aligned}$$

where we used the assumption $\|S_t\| \leq 1.01 \sqrt{\sigma_1}$ and $\|E_t\| \leq \sqrt{\sigma_1}$. Similar to (88), we have

$$\begin{aligned} \left\| \tilde{S}_{t+1} U_t^\top R_t^\top V_\perp \right\| &\leq \left\| \tilde{S}_{t+1} U_t^\top \right\| \|R_t\| \\ &\leq \left\| S_t - \bar{\eta}_t \left(\left(S_t S_t^\top - \Sigma \right) S_t + S_t E_t^\top E_t \right) \right\| \|R_t\| \\ &\leq 6 \bar{\eta}_t \delta \sqrt{\sigma_1} \|\Delta_t\|_F, \end{aligned}$$

and

$$\left\| V^\top R_t U_t U_t^\top R_t^\top V_\perp \right\| \leq \left\| U_t U_t^\top \right\| \left\| R_t \right\|^2 \leq 10\sigma_1 \bar{\eta}_t^2 \delta^2 \|\Delta_t\|_F^2 \leq \bar{\eta}_t \delta \sigma_1 \|\Delta_t\|_F.$$

Moreover, we have already shown in Proposition 17 that

$$\left\| \tilde{S}_{t+1} \tilde{E}_{t+1}^\top \right\| \leq \left(1 - \bar{\eta}_t \lambda_{\min} \left(S_t S_t^\top \right) + 2\bar{\eta}_t \left\| \Sigma - S_t S_t^\top \right\| + 2\bar{\eta}_t \|E_t E_t\| \right) \left\| S_t E_t^\top \right\|. \quad (89)$$

Combining the above inequalities leads to

$$\begin{aligned} \left\| S_{t+1} E_{t+1}^\top \right\| &\leq \left(1 - \bar{\eta}_t \lambda_{\min} \left(S_t S_t^\top \right) + 2\bar{\eta}_t \left\| \Sigma - S_t S_t^\top \right\| + 2\bar{\eta}_t \|E_t E_t\| \right) \left\| S_t E_t^\top \right\| \\ &\quad + 22\bar{\eta}_t \delta \sigma_1 \|\Delta_t\|_F. \end{aligned}$$

Bounding $\|S_{t+1} S_{t+1}^\top\|$: First, note that

$$S_{t+1} S_{t+1}^\top = \tilde{S}_{t+1} \tilde{S}_{t+1}^\top + \underbrace{V^\top R_t U_t S_{t+1}^\top + S_{t+1} U_t^\top R_t^\top V + V^\top R_t U_t U_t^\top R_t^\top V}_{\text{deviation}}.$$

Similar to our previous arguments, we will provide an upper bound on the deviation term, and then resort to Proposition 17 to provide an upper bound for $\tilde{S}_{t+1} \tilde{S}_{t+1}^\top$. First, note that

$$\begin{aligned} \left\| V^\top R_t U_t S_{t+1}^\top \right\| &\leq \left\| V^\top R_t V S_t S_{t+1}^\top \right\| + \left\| V^\top R_t V_\perp E_t S_{t+1}^\top \right\| \\ &\leq 3\bar{\eta}_t \delta \|\Delta_t\|_F \left(\left\| S_t S_{t+1}^\top \right\| + \left\| E_t S_{t+1}^\top \right\| \right), \end{aligned} \quad (90)$$

where we used Lemma 53 in the last inequality. On the other hand

$$\begin{aligned} \left\| S_t S_{t+1}^\top \right\| &= \left\| S_t \left(S_t^\top - \bar{\eta}_t \left(S_t^\top \left(S_t S_t^\top - \Sigma \right) + E_t^\top E_t S_t^\top \right) + U_t^\top R_t^\top V \right) \right\| \\ &\leq \left\| S_t S_t^\top \right\| \left(1 + \bar{\eta}_t \left\| S_t S_t^\top - \Sigma \right\| + \bar{\eta}_t \left\| E_t E_t^\top \right\| \right) + \left\| S_t U_t^\top R_t^\top V \right\| \\ &\leq 2 \left\| S_t S_t^\top \right\| + \left\| S_t S_t^\top V^\top R_t^\top V \right\| + \left\| S_t E_t^\top V_\perp^\top R_t^\top V \right\| \\ &\leq 3 \left\| S_t S_t^\top \right\| + 3\bar{\eta}_t \delta \|\Delta_t\|_F \left\| S_t E_t^\top \right\|. \end{aligned}$$

Similarly, we have

$$\left\| E_t S_{t+1}^\top \right\| \leq 3 \left\| E_t S_t^\top \right\| + 3\bar{\eta}_t \delta \|\Delta_t\|_F \left\| E_t E_t^\top \right\|.$$

The above two inequalities combined with (90) results in

$$2 \left\| V^\top R_t U_t S_{t+1}^\top \right\| \leq 216\bar{\eta}_t \sigma_1^2 \sqrt{r} \delta.$$

which follows from our assumption on $\|S_t\|$, $\|E_t\|$ and δ . Similarly, one can show that

$$\left\| V^\top R_t U_t U_t^\top R_t^\top V \right\| \leq \bar{\eta}_t \sigma_1^2 \sqrt{r} \delta.$$

Therefore, the norm of the deviation term is upper bounded by $217\bar{\eta}_t\sigma_1^2\sqrt{r}\delta$. Moreover, Proposition 17 implies

$$\left\|\tilde{S}_{t+1}\tilde{S}_{t+1}^\top\right\|\leq(1+2\sigma_1\bar{\eta}_t+4\sigma_1^2\bar{\eta}_t^2)\left\|S_tS_t^\top\right\|-2\bar{\eta}_t\left\|S_tS_t^\top\right\|^2+\bar{\eta}_t^2\left\|S_tS_t^\top\right\|^3,$$

which in turn leads to

$$\left\|S_{t+1}S_{t+1}^\top\right\|\leq(1+2\sigma_1\bar{\eta}_t+4\sigma_1^2\bar{\eta}_t^2)\left\|S_tS_t^\top\right\|-2\bar{\eta}_t\left\|S_tS_t^\top\right\|^2+\bar{\eta}_t^2\left\|S_tS_t^\top\right\|^3+217\bar{\eta}_t\sigma_1^2\sqrt{r}\delta.$$

The rest of the proof is analogous to that of Proposition 17, and hence, omitted for brevity. \square

D.4 Proof of Proposition 21

Bounding $\|G_{t+1}\|$: To provide an upper bound for $\|G_{t+1}\| = \left\|E_{t+1}P_{\tilde{S}_{t+1}}^\perp\right\|$ in terms of $\|G_t\| = \left\|E_tP_{\tilde{S}_t}^\perp\right\|$, it is crucial to characterize the relationship between the projection operators $P_{\tilde{S}_{t+1}}^\perp$ and $P_{\tilde{S}_t}^\perp$. To this goal, we decompose $P_{\tilde{S}_{t+1}}^\perp$ as follows

$$P_{\tilde{S}_{t+1}}^\perp = P_{S_t}P_{\tilde{S}_{t+1}}^\perp + P_{\tilde{S}_t}^\perp P_{\tilde{S}_{t+1}}^\perp.$$

Based on the above decomposition, G_{t+1} can be written as

$$G_{t+1} = \underbrace{E_{t+1}P_{S_t}P_{\tilde{S}_{t+1}}^\perp}_{(A)} + \underbrace{E_{t+1}P_{\tilde{S}_t}^\perp P_{\tilde{S}_{t+1}}^\perp}_{(B)}.$$

We first study (A). Let $M_tD_tN_t^\top$ be the singular value decomposition of S_t , where $M_t \in \mathbb{R}^{r \times r}$ and $N_t \in \mathbb{R}^{r' \times r}$ are (row) orthonormal matrices, and $D_t \in \mathbb{R}^{r \times r}$ is a diagonal matrix collecting the singular values of S_t . Based on this definition, we have $P_{S_t} = N_tN_t^\top$. On the other hand, we have $S_{t+1}P_{S_t}P_{\tilde{S}_{t+1}}^\perp = -S_{t+1}P_{\tilde{S}_t}^\perp P_{\tilde{S}_{t+1}}^\perp$, which is equivalent to

$$S_{t+1}N_tN_t^\top P_{\tilde{S}_{t+1}}^\perp = -S_{t+1}P_{\tilde{S}_t}^\perp P_{\tilde{S}_{t+1}}^\perp. \quad (91)$$

Our next technical lemma shows that $S_{t+1}N_t$ is invertible.

Lemma 55 *The matrix $S_{t+1}N_t$ is invertible.*

The proof of this lemma can be found in Appendix F.3. Lemma (55) combined with (91) leads to

$$N_t^\top P_{\tilde{S}_{t+1}}^\perp = -(S_{t+1}N_t)^{-1}S_{t+1}P_{\tilde{S}_t}^\perp P_{\tilde{S}_{t+1}}^\perp.$$

Therefore, we have

$$\begin{aligned} (A) &= E_{t+1}P_{S_t}P_{\tilde{S}_{t+1}}^\perp \\ &= E_{t+1}N_tN_t^\top P_{\tilde{S}_{t+1}}^\perp \\ &= -E_{t+1}N_t(S_{t+1}N_t)^{-1}S_{t+1}P_{\tilde{S}_t}^\perp P_{\tilde{S}_{t+1}}^\perp \\ &= \underbrace{-E_{t+1}N_t(S_{t+1}N_t)^{-1}\tilde{S}_{t+1}P_{\tilde{S}_t}^\perp P_{\tilde{S}_{t+1}}^\perp}_{(A_1)} \\ &\quad - \underbrace{E_{t+1}N_t(S_{t+1}N_t)^{-1}\left(S_{t+1}-\tilde{S}_{t+1}\right)P_{\tilde{S}_t}^\perp P_{\tilde{S}_{t+1}}^\perp}_{(A_2)}. \end{aligned}$$

We first control (A_1) . Observe that

$$\tilde{S}_{t+1} \mathbf{P}_{\tilde{S}_t}^\perp = \left(S_t - \bar{\eta}_t \left(S_t S_t^\top - \Sigma \right) S_t - \bar{\eta}_t S_t E_t^\top E_t \right) \mathbf{P}_{\tilde{S}_t}^\perp = -\bar{\eta}_t S_t E_t^\top G_t.$$

Therefore, we have

$$(A_1) = \bar{\eta}_t E_{t+1} N_t (S_{t+1} N_t)^{-1} S_t E_t^\top G_t \mathbf{P}_{\tilde{S}_{t+1}}^\perp. \quad (92)$$

On the other hand, note that $E_{t+1} = E_t - \bar{\eta}_t E_t (S_t^\top S_t + E_t^\top E_t) + V_\perp^\top R_t U_t$. Hence, we have

$$(B) = \left(I - \bar{\eta}_t E_t E_t^\top \right) G_t \mathbf{P}_{\tilde{S}_{t+1}}^\perp + V_\perp^\top R_t V_\perp G_t \mathbf{P}_{\tilde{S}_{t+1}}^\perp. \quad (93)$$

Combining equations (92) and (93), we obtain

$$G_{t+1} = \left(I - \bar{\eta}_t E_t E_t^\top + \bar{\eta}_t E_{t+1} N_t (S_{t+1} N_t)^{-1} S_t E_t^\top \right) G_t \mathbf{P}_{\tilde{S}_{t+1}}^\perp + V_\perp^\top R_t V_\perp G_t \mathbf{P}_{\tilde{S}_{t+1}}^\perp + (A_2),$$

which results in

$$\|G_{t+1}\| \leq \underbrace{\left\| I - \bar{\eta}_t E_t E_t^\top + \bar{\eta}_t E_{t+1} N_t (S_{t+1} N_t)^{-1} S_t E_t^\top \right\|}_{(C)} \|G_t\| + \|R_t\| \|G_t\| + \|(A_2)\|.$$

Therefore, it remains to control the terms $\|(C)\|$ and $\|(A_2)\|$. First, we provide an upper bound on $\|(C)\|$. Define $S_{t+1} N_t = (I + \Xi) S_t N_t$, where $\Xi = S_{t+1} N_t (S_t N_t)^{-1} - I$ (note that $S_t N_t = M_t D_t$ which implies that $S_t N_t$ is invertible). Hence, we have

$$\begin{aligned} \|(C)\| &\leq \underbrace{\left\| I - \bar{\eta}_t E_t E_t^\top + \bar{\eta}_t E_{t+1} N_t (S_t N_t)^{-1} S_t E_t^\top \right\|}_{(C_1)} \\ &\quad + \underbrace{\bar{\eta}_t \left\| E_{t+1} N_t (S_t N_t)^{-1} \right\| \left\| (I + \Xi)^{-1} - I \right\| \left\| S_t E_t^\top \right\|}_{(C_2)}. \end{aligned}$$

It is shown in the proof of Lemma 55 that $\|\Xi\| \leq 3\bar{\eta}_t \|\Delta_t\| \leq 1/2$. Therefore, one can write $\|(I + \Xi)^{-1} - I\| \leq \|\Xi\| \|(I + \Xi)^{-1}\| \leq 6\bar{\eta}_t \|\Delta_t\|$. To provide an upper bound for $\left\| E_{t+1} N_t (S_t N_t)^{-1} \right\|$, one can write

$$\begin{aligned} E_{t+1} N_t (S_t N_t)^{-1} &= E_{t+1} S_t^\top \left(S_t S_t^\top \right)^{-1} \\ &= \left(E_t - \bar{\eta}_t E_t \left(S_t^\top S_t + E_t^\top E_t \right) + V_\perp^\top R_t U_t \right) S_t^\top \left(S_t S_t^\top \right)^{-1} \\ &= \left(I - \bar{\eta}_t E_t E_t^\top \right) H_t - \bar{\eta}_t E_t S_t^\top + V_\perp^\top R_t U_t S_t^\top \left(S_t S_t^\top \right)^{-1} \\ &= \left(I - \bar{\eta}_t E_t E_t^\top \right) H_t - \bar{\eta}_t E_t S_t^\top + V_\perp^\top R_t V + V_\perp^\top R_t V_\perp H_t, \end{aligned}$$

where $H_t = E_t S_t^\top (S_t S_t^\top)^{-1}$. Therefore, we have

$$\begin{aligned} \left\| E_{t+1} N_t (S_t N_t)^{-1} \right\| &\leq \|H_t\| + \bar{\eta}_t \left\| E_t S_t^\top \right\| + \|R_t\| (1 + \|H_t\|) \\ &\leq \|H_t\| + \bar{\eta}_t \left\| E_t S_t^\top \right\| + 4\bar{\eta}_t \delta \|\Delta_t\|_F. \end{aligned}$$

Hence, we obtain

$$\|C_2\| \leq 6\bar{\eta}_t^2 \|\Delta_t\| \left(\|H_t\| + \bar{\eta}_t \left\| E_t S_t^\top \right\| + \bar{\eta}_t \delta \|\Delta_t\|_F \right) \left\| E_t S_t^\top \right\|.$$

To control $\|(C_1)\|$, we use triangle inequality to arrive at the following decomposition

$$\begin{aligned} (C_1) &\leq \underbrace{\left\| I - \bar{\eta}_t E_t E_t^\top + \bar{\eta}_t E_t N_t (S_t N_t)^{-1} S_t E_t^\top \right\|}_{(C_{11})} + \underbrace{\bar{\eta}_t^2 \left\| S_t S_t^\top E_t S_t^\top (S_t S_t^\top)^{-1} S_t E_t^\top \right\|}_{(C_{12})} \\ &\quad + \underbrace{\bar{\eta}_t^2 \left\| E_t E_t^\top E_t S_t^\top (S_t S_t^\top)^{-1} S_t E_t^\top \right\|}_{(C_{13})} + \underbrace{\bar{\eta}_t \left\| R_t U_t N_t (S_t N_t)^{-1} S_t E_t^\top \right\|}_{(C_{14})}. \end{aligned}$$

It is easy to see that $\|(C_{11})\| = \|I - \bar{\eta}_t G_t G_t^\top\| \leq 1$, due to the assumed upper bound on $\bar{\eta}_t$ and $\|E_t E_t^\top\|$. Moreover, one can verify that $\|(C_{12})\| \leq \bar{\eta}_t^2 \|E_t S_t^\top\|^2$, $\|(C_{13})\| \leq \bar{\eta}_t^2 \|E_t\|^4$, and $\|(C_{14})\| \leq 3\bar{\eta}_t^2 \delta \|\Delta_t\|_F (1 + \|H_t\|) \|S_t E_t^\top\| \leq 6\bar{\eta}_t^2 \delta \|\Delta_t\|_F \|S_t E_t^\top\|$. Combining the derived upper bounds for $\|C_1\|$ and $\|C_2\|$, we have

$$(C) \leq 1 + \bar{\eta}_t^2 \left(2 \left\| E_t S_t^\top \right\|^2 + \|E_t\|^4 + 7\delta \|\Delta_t\|_F \left\| E_t S_t^\top \right\| + 6 \|H_t\| \|\Delta_t\| \left\| E_t S_t^\top \right\| \right).$$

To complete the proof, it remains to provide an upper bound for $\|(A_2)\|$. First, note that $(S_{t+1} - \tilde{S}_{t+1}) P_{S_t}^\perp = V^\top R_t V_\perp G_t$. Given this equality, one can write

$$\begin{aligned} \|A_2\| &\leq \left\| E_{t+1} N_t (S_{t+1} N_t)^{-1} V^\top R_t V_\perp G_t P_{S_{t+1}}^\perp \right\| \\ &\leq \left\| E_{t+1} N_t (S_{t+1} N_t)^{-1} \right\| \|R_t\| \|G_t\| \\ &\leq 3\bar{\eta}_t \delta \left\| E_{t+1} N_t (S_{t+1} N_t)^{-1} \right\| \|\Delta_t\|_F \|G_t\|. \end{aligned}$$

Therefore, it suffices to provide an upper bound for $\left\| E_{t+1} N_t (S_{t+1} N_t)^{-1} \right\|$:

$$\begin{aligned} \left\| E_{t+1} N_t (S_{t+1} N_t)^{-1} \right\| &\leq \left\| E_{t+1} N_t (S_t N_t)^{-1} \right\| \left\| S_t N_t (S_{t+1} N_t)^{-1} \right\| \\ &\leq \left\| E_{t+1} N_t (S_t N_t)^{-1} \right\| \|(I + \Xi)^{-1}\| \\ &\leq 2 \left\| E_{t+1} N_t (S_t N_t)^{-1} \right\| \\ &\leq 2. \end{aligned}$$

Combining the above inequalities leads to

$$\|(A_2)\| \leq 6\bar{\eta}_t \delta \|\Delta_t\|_F \|G_t\|.$$

Finally, combining the derived upper bounds for (C) and (A₂) gives rise to the following inequalities

$$\begin{aligned} \|G_{t+1}\| &\leq \left(1 + 2\bar{\eta}_t^2 \left\| E_t S_t^\top \right\|^2 + \bar{\eta}_t^2 \|E_t\|^4 + 6\bar{\eta}_t^2 \|H_t\| \|\Delta_t\| \left\| E_t S_t^\top \right\| + 7\bar{\eta}_t \delta \|\Delta_t\|_F \right) \|G_t\| \\ &\leq \left(1 + 2\bar{\eta}_t^2 \left\| E_t S_t^\top \right\|^2 + \bar{\eta}_t^2 \|E_t\|^4 + 2\bar{\eta}_t^2 \|\Delta_t\| \left\| E_t S_t^\top \right\| + 7\bar{\eta}_t \delta \|\Delta_t\|_F \right) \|G_t\|. \end{aligned}$$

Bounding $\|F_{t+1}\|$: Similar to the previous part, we use the decomposition $\mathbf{P}_{S_{t+1}} = \mathbf{P}_{S_t} \mathbf{P}_{S_{t+1}} + \mathbf{P}_{S_t}^\perp \mathbf{P}_{S_{t+1}}$ to write

$$F_{t+1} = E_{t+1} \mathbf{P}_{S_{t+1}} = \underbrace{E_{t+1} \mathbf{P}_{S_t} \mathbf{P}_{S_{t+1}}}_{(A)} + \underbrace{E_{t+1} \mathbf{P}_{S_t}^\perp \mathbf{P}_{S_{t+1}}}_{(B)}.$$

First, we provide an upper bound for $\|(B)\|$:

$$\begin{aligned} \|(B)\| &= \left\| E_{t+1} \mathbf{P}_{S_t}^\perp (\mathbf{P}_{S_{t+1}} - \mathbf{P}_{S_t}) \right\| \\ &\leq \left\| E_{t+1} \mathbf{P}_{S_t}^\perp \right\| \left\| \mathbf{P}_{S_{t+1}} - \mathbf{P}_{S_t} \right\| \\ &= \left\| \left(I - \bar{\eta}_t E_t E_t^\top \right) G_t + V_\perp^\top R_t U_t \mathbf{P}_{S_t}^\perp \right\| \left\| \mathbf{P}_{S_{t+1}} - \mathbf{P}_{S_t} \right\| \\ &\leq \left(\|G_t\| + \left\| V_\perp^\top R_t V_\perp G_t \right\| \right) \left\| \mathbf{P}_{S_{t+1}} - \mathbf{P}_{S_t} \right\| \\ &\leq (\|G_t\| + 3\bar{\eta}_t \delta \|\Delta_t\|_F \|G_t\|) \left\| \mathbf{P}_{S_{t+1}} - \mathbf{P}_{S_t} \right\| \\ &\leq 2 \|G_t\| \left\| \mathbf{P}_{S_{t+1}} - \mathbf{P}_{S_t} \right\|. \end{aligned} \tag{94}$$

To control $\|\mathbf{P}_{S_{t+1}} - \mathbf{P}_{S_t}\|$, we use the following technical lemma.

Lemma 56 (Theorem 2.4, Chen et al. (2016)) *Let $A \in \mathbb{R}^{m \times n}$, and $B = A + E \in \mathbb{R}^{m \times n}$ have the same rank. Then, we have*

$$\|\mathbf{P}_A - \mathbf{P}_B\| \leq \left\| EA^\dagger \right\| \vee \left\| EB^\dagger \right\|.$$

Due to Proposition 19 and our assumptions, we have $\lambda_{\min}(S_{t+1} S_{t+1}^\top) \geq \lambda_{\min}(S_t S_t^\top) > 0$, and hence, both $S_{t+1} S_{t+1}^\top$ and $S_t S_t^\top$ are rank- r . Invoking Lemma 56, we have

$$\begin{aligned} \|\mathbf{P}_{S_{t+1}} - \mathbf{P}_{S_t}\| &\leq \left\| (S_{t+1} - S_t) S_t^\top \left(S_t S_t^\top \right)^{-1} \right\| \\ &= \left\| \left(-\bar{\eta}_t \left((S_t S_t^\top - \Sigma) S_t + S_t E_t^\top E_t \right) + V^\top R_t U_t \right) S_t^\top \left(S_t S_t^\top \right)^{-1} \right\| \\ &\leq \bar{\eta}_t \left\| S_t S_t^\top - \Sigma \right\| + \bar{\eta}_t \left\| S_t E_t^\top \right\| \|H_t\| + \|R_t\| (1 + \|H_t\|) \\ &\leq \bar{\eta}_t \left\| S_t S_t^\top - \Sigma \right\| + \bar{\eta}_t \left\| S_t E_t^\top \right\| + 6\bar{\eta}_t \delta \|\Delta_t\|_F \\ &\leq 3\bar{\eta}_t \|\Delta_t\|, \end{aligned} \tag{95}$$

where in the last inequality, we used the following auxiliary lemma.

Lemma 57 *If $\|\Delta_t\| \geq \sqrt{d} \|G_t\|^2$, then we have $\|\Delta_t\|_F \leq 5\sqrt{r} \|\Delta_t\|$.*

Proof Recall the signal-residual decomposition

$$\Delta_t = V \left(S_t S_t^\top - \Sigma \right) V^\top + V S_t E_t^\top V_\perp^\top + V_\perp E_t S_t^\top V^\top + V_\perp F_t F_t^\top V_\perp^\top + V_\perp G_t G_t^\top V_\perp^\top.$$

One can write

$$\begin{aligned}\|\Delta_t\|_F &\leq \sqrt{r} \left(\|S_t S_t^\top - \Sigma\| + 2 \|S_t E_t^\top\| + \|F_t F_t^\top\| \right) + \sqrt{d} \|G_t\|^2 \\ &= 4\sqrt{r} \|\Delta_t\| + \|\Delta_t\| \\ &\leq 5\sqrt{r} \|\Delta_t\|,\end{aligned}$$

which completes the proof. \blacksquare

Combining (95) with (94) leads to

$$\|(B)\| \leq 6\bar{\eta}_t \|\Delta_t\| \|G_t\|.$$

Next, we will provide an upper bound for (A). One can write

$$\begin{aligned}\|(A)\| &\leq \|E_{t+1} \mathbf{P}_{S_t}\| \\ &\leq \left\| \left(E_t - \bar{\eta}_t E_t \left(E_t^\top E_t + S_t^\top S_t \right) + V_\perp^\top R_t U_t \right) \mathbf{P}_{S_t} \right\| \\ &\leq \left\| F_t - \bar{\eta}_t E_t E_t^\top F_t - \bar{\eta}_t F_t S_t^\top S_t \right\| + \left\| V_\perp^\top R_t U_t \mathbf{P}_{S_t} \right\|.\end{aligned}\tag{96}$$

The first term in the above inequality can be bounded as follows

$$\begin{aligned}\left\| F_t - \bar{\eta}_t E_t E_t^\top F_t - \bar{\eta}_t F_t S_t^\top S_t \right\| &\leq \left\| \left(0.5I - \bar{\eta}_t E_t E_t^\top \right) F_t \right\| + \left\| F_t \left(0.5I - \bar{\eta}_t S_t^\top S_t \right) \right\| \\ &\leq \left(\left\| 0.5I - \bar{\eta}_t E_t E_t^\top \right\| + \left\| 0.5I - \bar{\eta}_t S_t^\top S_t \right\| \right) \|F_t\| \\ &\leq \left(1 - \bar{\eta}_t \lambda_{\min} \left(S_t S_t^\top \right) \right) \|F_t\|.\end{aligned}$$

Moreover, we have

$$\left\| V_\perp^\top R_t U_t \mathbf{P}_{S_t} \right\| \leq \|R_t\| (\|S_t\| + \|F_t\|) \leq 3\bar{\eta}_t \delta \|\Delta_t\|_F (\|S_t\| + \|F_t\|).$$

Therefore, we have

$$\|(A)\| \leq \left(1 - \bar{\eta}_t \lambda_{\min} \left(S_t S_t^\top \right) + 3\bar{\eta}_t \delta \|\Delta_t\|_F \right) \|F_t\| + 3\bar{\eta}_t \delta \|\Delta_t\|_F \|S_t\|.$$

Finally, combining the derived upper bounds for (A) and (B) leads to

$$\|F_{t+1}\| \leq \left(1 - \bar{\eta}_t \lambda_{\min} \left(S_t S_t^\top \right) + 3\bar{\eta}_t \delta \|\Delta_t\|_F \right) \|F_t\| + 3\bar{\eta}_t \delta \|\Delta_t\|_F \|S_t\| + 6\bar{\eta}_t \|\Delta_t\| \|G_t\|,$$

which completes the proof. \square

D.5 Proof of Lemma 33

Recall that $U_0 = \alpha B$, where BB^\top is the best rank- r' approximation of $C \in \frac{1}{2m} \sum_{i=1}^m \text{Sign}(y_i) (A_i + A_i^\top)$. Since $\text{rank}(X^*) = r$, Sign-RIP implies

$$\left\| C - \varphi(X^*) \frac{X^*}{\|X^*\|_F} \right\| \leq \varphi(X^*) \delta.$$

Note that BB^\top is the best rank- r' approximation of C . We have

$$\begin{aligned} \left\| BB^\top - \varphi(X^*) \frac{X^*}{\|X^*\|_F} \right\| &\leq \left\| BB^\top - C \right\| + \left\| C - \varphi(X^*) \frac{X^*}{\|X^*\|_F} \right\| \\ &\leq |\lambda_{r'+1}(C)| + \varphi(X^*)\delta \\ &\leq \left\| C - \varphi(X^*) \frac{X^*}{\|X^*\|_F} \right\| + \varphi(X^*)\delta \\ &\leq 2\varphi(X^*)\delta. \end{aligned}$$

Therefore, based on the definition of U_0 , we have

$$\left\| U_0 U_0^\top - \alpha^2 \varphi(X^*) \frac{X^*}{\|X^*\|_F} \right\| \leq 2\alpha^2 \varphi(X^*)\delta.$$

Given this bound, one can write

$$\begin{aligned} \left\| S_0 S_0^\top - \alpha^2 \varphi(X^*) \frac{\Sigma}{\|X^*\|_F} \right\| &= \left\| V^\top \left(U_0 U_0^\top - \alpha^2 \varphi(X^*) \frac{X^*}{\|X^*\|_F} \right) V \right\| \\ &\leq \left\| U_0 U_0^\top - \alpha^2 \varphi(X^*) \frac{X^*}{\|X^*\|_F} \right\| \\ &\leq 2\alpha^2 \varphi(X^*)\delta. \end{aligned}$$

Similarly, we have

$$\begin{aligned} \left\| S_0 E_0^\top \right\| &= \left\| V^\top \left(U_0 U_0^\top - \alpha^2 \varphi(X^*) \frac{X^*}{\|X^*\|_F} \right) V_\perp \right\| \leq 2\alpha^2 \varphi(X^*)\delta, \\ \left\| E_0 E_0^\top \right\| &= \left\| V_\perp^\top \left(U_0 U_0^\top - \alpha^2 \varphi(X^*) \frac{X^*}{\|X^*\|_F} \right) V_\perp \right\| \leq 2\alpha^2 \varphi(X^*)\delta. \end{aligned}$$

This completes the proof. \square

D.6 Proof of Lemma 34

We prove this lemma by induction on t . First, due to Lemma 33, it is easy to verify that (39)-(42) hold for $t = 0$. Now, suppose that (39)-(41) are satisfied for $t < T_{end}$. Moreover, without loss of generality, we assume that $\|\Delta_t\|_F \gtrsim d\alpha^{2-\mathcal{O}(\sqrt{r}\kappa^2\delta)}$ for every $0 \leq t \leq T_{end}$; otherwise, the statement of the lemma holds. Together with the induction hypothesis on $\|G_t\|$, this implies that $\|\Delta_t\|_F \geq \zeta$, for $\zeta > 0$ defined in Propositions 19, 20, and 21.

Bounding $\|F_{t+1}\|$: In order to apply Proposition 21, first we verify its assumptions. One can write $\|E_t E_t^\top\| = \|F_t\|^2 + \|G_t\|^2$. Therefore, we have $\|E_t E_t^\top\| \leq \sigma_1$, due to the induction hypothesis on $\|F_t\|$ and $\|G_t\|$. On the other hand, $\|S_t S_t^\top\| \leq 1.01\sigma_1$ and $\|E_t S_t^\top (S_t S_t^\top)^{-1}\| \leq 1/3$ due to our induction hypothesis. It remains to show that $(4r, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP with $\varepsilon \geq \sqrt{d}\bar{\varphi}\alpha$ implies $(4r, \delta, \sqrt{d}\|G_t\|^2, \mathcal{S})$ -Sign-RIP. To show this, first note that $(4r, \delta, \varepsilon_1, \mathcal{S})$ -Sign-RIP implies $(4r, \delta, \varepsilon_2, \mathcal{S})$ -Sign-RIP, for any $\varepsilon_2 \leq \varepsilon_1$. Using this fact, it suffices to show that $\sqrt{d}\|G_t\|^2 \leq \sqrt{d}\bar{\varphi}\alpha \lesssim \varepsilon$, which is immediate due to our induction hypothesis on $\|G_t\|$,

and our choice of δ . Therefore, Proposition 21 holds and we have

$$\|F_{t+1}\| \leq \left(1 - \bar{\eta}_t \lambda_{\min} \left(S_t S_t^\top\right) + 3\bar{\eta}_t \delta \|\Delta_t\|_F\right) \|F_t\| + 3\bar{\eta}_t \delta \|\Delta_t\|_F \|S_t\| + 6\bar{\eta}_t \|\Delta_t\| \|G_t\|. \quad (97)$$

Due to our induction hypothesis, (97) can be simplified as

$$\begin{aligned} \|F_{t+1}\| &\leq (1 + 75\sqrt{r}\sigma_1\bar{\eta}_t\delta) \|F_t\| + 80\sqrt{r}\sigma_1^{1.5}\bar{\eta}_t\delta + 30\sigma_1\bar{\eta}_t \|G_t\| \\ &\leq \|F_t\| + 100\sqrt{r}\sigma_1^{1.5}\bar{\eta}_t\delta + 30\sigma_1\bar{\eta}_t \|G_t\| \\ &\leq \|F_t\| + 100\sqrt{r}\sigma_1^{1.5}\bar{\eta}_t\delta + 30\sigma_1\bar{\eta}_t\sqrt{\alpha\bar{\varphi}\delta} \\ &\leq \eta\bar{\varphi}^2 \left(100\sqrt{r}\sigma_1^{1.5}\delta + 30\sigma_1\sqrt{\alpha\bar{\varphi}\delta}\right) (t+2), \end{aligned}$$

where in the first inequality, we use Lemma 57 and the induction hypothesis on $\|\Delta_t\|$. Moreover, in the last inequality, we used the induction hypothesis on $\|G_t\|$.

Bounding $\|S_{t+1}S_{t+1}^\top\|$: Similar to our previous argument, it is easy to verify that the assumptions of Proposition 20 are satisfied at iteration t . Therefore, $\|S_{t+1}S_{t+1}^\top\| \leq 1.01\sigma_1$ readily follows from Proposition 20.

Bounding $\|U_{t+1}U_{t+1}^\top - X^*\|$: Note that

$$\|U_{t+1}U_{t+1}^\top - X^*\| \leq \left\|\Sigma - S_{t+1}S_{t+1}^\top\right\| + 2\left\|S_{t+1}E_{t+1}^\top\right\| + \left\|E_{t+1}E_{t+1}^\top\right\|.$$

On the other hand, we have

$$\left\|E_{t+1}E_{t+1}^\top\right\| \leq \|F_{t+1}\|^2 + \|G_{t+1}\|^2 \leq 0.5\sigma_1.$$

where the last inequality is due to our choice of δ and α . Similarly, we can show that $\|S_{t+1}E_{t+1}^\top\| \leq \sigma_1$. This together with $\|\Sigma - S_{t+1}S_{t+1}^\top\| \leq \|S_{t+1}S_{t+1}^\top\| + \sigma_1 \leq 2.01\sigma_1$ leads to

$$\left\|U_{t+1}U_{t+1}^\top - X^*\right\| \leq 5\sigma_1.$$

Establishing $S_{t+1}S_{t+1}^\top \succ 0$: The proof follows directly from the application of Proposition 19. The details are omitted due to their similarity to the proof of (32) in Lemma 32.

Bounding $\left\|E_{t+1}S_{t+1}^\top (S_{t+1}S_{t+1}^\top)^{-1}\right\|$: To streamline the proof, let $H_t = E_t S_t^\top (S_t S_t^\top)^{-1}$. Our goal is to prove $\|H_t\| \leq 1/3$ by showing the following recursive relationship.

Lemma 58 *For every $0 \leq s \leq t$, we have*

$$\|H_{s+1}\| \leq (1 - c\bar{\eta}_t\sigma_r) \|H_s\| + c'\sqrt{r}\sigma_1\bar{\eta}_t\delta, \quad (98)$$

where $c, c' > 0$ are some universal constants.

The proof of Lemma 58 can be found in Appendix F.4. Equipped with this lemma, we are ready to derive the desired result. First, due to Lemma 33, we have

$$\|H_0\| \leq \frac{\|E_0 S_0^\top\|}{\lambda_{\min}(S_0 S_0^\top)} \leq \frac{2\alpha^2 \delta \varphi(X^*)}{\alpha^2 \varphi(X^*) \left(\frac{1}{\sqrt{r\kappa}} - 2\delta\right)} \leq 4\sqrt{r\kappa}\delta,$$

provided that $\delta \leq \frac{1}{4\sqrt{r\kappa}}$. On the other hand, (98) implies that

$$\|H_{t+1}\| - \frac{c'}{c}\sqrt{r\kappa}\delta \leq (1 - c\bar{\eta}_t\sigma_r)^{t+1} \left(\|H_0\| - \frac{c'}{c}\sqrt{r\kappa}\delta \right) \leq \|H_0\| + \frac{c'}{c}\sqrt{r\kappa}\delta.$$

Therefore, due to our choice of δ , we have

$$\|H_{t+1}\| \leq \left(4\sqrt{r\kappa} + \frac{2c'}{c}\sqrt{r\kappa} \right) \delta \lesssim \sqrt{r\kappa}\delta \leq 1/3.$$

Bounding $\|G_{t+1}\|$: Due to Proposition 21, we have

$$\begin{aligned} \|G_{t+1}\| \leq & \left(1 + \bar{\eta}_t^2 \left(2\|E_t S_t^\top\|^2 + \|E_t\|^4 + 6\|E_t S_t (S_t S_t)^{-1}\| \|\Delta_t\| \|E_t S_t^\top\| \right) \right) \\ & + 7\bar{\eta}_t \delta \|\Delta_t\|_F \|G_t\|. \end{aligned}$$

Moreover, one can write

$$\bar{\eta}_t^2 \|E_t S_t^\top\|^2 \lesssim \bar{\eta}_t \|E_t\|^2 \lesssim r\sigma_1^3 \bar{\eta}_t^2 \delta^2 t^2 + \sigma_1^2 \bar{\eta}_t^2 \alpha \bar{\varphi} \delta t^2 \lesssim \sqrt{r\kappa}\delta \log\left(\frac{1}{\alpha}\right).$$

Similarly, it can be shown that

$$\bar{\eta}_t^2 \|E_t\|^4 \vee \bar{\eta}_t^2 \|H_t\| \|\Delta_t\| \|E_t S_t^\top\| \vee \bar{\eta}_t \delta \|\Delta_t\|_F \lesssim \sqrt{r\kappa}\delta \log\left(\frac{1}{\alpha}\right).$$

Therefore, for some universal constant $C > 0$, we have

$$\|G_{t+1}\| \leq \left(1 + C\sqrt{r\kappa}\delta \log\left(\frac{1}{\alpha}\right) \right) \|G_t\|.$$

Hence, we have

$$\begin{aligned} \|G_{t+1}\| & \leq \left(1 + C\sqrt{r\kappa}\delta \log\left(\frac{1}{\alpha}\right) \right)^{t+1} \|G_0\| \\ & \leq \left(1 + C\sqrt{r\kappa}\delta \log\left(\frac{1}{\alpha}\right) \right)^{C' \frac{\log(1/\alpha)}{\bar{\eta}_t \sigma_r}} \|G_0\| \\ & \leq \exp\left(C'' \sqrt{r\kappa}^2 \delta \log\left(\frac{1}{\alpha}\right) \right) \|G_0\| \\ & \leq \alpha^{-\mathcal{O}(\sqrt{r\kappa}^2 \delta)} \|G_0\| \\ & \leq \alpha^{1-\mathcal{O}(\sqrt{r\kappa}^2 \delta)} \sqrt{\bar{\varphi}} \delta, \end{aligned}$$

where in the last inequality, we used the upper bound $\|G_0\| \leq \|E_0\|$ and Lemma 33.

Appendix E. Proofs of Empirical Loss with Noisy Measurements

E.1 Preliminaries

Given the update rule $U_{t+1} = U_t - 2\eta_t Q_t U_t$, we consider the following decomposition

$$U_{t+1} = \tilde{U}_{t+1} + R_t U_t, \text{ where } \tilde{U}_{t+1} = U_t - \frac{2\eta_t \rho^t}{\|\Delta_t\|} \Delta_t U_t, \quad R_t = \frac{2\eta_t \rho^t}{\|\Delta_t\|} \Delta_t - 2\eta_t Q_t, \quad (99)$$

In the above decomposition, \tilde{U}_{t+1} resembles one iteration of GD on $\bar{f}_{\ell_2}(U)$ with the ‘‘effective’’ step-size $\frac{\eta_t \rho^t}{2\|\Delta_t\|}$. Moreover, the term $R_t U_t$ captures the deviation of SubGM and GD. Similar to the noiseless setting, the main idea behind our proof technique is to show that R_t remains small throughout the iterations of SubGM, and consequently, SubGM behaves similarly to GD. To this goal, we first provide an upper bound on $\|\Delta_t\|_F$ in terms of $\|\Delta_t\|$.

Lemma 59 *Suppose that $\sqrt{d} \|G_t\|^2 \leq \|\Delta_t\|$. Then, we have $\|\Delta_t\|_F \leq 2(1 + \sqrt{r}) \|\Delta_t\|$.*

Proof Due to our proposed signal-residual decomposition, one can write

$$\begin{aligned} \|\Delta_t\|_F &\leq \left\| \underbrace{V \left(S_t S_t^\top - \Sigma \right) V^\top + V S_t E_t^\top V_\perp^\top + V_\perp E_t S_t^\top V^\top + V_\perp F_t F_t^\top V_\perp^\top}_{\text{rank-}4r} \right\|_F + \left\| \underbrace{V_\perp G_t G_t^\top V_\perp^\top}_{\text{small norm}} \right\|_F \\ &\leq \sqrt{4r} \left\| V \left(S_t S_t^\top - \Sigma \right) V^\top + V S_t E_t^\top V_\perp^\top + V_\perp E_t S_t^\top V^\top + V_\perp F_t F_t^\top V_\perp^\top \right\| + \sqrt{d} \|G_t\|^2 \\ &\leq \sqrt{4r} \|\Delta_t\| + \sqrt{4r} \left\| V_\perp G_t G_t^\top V_\perp^\top \right\| + \sqrt{d} \|G_t\|^2 \\ &\leq 2(1 + \sqrt{r}) \|\Delta_t\|, \end{aligned}$$

where the last inequality follows from the assumption $4r \leq d$ and $\sqrt{d} \|G_t\|^2 \leq \|\Delta_t\|$. This completes the proof. \blacksquare

Equipped with this technical lemma, we next provide an upper bound on $\|R_t\|$.

Lemma 60 *Suppose that the measurements satisfy $(4r, \delta, \varepsilon, \mathcal{S})$ -Sign-RIP with $\delta < \frac{1}{4(1+\sqrt{r})}$, $\varepsilon = \sqrt{d} \|G_t\|^2$, $\mathcal{S} = \{X : \|X\|_F \geq \zeta\}$ for $\zeta = \sqrt{d} \|G_t\|^2 \left(\frac{1}{\delta} \vee \sqrt{d} \right)$, and $\sqrt{d} \|G_t\|^2 \leq \|\Delta_t\|$. Then, we have $\|R_t\| \leq 8(1 + \sqrt{r}) \eta_t \rho^t \delta$.*

Proof One can write

$$\begin{aligned} R_t &= \frac{2\eta_t \rho^t}{\|\Delta_t\|} \Delta_t - \frac{2\eta_t \rho^t}{\|Q_t\|} Q_t \\ &= -\frac{2\eta_t \rho^t}{\|Q_t\|} \left(Q_t - \varphi(\Delta_t) \frac{\Delta_t}{\|\Delta_t\|_F} \right) - \frac{2\eta_t \rho^t \varphi(\Delta_t)}{\|Q_t\| \|\Delta_t\|_F} \Delta_t + \frac{2\eta_t \rho^t}{\|\Delta_t\|} \Delta_t \\ &= -\frac{2\eta_t \rho^t}{\|Q_t\|} \left(Q_t - \varphi(\Delta_t) \frac{\Delta_t}{\|\Delta_t\|_F} \right) + \frac{\|Q_t\| - \varphi(\Delta_t) \frac{\|\Delta_t\|}{\|\Delta_t\|_F}}{\|Q_t\|} \frac{2\eta_t \rho^t}{\|\Delta_t\|} \Delta_t. \end{aligned}$$

The above equality implies that

$$\begin{aligned} \|R_t\| &\leq \frac{2\eta\rho^t}{\|Q_t\|} \left\| Q_t - \varphi(\Delta_t) \frac{\Delta_t}{\|\Delta_t\|_F} \right\| + 2\eta\rho^t \cdot \frac{\left| \|Q_t\| - \varphi(\Delta_t) \frac{\|\Delta_t\|}{\|\Delta_t\|_F} \right|}{\|Q_t\|} \\ &\leq \left(\frac{1}{\|Q_t\|} \right) 2\eta\rho^t \varphi(\Delta_t) \delta, \end{aligned} \quad (100)$$

where in the last inequality, we used Sign-RIP. Next, we provide an upper bound for $1/\|Q_t\|$. Due to Sign-RIP, we have $\|Q_t\| \geq \left(\frac{\|\Delta_t\|}{\|\Delta_t\|_F} - \delta \right) \varphi(\Delta_t)$. On the other hand, due to Lemma 59, we have $\frac{\|\Delta_t\|}{\|\Delta_t\|_F} \leq \frac{1}{2(1+\sqrt{r})}$. Combining these inequalities with (100), we have

$$\|R_t\| \leq \frac{2}{\frac{1}{2(1+\sqrt{r})} - \delta} \cdot \eta\rho^t \delta \leq 8(1+\sqrt{r})\eta\rho^t \delta,$$

where the last inequality is due to the assumption $\delta \leq \frac{1}{4(1+\sqrt{r})}$. \blacksquare

E.2 Proof of Proposition 24

The proof is almost a line-by-line reconstruction of the proof of Proposition 19 in Appendix D.2. For brevity, we only provide a sketch of the proof. Similar to (82), one can write

$$S_{t+1} = \tilde{S}_{t+1} + V^\top R_t U_t. \quad (101)$$

Given this decomposition, we characterize the growth rate of $\lambda_{\min}(S_{t+1}S_{t+1}^\top)$ by first resorting to a more tractable lower bound. In particular, we define $M := (I + \Xi) \tilde{S}_{t+1} \tilde{S}_{t+1}^\top (I + \Xi)^\top$, where $\Xi := V^\top R_t U_t \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1}$. Based on the definition of M , a series of inequalities analogous to (83) can be used to show that $\lambda_{\min}(S_{t+1}S_{t+1}^\top) \geq \lambda_{\min}(M)$. Therefore, providing a lower bound for $\lambda_{\min}(M)$ suffices. Similar to the proof of Proposition 19, we first show that $\lambda_{\min}(M) \approx \lambda_{\min}(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top)$:

$$\begin{aligned} \left| \lambda_{\min}(M) - \lambda_{\min}(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top) \right| &\leq 3 \|\Xi\| \lambda_{\min}(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top) \\ &\leq 192\sqrt{r}\eta\rho^t \lambda_{\min}(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top). \end{aligned}$$

Combining the above inequality with the one-step dynamic of $\lambda_{\min}(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top)$ from Proposition 16 completes the proof. \square

E.3 Proof of Proposition 25

Similar to the minimum eigenvalue dynamics, the proof is identical to the proof of Proposition 20. Hence, we only provide a sketch. Similar to (85), one can write

$$\Sigma - S_{t+1} S_{t+1}^\top = \Sigma - \tilde{S}_{t+1} \tilde{S}_{t+1}^\top - V^\top R_t U_t \tilde{S}_{t+1}^\top - \tilde{S}_{t+1} U_t^\top R_t^\top V - V^\top R_t U_t U_t^\top R_t^\top V.$$

Lemma 60 combined with an argument similar to Appendix D.3 leads to

$$\left\| V^\top R_t U_t \tilde{S}_{t+1}^\top + \tilde{S}_{t+1} U_t^\top R_t^\top V + V^\top R_t U_t U_t^\top R_t^\top V \right\| \leq 193\sqrt{r}\eta\rho^t\sigma_1\delta.$$

The above bound combined with the one-step dynamics of $\Sigma - \tilde{S}_{t+1}\tilde{S}_{t+1}^\top$ in Proposition 17 completes the proof for the one-step dynamics of $\Sigma - S_{t+1}S_{t+1}^\top$. The dynamics of the cross term (27) and the upper bound on $\|S_{t+1}S_{t+1}^\top\|$ (28) can be deduced in a similar fashion. The details are omitted for brevity. \square

E.4 Proof of Proposition 26

The proof of Proposition 26 is identical to that of Proposition 21, with a key difference that $\|R_t\| \leq 16\sqrt{r}\eta\rho^t$. The details are omitted for brevity. \square

E.5 Proof of Lemma 36

The proof is based on an inductive argument similar to the proof of Lemma 34 in Appendix D.6. Due to our special initialization, it is easy to verify that the statements of the lemma are satisfied for $t = 0$. Now suppose that (50)-(55) are satisfied for t . Due to Proposition 26, one can write

$$\begin{aligned} \|G_{t+1}\| &\leq (1 + 5\eta^2\rho^{2t} + 49\sqrt{r}\eta\rho^t\delta) \|G_t\|, \\ &\leq \exp(5\eta^2\rho^{2t} + 49\sqrt{r}\eta\rho^t\delta) \|G_t\|, \\ &\leq \|G_0\| \prod_{s=0}^t \exp(5\eta^2\rho^{2s} + 49\sqrt{r}\eta\rho^s\delta) \\ &\leq \|G_0\| \exp\left(\sum_{s=0}^t 5\eta^2\rho^{2s} + 49\sqrt{r}\eta\rho^s\delta\right) \\ &\leq \|G_0\| \exp\left(\sum_{s=0}^{\infty} 5\eta^2\rho^{2s} + 49\sqrt{r}\eta\rho^s\delta\right) \\ &\leq \|G_0\| \exp\left(\frac{5\eta^2}{1-\rho^2} + \frac{49\sqrt{r}\eta\delta}{1-\rho}\right). \end{aligned}$$

Due to $\rho = 1 - \Theta(\eta/(\kappa \log(1/\alpha)))$ and $\eta \lesssim 1/(\kappa \log(1/\alpha))$, we have

$$\frac{5\eta^2}{1-\rho^2} \lesssim \eta\kappa \log(1/\alpha) \leq 1/2, \quad \text{and} \quad \frac{49\sqrt{r}\eta\delta}{1-\rho} \lesssim \sqrt{r}\kappa\delta \log(1/\alpha).$$

Combining the above inequalities leads

$$\|G_{t+1}\| \leq 2\|G_0\| \alpha^{-\mathcal{O}(\sqrt{r}\kappa\delta)} \leq 2\sqrt{2}\alpha^{1-\mathcal{O}(\sqrt{r}\kappa\delta)} \sqrt{\bar{\varphi}\delta},$$

where the last inequality is due to our special initialization technique and Lemma 33. The remaining bounds in Lemma 36 can be established similarly to Lemma 34. The details are omitted for brevity. \square

Appendix F. Omitted Proofs

F.1 Proof of Lemma 50

To prove this lemma, we use the one-step discretization technique. First note that, for any $X \in \mathcal{S}_{k,\varepsilon}$, there exists a matrix in $X' \in \mathcal{S}_k$ such that $\|X - X'\|_F \leq \varepsilon$. Suppose that $\mathcal{N}_{k,\xi}$ is a ξ -net of \mathcal{S}_k where $\xi \geq \varepsilon$. Based on the definition of ξ -net, there exists $X'' \in \mathcal{N}_{k,\xi}$ such that $\|X' - X''\|_F \leq \xi$. This implies that $\|X - X''\|_F \leq \|X - X'\|_F + \|X' - X''\|_F \leq 2\xi$, and hence, $\mathcal{N}_{k,\xi}$ is a 2ξ -net of $\mathcal{S}_{k,\varepsilon}$. Given this fact, one can write the following chain of inequalities for every $Y \in \mathbb{S}$:

$$\begin{aligned}
 \mathbb{E}[\mathcal{G}_Y] &\leq \mathbb{E} \left[\underbrace{\sup_{X' \in \mathcal{N}_{k,\xi}} \frac{1}{m} \sum_{i=1}^m \text{Sign}(\langle A_i, X' \rangle - s_i) \langle A_i, Y \rangle - \varphi(X') \left\langle \frac{X'}{\|X'\|_F}, Y \right\rangle}_{(A)} \right] \\
 &+ \mathbb{E} \left[\underbrace{\sup_{\|X - X'\|_F \leq 2\xi} \frac{1}{m} \sum_{i=1}^m (\text{Sign}(\langle A_i, X \rangle - s_i) - \text{Sign}(\langle A_i, X' \rangle - s_i)) \langle A_i, Y \rangle}_{(B)} \right] \\
 &+ \underbrace{\sup_{\|X - X'\|_F \leq 2\xi} \left\langle \varphi(X) \frac{X}{\|X\|_F} - \varphi(X') \frac{X'}{\|X'\|_F}, Y \right\rangle}_{(C)}. \tag{102}
 \end{aligned}$$

We control each term in the above inequality separately.

Bounding (A). To control (A), note that $\frac{1}{m} \sum_{i=1}^m \text{Sign}(\langle A_i, X' \rangle - s_i) \langle A_i, Y \rangle - \varphi(X') \left\langle \frac{X'}{\|X'\|_F}, Y \right\rangle$ is $\mathcal{O}(1/m)$ -sub-Gaussian and (A) is the supremum of the sub-Gaussian random variable over a finite set $\mathcal{N}_{k,\xi}$. Hence, the Maximum Inequality implies that

$$(A) \lesssim \sqrt{\frac{dk}{m} \log \left(\frac{R}{\xi} \right)}. \tag{103}$$

Bounding (B). Invoking Hölder's inequality, one can write

$$(B) \leq \mathbb{E} \left[\sup_{\|X - X'\|_F \leq 2\xi} \left(\frac{1}{m} \sum_{i=1}^m |\text{Sign}(\langle A_i, X \rangle - s_i) - \text{Sign}(\langle A_i, X' \rangle - s_i)| \right) \max_{1 \leq i \leq m} |\langle A_i, Y \rangle| \right].$$

Note that if $|\langle A_i, X - X' \rangle| \leq |\langle A_i, X' - \lambda s_i \rangle|$, then $\text{Sign}(\langle A_i, X' \rangle - \lambda s_i) = \text{Sign}(\langle A_i, X \rangle - \lambda s_i)$. Therefore, the above term can be further bounded by

$$\begin{aligned}
 & (B) \\
 & \leq \mathbb{E} \left[\sup_{\|X - X'\|_F \leq 2\xi} \left(\frac{1}{m} \sum_{i=1}^m \mathbb{1}(|\langle A_i, X - X' \rangle| \geq |\langle A_i, X' \rangle - s_i|) \right) \max_{1 \leq i \leq m} |\langle A_i, Y \rangle| \right] \\
 & \leq \mathbb{E} \left[\sup_{\|X - X'\|_F \leq 2\xi} \left(\frac{1}{m} \sum_{i=1}^m \mathbb{1}(|\langle A_i, X - X' \rangle| \geq t) + \mathbb{1}(|\langle A_i, X' \rangle - s_i| \leq t) \right) \max_{1 \leq i \leq m} |\langle A_i, Y \rangle| \right] \\
 & \leq \underbrace{\mathbb{E} \left[\sup_{\|X - X'\|_F \leq 2\xi} \left(\frac{1}{m} \sum_{i=1}^m \mathbb{1}(|\langle A_i, X - X' \rangle| \geq t) \right) \max_{1 \leq i \leq m} |\langle A_i, Y \rangle| \right]}_{(B_1)} \\
 & \quad + \underbrace{\mathbb{E} \left[\sup_{X' \in \mathcal{N}_{k, \xi}} \left(\frac{1}{m} \sum_{i=1}^m \mathbb{1}(|\langle A_i, X' \rangle - s_i| \leq t) \right) \max_{1 \leq i \leq m} |\langle A_i, Y \rangle| \right]}_{(B_2)}.
 \end{aligned} \tag{104}$$

For (B_1) , we have

$$\begin{aligned}
 (B_1) & \leq \mathbb{E} \left[\left(\frac{1}{m} \sum_{i=1}^m \mathbb{1} \left(\|A_i\|_F \geq \frac{t}{2\xi} \right) \right) \max_{1 \leq i \leq m} |\langle A_i, Y \rangle| \right] \\
 & \leq \mathbb{E} \left[\mathbb{1} \left(\|A_i\|_F \geq \frac{t}{2\xi} \right) \right] \mathbb{E} \left[\max_{j \neq i} |\langle A_j, Y \rangle| \right] \\
 & \quad + \mathbb{E} \left[\mathbb{1} \left(\|A_i\|_F \geq \frac{t}{2\xi} \right) |\langle A_i, Y \rangle| \right] \\
 & \leq \mathcal{O} \left(e^{-C \frac{t^2}{\xi^2}} \sqrt{\log(m)} \right) + \mathbb{E} \left[\mathbb{1} \left(\|A_i\|_F \geq \frac{t}{2\xi} \right) |\langle A_i, Y \rangle| \right],
 \end{aligned}$$

where $C > 0$ is a universal constant and $t/\xi \geq \sqrt{d}$. Furthermore, applying Cauchy-Schwarz inequality, we have

$$\mathbb{E} \left[\mathbb{1} \left(\|A_i\|_F \geq \frac{t}{2\xi} \right) |\langle A_i, Y \rangle| \right] \leq \sqrt{\mathbb{P}(2\xi \|A_i\|_F \geq t)} \sqrt{\mathbb{E} [\langle A_i, Y \rangle^2]} \lesssim e^{-C \frac{t^2}{\xi^2}}.$$

Hence, we conclude that $(B_1) \lesssim e^{-C \frac{t^2}{\xi^2}} \sqrt{\log(m)}$.

Now we turn to bound (B_2) . Note that $(\frac{1}{m} \sum_{i=1}^m \mathbb{1}(|\langle A_i, X' \rangle - s_i| \leq t)) \max_{1 \leq i \leq m} |\langle A_i, Y \rangle|$ is $\mathcal{O}(\log(m)/m)$ -sub-Gaussian, and $\mathcal{N}_{k, \xi}$ is a finite set. Hence, the Maximum Inequality yields

$$\begin{aligned}
 (B_2) & \leq \sup_{X' \in \mathcal{N}_{k, \xi}} \mathbb{E} \left[\left(\frac{1}{m} \sum_{i=1}^m \mathbb{1}(|\langle A_i, X' \rangle - s_i| \leq t) \right) \max_{1 \leq i \leq m} |\langle A_i, Y \rangle| \right] \\
 & \quad + \mathcal{O} \left(\sqrt{\frac{dk \log(m) \log(R/\xi)}{m}} \right).
 \end{aligned} \tag{105}$$

For the first part, one can write

$$\begin{aligned}
 & \mathbb{E} \left[\left(\frac{1}{m} \sum_{i=1}^m \mathbb{1} (|\langle A_i, X' \rangle - s_i| \leq t) \right) \max_{1 \leq i \leq m} |\langle A_i, Y \rangle| \right] \\
 & \leq \mathbb{E} [\mathbb{1} (|\langle A_i, X' \rangle - s_i| \leq t)] \mathbb{E} \left[\max_{j \neq i} |\langle A_i, Y \rangle| \right] + \mathbb{E} [\mathbb{1} (|\langle A_i, X' \rangle - s_i| \leq t) |\langle A_i, Y \rangle|] \\
 & \leq \sqrt{\log(m)} \frac{t}{\zeta}.
 \end{aligned}$$

Therefore, we conclude that (B) $\lesssim \sqrt{\log(m)} \frac{t}{\zeta} + e^{-C \frac{t^2}{\xi^2}} \sqrt{\log(m)} + \sqrt{\frac{dk \log(m) \log(R/\xi)}{m}}$. Finally, it remains to bound (C) in (102).

Bounding (C). We have

$$\begin{aligned}
 (C) &= \sup_{\|X - X'\|_F \leq 2\xi} \left\{ (\varphi(X) - \varphi(X')) \left\langle \frac{X}{\|X\|_F}, Y \right\rangle + \varphi(X') \left\langle \frac{X}{\|X\|_F} - \frac{X'}{\|X'\|_F}, Y \right\rangle \right\} \\
 &\leq \sup_{\|X - X'\|_F \leq 2\xi} \{ \varphi(X) - \varphi(X') \} + \sup_{\|X - X'\|_F \leq 2\xi} \left\| \frac{X}{\|X\|_F} - \frac{X'}{\|X'\|_F} \right\|_F.
 \end{aligned}$$

For the first part, we use the mean value theorem to write

$$|\varphi(X') - \varphi(X)| \leq \|\nabla \varphi(Z)\|_F \|X' - X\|_F \leq 2 \|\nabla \varphi(Z)\|_F \xi,$$

where $Z = \lambda X + (1 - \lambda)X'$, $\lambda \in [0, 1]$. Note that $\nabla \varphi(Z) = \sqrt{\frac{2}{\pi}} p \mathbb{E} \left[\frac{s^2 Z}{\|Z\|_F^4} e^{-\frac{s^2}{2\|Z\|_F^2}} \right]$. Hence, we have

$$\begin{aligned}
 \sup_{\|Z\|_F \geq \zeta} \|\nabla \varphi(Z)\|_F &\lesssim \sup_{\|Z\|_F \geq \zeta} \mathbb{E} \left[\frac{s^2}{\|Z\|_F^3} e^{-\frac{s^2}{2\|Z\|_F^2}} \right] \\
 &\leq \frac{1}{\zeta} \sup_{\|Z\|_F \geq \zeta} \mathbb{E} \left[\frac{s^2}{\|Z\|_F^2} e^{-\frac{s^2}{2\|Z\|_F^2}} \right] \\
 &\lesssim \frac{1}{\zeta}.
 \end{aligned}$$

For the second part, we have

$$\begin{aligned}
 & \sup_{\|X - X'\|_F \leq 2\xi} \left\| \frac{X}{\|X\|_F} - \frac{X'}{\|X'\|_F} \right\|_F \\
 & \leq \sup_{\|X - X'\|_F \leq 2\xi} \left\| \frac{X - X'}{\|X\|_F} \right\|_F + \left\| \frac{X (\|X'\|_F - \|X\|_F)}{\|X\|_F \|X'\|_F} \right\|_F \\
 & \leq \frac{4\xi}{\zeta}.
 \end{aligned}$$

Therefore, we conclude that (C) $\lesssim \frac{\xi}{\zeta}$.

Combining the derived upper bounds for (A), (B), and (C), we have

$$\begin{aligned} \mathbb{E}[\mathcal{G}_Y] &\leq (A) + (B) + (C) \\ &\lesssim \sqrt{\frac{dk}{m} \log\left(\frac{R}{\xi}\right)} + \sqrt{\log(m)} \frac{t}{\zeta} + e^{-C \frac{t^2}{\xi^2}} \sqrt{\log(m)} \\ &\quad + \sqrt{\frac{dk \log(m) \log(R/\xi)}{m}} + \frac{\xi}{\zeta}, \end{aligned} \quad (106)$$

provided that $\xi \geq \varepsilon$ and $t/\xi \geq \sqrt{d}$. Let $\xi \asymp \zeta \sqrt{k/m}$ and $t \asymp \zeta \sqrt{dk \log(m)/m}$. Clearly, these choices of parameters satisfy $t/\xi \geq \sqrt{d}$. Moreover, $\xi \asymp \zeta \sqrt{k/m}$ together with the assumption $\varepsilon \lesssim \zeta \sqrt{k/m}$ implies that $\varepsilon \leq \xi$. Finally, plugging these values in (106) leads to

$$\mathbb{E}[\mathcal{G}_Y] \lesssim \sqrt{\frac{dk}{m} \log^2(m) \log\left(\frac{R}{\zeta}\right)},$$

for every $Y \in \mathbb{S}$. This in turn implies

$$\sup_{Y \in \mathbb{S}_k} \mathbb{E}[\mathcal{G}_Y] \leq \sup_{Y \in \mathbb{S}} \mathbb{E}[\mathcal{G}_Y] \lesssim \sqrt{\frac{dk}{m} \log^2(m) \log\left(\frac{R}{\zeta}\right)},$$

which completes the proof. \square

F.2 Proof of Lemma 54

To prove $0.5 \leq \left\| S_t S_t^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\| \leq 2$, it suffices to show that $0.5 \leq \lambda_{\min} \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \left(S_t S_t^\top \right)^{-1} \right) \leq \left\| \tilde{S}_{t+1} \tilde{S}_{t+1}^\top \left(S_t S_t^\top \right)^{-1} \right\| \leq 2$. For brevity, we only show $0.5 \leq \lambda_{\min} \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \left(S_t S_t^\top \right)^{-1} \right)$, as the other part of the inequality can be proven in a similar fashion. One can write

$$\begin{aligned} \tilde{S}_{t+1} \tilde{S}_{t+1}^\top &= S_t S_t^\top - \bar{\eta}_t S_t S_t^\top \left(S_t S_t^\top - \Sigma \right) - \bar{\eta}_t \left(S_t S_t^\top - \Sigma \right) S_t S_t^\top - 2\bar{\eta}_t S_t E_t E_t^\top S_t^\top \\ &\quad + \bar{\eta}_t^2 \left(S_t S_t^\top - \Sigma \right) S_t S_t^\top \left(S_t S_t^\top - \Sigma \right) + \bar{\eta}_t^2 S_t E_t E_t^\top E_t E_t^\top E_t S_t^\top \\ &\quad + \bar{\eta}_t^2 \left(S_t S_t^\top - \Sigma \right) S_t E_t E_t^\top S_t^\top + \bar{\eta}_t^2 S_t E_t E_t^\top S_t^\top \left(S_t S_t^\top - \Sigma \right). \end{aligned} \quad (107)$$

Note that the eigenvalues of $\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \left(S_t S_t^\top \right)^{-1}$ are real and nonnegative, due to its similarity to $\left(S_t S_t^\top \right)^{-1/2} \tilde{S}_{t+1} \tilde{S}_{t+1}^\top \left(S_t S_t^\top \right)^{-1/2}$. On the other hand, one can write

$$\begin{aligned} &\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \left(S_t S_t^\top \right)^{-1} \\ &= I + \bar{\eta}_t \Sigma + \bar{\eta}_t S_t S_t^\top \Sigma \left(S_t S_t^\top \right)^{-1} - 2\bar{\eta}_t S_t S_t^\top - 2\bar{\eta}_t S_t E_t E_t^\top S_t^\top \left(S_t S_t^\top \right)^{-1} \\ &\quad + \bar{\eta}_t^2 \left(S_t S_t^\top - \Sigma \right) S_t S_t^\top \left(S_t S_t^\top - \Sigma \right) \left(S_t S_t^\top \right)^{-1} + \bar{\eta}_t^2 S_t E_t E_t^\top E_t E_t^\top E_t S_t^\top \left(S_t S_t^\top \right)^{-1} \\ &\quad + \bar{\eta}_t^2 \left(S_t S_t^\top - \Sigma \right) S_t E_t E_t^\top S_t^\top \left(S_t S_t^\top \right)^{-1} + \bar{\eta}_t^2 S_t E_t E_t^\top S_t^\top \left(S_t S_t^\top - \Sigma \right) \left(S_t S_t^\top \right)^{-1}. \end{aligned}$$

We will show that every term in the above decomposition, except for the first term, is in the order of $\mathcal{O}(\bar{\eta}_t \sigma_1)$. First, note that

$$\|\bar{\eta}_t \Sigma\| = \left\| \bar{\eta}_t S_t S_t^\top \Sigma \left(S_t S_t^\top \right)^{-1} \right\| = \mathcal{O}(\bar{\eta}_t \sigma_1).$$

Similarly, we have

$$\begin{aligned} & \left\| 2\bar{\eta}_t S_t S_t^\top + 2\bar{\eta}_t S_t E_t E_t^\top S_t^\top \left(S_t S_t^\top \right)^{-1} \right\| \\ & \leq 2\bar{\eta}_t \left\| S_t S_t^\top \right\| + 2\bar{\eta}_t \left\| S_t E_t E_t^\top S_t^\top \left(S_t S_t^\top \right)^{-1} \right\| \\ & = 2\bar{\eta}_t \left\| S_t S_t^\top \right\| + 2\bar{\eta}_t \left\| E_t E_t^\top \right\| \left\| S_t^\top \left(S_t S_t^\top \right)^{-1} S_t \right\| \\ & \leq 2\bar{\eta}_t \left\| S_t S_t^\top \right\| + 2\bar{\eta}_t \left\| E_t E_t^\top \right\| \\ & = \mathcal{O}(\bar{\eta}_t \sigma_1). \end{aligned}$$

It is easy to see that all the remaining terms are in the order of $\mathcal{O}(\bar{\eta}_t \sigma_1)$; we omit their proofs for brevity. Combining the above bounds, we obtain

$$\lambda_{\min} \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \left(S_t S_t^\top \right)^{-1} \right) = 1 - \mathcal{O}(\bar{\eta}_t \sigma_1) \geq 1/2,$$

where the last inequality is due to our choice of $\bar{\eta}_t$. This in turn implies that $\left\| S_t S_t^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\| \leq 2$. Similarly, we can show that $\left\| S_t S_t^\top \left(S_{t+1} S_{t+1}^\top \right)^{-1} \right\| \leq 3$, by proving $\lambda_{\min} \left(S_{t+1} S_{t+1}^\top \left(S_t S_t^\top \right)^{-1} \right) \geq 1/3$. This can be shown in an analogous fashion, after noting that

$$S_{t+1} S_{t+1}^\top = \tilde{S}_{t+1} \tilde{S}_{t+1}^\top + \underbrace{V^\top R_t U_t \tilde{S}_{t+1}^\top + \tilde{S}_{t+1} U_t^\top R_t^\top V + V^\top R_t U_t U_t^\top R_t^\top V}_{\text{perturbation}}.$$

Similar to the proof of Proposition 20, it can be shown that the norm of the perturbation term is upper bounded by $4\bar{\eta}_t \delta \|\Delta_t\|_F \leq 1/6$, due to our choice of δ and η , and our assumptions on $\|E_t E_t^\top\|$ and $\|S_t S_t^\top\|$.

Finally, we prove the second statement by providing an upper bound for $\left\| \tilde{E}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\|$. Due to the first part of the lemma, one can write

$$\begin{aligned} \left\| \tilde{E}_{t+1} \tilde{S}_{t+1}^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\| & \leq \left\| \tilde{E}_{t+1} \tilde{S}_{t+1}^\top \left(S_t S_t^\top \right)^{-1} \right\| \left\| S_t S_t^\top \left(\tilde{S}_{t+1} \tilde{S}_{t+1}^\top \right)^{-1} \right\| \\ & \leq 2 \left\| \tilde{E}_{t+1} \tilde{S}_{t+1}^\top \left(S_t S_t^\top \right)^{-1} \right\|. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} \tilde{E}_{t+1} \tilde{S}_{t+1}^\top & = E_t S_t^\top + \bar{\eta}_t E_t S_t^\top (\Sigma - S_t S_t^\top) - \bar{\eta}_t E_t (S_t^\top S_t + E_t^\top E_t) S_t^\top \\ & \quad + \bar{\eta}_t^2 E_t (S_t^\top S_t + E_t^\top E_t) S_t^\top \left(S_t S_t^\top - \Sigma \right) \\ & \quad - \bar{\eta}_t E_t E_t^\top E_t S_t^\top + \bar{\eta}_t^2 E_t \left(S_t^\top S_t + E_t^\top E_t \right) E_t^\top E_t S_t^\top. \end{aligned}$$

Let us define $H_t = E_t S_t^\top (S_t S_t^\top)^{-1}$. Based on this definition, one can verify that

$$\begin{aligned} & \tilde{E}_{t+1} \tilde{S}_{t+1}^\top (S_t S_t^\top)^{-1} \\ &= H_t + \bar{\eta}_t H_t S_t S_t^\top (\Sigma - S_t S_t^\top) (S_t S_t^\top)^{-1} - \bar{\eta}_t H_t S_t S_t^\top - 2\bar{\eta}_t E_t E_t^\top H_t \\ & \quad + \bar{\eta}_t^2 H_t (S_t S_t^\top)^2 (S_t S_t^\top - \Sigma) (S_t S_t^\top)^{-1} + \bar{\eta}_t^2 E_t (S_t^\top S_t + E_t^\top E_t) E_t^\top H_t \\ & \quad + \bar{\eta}_t^2 E_t E_t^\top H_t S_t S_t^\top (S_t S_t^\top - \Sigma) (S_t S_t^\top)^{-1}. \end{aligned}$$

Now, note that

$$\begin{aligned} \left\| \bar{\eta}_t H_t S_t S_t^\top (\Sigma - S_t S_t^\top) (S_t S_t^\top)^{-1} \right\| &\leq \bar{\eta}_t \|H_t\| \|\Sigma - S_t S_t^\top\| \\ &\lesssim \bar{\eta}_t \sigma_1 \|H_t\| \leq \frac{1}{12} \|H_t\|. \end{aligned}$$

Similarly, we have $\|\bar{\eta}_t H_t S_t S_t^\top\| \leq \frac{1}{12} \|H_t\|$ and $\|\bar{\eta}_t E_t E_t^\top H_t\| \leq \frac{1}{12} \|H_t\|$. Moreover, we have

$$\begin{aligned} \left\| \bar{\eta}_t^2 H_t (S_t S_t^\top)^2 (S_t S_t^\top - \Sigma) (S_t S_t^\top)^{-1} \right\| &\leq \bar{\eta}_t^2 \|H_t\| \|S_t S_t^\top\| \|S_t S_t^\top - \Sigma\| \\ &\lesssim (\bar{\eta}_t \sigma_1)^2 \|H_t\| \leq \frac{1}{12} \|H_t\|. \end{aligned}$$

In a similar fashion, it can be shown that $\|\bar{\eta}_t^2 E_t (S_t^\top S_t + E_t^\top E_t) E_t^\top H_t\| \leq \frac{1}{12} \|H_t\|$ and $\left\| \bar{\eta}_t^2 E_t E_t^\top H_t S_t S_t^\top (S_t S_t^\top - \Sigma) (S_t S_t^\top)^{-1} \right\| \leq \frac{1}{12} \|H_t\|$. Combining the derived bounds completes the proof. \square

F.3 Proof of Lemma 55

To prove this lemma, we show that

$$S_{t+1} N_t = (I + \Xi) S_t N_t \tag{108}$$

for some matrix Ξ with $\|\Xi\| < 1$. Before proceeding, we show that the above inequality is enough to prove the invertibility of $S_{t+1} N_t$. First, note that $S_t N_t = M_t D_t$. Therefore, the matrix $M_t D_t$ is invertible due to the assumption $S_t S_t^\top \succ 0$. On the other hand, $\|\Xi\| < 1$ implies that $I + \Xi$ is invertible, thereby completing the proof. To verify (108), it suffices to show that $\Xi = S_{t+1} N_t (S_t N_t)^{-1} - I$ has norm less than one. To this goal, we write

$$\begin{aligned} \Xi &= S_{t+1} N_t (S_t N_t)^{-1} - I \\ &= V^\top \left(U_t - \bar{\eta}_t (U_t U_t^\top - X^\star) U_t + R_t U_t \right) N_t \left(V^\top U_t N_t \right)^{-1} - \left(V^\top U_t N_t \right) \left(V^\top U_t N_t \right)^{-1} \\ &= V^\top \left(-\bar{\eta}_t (U_t U_t^\top - X^\star) + R_t \right) U_t N_t \left(V^\top U_t N_t \right)^{-1} \\ &\stackrel{(a)}{=} V^\top \left(-\bar{\eta}_t (U_t U_t^\top - X^\star) + R_t \right) U_t S_t^\top (S_t S_t^\top)^{-1} \\ &= V^\top \left(-\bar{\eta}_t (U_t U_t^\top - X^\star) + R_t \right) V + V^\top \left(-\bar{\eta}_t (U_t U_t^\top - X^\star) + R_t \right) V_\perp H_t, \end{aligned}$$

where H_t is defined as $E_t S_t^\top (S_t S_t^\top)^{-1}$. Moreover, in (a), we used the following chain of equalities

$$N_t \left(V^\top U_t N_t \right)^{-1} = N_t (S_t N_t)^{-1} = N_t D_t^{-1} M_t^\top = N_t D_t M_t^\top M_t D_t^{-2} M_t^\top = S_t^\top (S_t S_t^\top)^{-1}$$

Therefore, we have

$$\begin{aligned} \|\Xi\| &\leq \left\| -\bar{\eta}_t \left(U_t U_t^\top - X^\star \right) + R_t \right\| (1 + \|H_t\|) \\ &\leq 2\bar{\eta}_t \|\Delta_t\| + 6\bar{\eta}_t \delta \|\Delta_t\|_F \leq 3\bar{\eta}_t \|\Delta_t\| \leq 1/2. \end{aligned} \quad (109)$$

Here we used Lemma 57, $\|\Delta_t\| \lesssim \sigma_1$, $\|R_t\| \leq 3\bar{\eta}_t \delta \|\Delta_t\|_F$, $\|H_t\| \leq 1/3$, and our assumption on η . This completes the proof. \square

F.4 Proof of Lemma 58

First, note that

$$\begin{aligned} E_{s+1} S_{s+1}^\top &= E_s S_s^\top + \bar{\eta}_t E_s S_s^\top (\Sigma - S_s S_s^\top) - \bar{\eta}_t E_s (S_s^\top S_s + E_s^\top E_s) S_s^\top \\ &\quad - \bar{\eta}_t E_s E_s^\top E_s S_s^\top + \bar{\eta}_t^2 E_s (S_s^\top S_s + E_s^\top E_s) S_s^\top (S_s S_s^\top - \Sigma) \\ &\quad + \bar{\eta}_t^2 E_s (S_s^\top S_s + E_s^\top E_s) E_s^\top E_s S_s^\top + V_\perp^\top R_s U_s S_{s+1}^\top \\ &\quad + E_{s+1} U_s^\top R_s^\top V + V_\perp^\top R_s U_s U_s^\top R_s^\top V. \end{aligned} \quad (110)$$

On the other hand, one can write

$$\begin{aligned} S_{s+1} S_{s+1}^\top &= S_s S_s^\top + \bar{\eta}_t S_s S_s^\top (\Sigma - S_s S_s^\top) + \bar{\eta}_t (\Sigma - S_s S_s^\top) S_s S_s^\top - 2\bar{\eta}_t S_s E_s^\top E_s S_s^\top \\ &\quad + \bar{\eta}_t^2 (\Sigma - S_s S_s^\top) S_s S_s^\top (\Sigma - S_s S_s^\top) + \bar{\eta}_t^2 S_s E_s^\top E_s E_s^\top E_s S_s^\top \\ &\quad - \bar{\eta}_t^2 (\Sigma - S_s S_s^\top) S_s E_s^\top E_s S_s^\top - \bar{\eta}_t^2 S_s E_s^\top E_s S_s^\top (\Sigma - S_s S_s^\top) \\ &\quad + V^\top R_s U_s S_{s+1}^\top + S_{s+1} U_s^\top R_s^\top V + V^\top R_s U_s U_s^\top R_s^\top V. \end{aligned} \quad (111)$$

Pre-multiplying (111) with H_s leads to a relationship between $S_{s+1} S_{s+1}^\top$ and $E_{s+1} S_{s+1}^\top$:

$$\begin{aligned} E_{s+1} S_{s+1}^\top &= H_s S_{s+1} S_{s+1}^\top + T \\ \implies H_{s+1} &= H_s + T (S_{s+1} S_{s+1})^{-1}, \end{aligned}$$

where simple algebra reveals that

$$\begin{aligned} T &= -H_s \left(\bar{\eta}_t \Sigma S_s S_s^\top - 2\bar{\eta}_t S_s E_s^\top E_s S_s^\top + \bar{\eta}_t^2 \Sigma S_s S_s^\top (\Sigma - S_s S_s^\top) \right) \\ &\quad + \bar{\eta}_t^2 S_s E_s^\top E_s E_s^\top E_s S_s^\top - \bar{\eta}_t^2 \Sigma S_s E_s^\top E_s S_s^\top - \bar{\eta}_t^2 S_s E_s^\top E_s S_s^\top (\Sigma - S_s S_s^\top) \\ &\quad - 2\bar{\eta}_t E_s E_s^\top E_s S_s^\top + \bar{\eta}_t^2 E_s E_s^\top E_s S_s^\top (S_s S_s^\top - \Sigma) + \bar{\eta}_t^2 E_s E_s^\top E_s E_s^\top E_s S_s^\top \\ &\quad + V_\perp^\top R_s U_s S_{s+1}^\top + E_{s+1} U_s^\top R_s^\top V + V_\perp^\top R_s U_s U_s^\top R_s^\top V \\ &\quad - H_s \left(V^\top R_s U_s S_{s+1}^\top + S_{s+1} U_s^\top R_s^\top V + V^\top R_s U_s U_s^\top R_s^\top V \right). \end{aligned}$$

For simplicity, we define $D = S_s S_s^\top (S_{s+1} S_{s+1}^\top)^{-1}$ in the sequel. Based on the definition of T , one can write

$$H_s + T(S_{s+1} S_{s+1})^{-1} = H_s A_s + B_s + C_s, \quad (112)$$

where the matrices A_s , B_s , and C_s are defined as

$$\begin{aligned} A_s &= I - \bar{\eta}_t \Sigma D - \bar{\eta}_t^2 \Sigma \left(S_s S_s^\top \right) \Sigma \left(S_s S_s^\top \right)^{-1} D + \bar{\eta}_t^2 \Sigma S_s S_s^\top D, \\ B_s &= -2\bar{\eta}_t S_s E_s^\top H_s D - \bar{\eta}_t^2 S_s E_s^\top E_s E_s^\top H_s D \\ &\quad - \bar{\eta}_t^2 \Sigma S_s E_s^\top H_s D - \bar{\eta}_t^2 S_s E_s^\top E_s S_s^\top \Sigma \left(S_s S_s^\top \right)^{-1} D - \bar{\eta}_t^2 S_s E_s^\top E_s S_s^\top D \\ &\quad - \left(V^\top R_s U_s S_{s+1}^\top + S_{s+1} U_s^\top R_s^\top V + V^\top R_s U_s U_s^\top R_s^\top V \right) \left(S_{s+1} S_{s+1}^\top \right)^{-1}, \\ C_s &= -2\bar{\eta}_t E_s E_s^\top H_s D + \bar{\eta}_t^2 E_s E_s^\top H_s S_s S_s^\top \left(S_s S_s^\top - \Sigma \right) \left(S_{s+1} S_{s+1}^\top \right)^{-1} \\ &\quad + \left(V_\perp^\top R_s U_s S_{s+1}^\top + E_{s+1} U_s^\top R_s^\top V + V_\perp^\top R_s U_s U_s^\top R_s^\top V \right) \left(S_{s+1} S_{s+1}^\top \right)^{-1} \\ &\quad + \bar{\eta}_t^2 E_s E_s^\top E_s E_s^\top H_s D. \end{aligned}$$

To provide an upper bound on $\|A_s\|$, we define $P = \left(I + \bar{\eta}_t S_s S_s^\top \Sigma \left(S_s S_s^\top \right)^{-1} - \bar{\eta}_t S_s S_s^\top \right) D$ and $Q = I + \bar{\eta}_t S_s S_s^\top \Sigma \left(S_s S_s^\top \right)^{-1} - \bar{\eta}_t S_s S_s^\top$. We have

$$\begin{aligned} \|A_s\|^2 &= \|I - \bar{\eta}_t \Sigma P\|^2 = \lambda_{\max} \left((I - \bar{\eta}_t P \Sigma)^\top (I - \bar{\eta}_t \Sigma P) \right) \\ &\leq 1 + \lambda_{\max} \left(-\bar{\eta}_t \Sigma P - \bar{\eta}_t P^\top \Sigma + \bar{\eta}_t^2 P^\top \Sigma^2 P \right) \\ &= 1 - \bar{\eta}_t \lambda_{\min} \left(\Sigma P + P^\top \Sigma - \bar{\eta}_t P^\top \Sigma^2 P \right) \\ &\leq 1 - \bar{\eta}_t \lambda_{\min} \left(\Sigma P + P^\top \Sigma - P^\top \Sigma P \right), \end{aligned}$$

where we used the fact that $\eta \lesssim \frac{1}{\sigma_1}$ in the last inequality. Now it suffices to provide a lower bound for $\lambda_{\min} (\Sigma P + P^\top \Sigma - P^\top \Sigma P)$. To this goal, we use the following intermediate lemma.

Lemma 61 (Theorem 4.1. in Eisenstat and Ipsen (1998)) *Given a diagonal matrix $\Lambda \in \mathbb{R}^{d \times d}$ and its perturbed variant $\Lambda' = \Lambda R$ for some $R \in \mathbb{R}^{d \times d}$, we have*

$$\min_k |\lambda_i(\Lambda') - \lambda_k(\Lambda)| \leq |\lambda_i(\Lambda')| \|I - R^{-1}\|.$$

for every $i = 1, 2, \dots, d$.

To apply this lemma, we choose $\Lambda = \Sigma$ and $R = P + \Sigma^{-1} P^\top \Sigma - \Sigma^{-1} P^\top \Sigma P$, which leads to the equality $\Sigma P + P^\top \Sigma - P^\top \Sigma P = \Sigma R$. Given this definition and Lemma 61, we have

$$\frac{\sigma_r}{1 + \|I - R^{-1}\|} \leq \lambda_{\min}(\Sigma R). \quad (113)$$

Now, we provide an upper bound for $\|I - R^{-1}\|$. First note that $\|I - R^{-1}\| \leq \|I - R\| \|R^{-1}\|$. On the other hand

$$\|I - R\| = \left\| \Sigma^{-1} \left(P^\top - I \right) \Sigma (I - P) \right\| \leq \|I - P\|^2 \leq (\|I - D\| + \|(I - Q)D\|)^2.$$

Using a similar approach to the proof of Lemma 54, we have

$$\|I - D\| = \left\| \left(S_{t+1} S_{t+1}^\top - S_t S_t^\top \right) \left(S_{t+1} S_{t+1}^\top \right)^{-1} \right\| \leq 0.1.$$

On the other hand, one can write

$$\begin{aligned} \|(I - Q)D\| &\leq \|I - Q\| \|D\| \\ &\stackrel{(a)}{\leq} 3 \|I - Q\| \\ &= \bar{\eta}_t \left\| S_t S_t^\top \left(\Sigma - S_t S_t^\top \right) \left(S_t S_t^\top \right)^{-1} \right\| \\ &\leq \bar{\eta}_t \left\| \Sigma - S_t S_t^\top \right\| \\ &\leq 0.1, \end{aligned}$$

where, in (a), we used Lemma 54. Therefore, we have $\|I - R\| \leq 0.04$. Next, we provide an upper bound for $\|R^{-1}\|$. Note that

$$R = P + \Sigma^{-1} P^\top \Sigma (I - P).$$

By Weyl's inequality, we have

$$\begin{aligned} \sigma_{\min}(R) &\geq \sigma_{\min}(P) - \left\| \Sigma^{-1} P^\top \Sigma (I - P) \right\| \\ &\geq \sigma_{\min}(P) - \|P\| \|I - P\| \\ &\geq 0.8 - 1.2 \times 0.2 = 0.56. \end{aligned}$$

Here we used the fact that $\|I - P\| \leq 0.2$. The above inequality implies that $\|R^{-1}\| = 1/\sigma_{\min}(R) \leq 2$. Combining the above bounds, we have

$$\|I - R^{-1}\| \leq \|I - R\| \|R^{-1}\| \leq 0.08.$$

This together with (113) implies that

$$\lambda_{\min}(\Sigma R) \geq 0.92\sigma_r.$$

Therefore, we have

$$\|A_s\|^2 \leq 1 - 0.92\bar{\eta}_t\sigma_r \quad \implies \quad \|A_s\| \leq 1 - 0.46\bar{\eta}_t\sigma_r. \quad (114)$$

Next, we provide an upper bound for $\|B_s\|$. Simple algebra reveals that

$$\begin{aligned} &\left\| 2\bar{\eta}_t S_s E_s^\top H_s D + \bar{\eta}_t^2 S_s E_s^\top E_s E_s^\top H_s D + \bar{\eta}_t^2 \Sigma S_s E_s^\top H_s D \right. \\ &\quad \left. + \bar{\eta}_t^2 S_s E_s^\top E_s S_s^\top \Sigma \left(S_s S_s^\top \right)^{-1} D + \bar{\eta}_t^2 S_s E_s^\top E_s S_s^\top D \right\| \\ &\lesssim \bar{\eta}_t \left\| S_s E_s^\top \right\| \|H_s\|. \end{aligned}$$

Next, we provide a bound for the remaining terms in B_s . We have

$$\begin{aligned}
 & \left\| V_{\perp}^{\top} R_s U_s S_{s+1}^{\top} \left(S_{s+1} S_{s+1}^{\top} \right)^{-1} \right\| \\
 & \leq \|R_s\| \left\| \left(I - \bar{\eta}_t \left(U_s U_s^{\top} - X^{\star} \right) + R_s \right)^{-1} U_{s+1} S_{s+1}^{\top} \left(S_{s+1} S_{s+1}^{\top} \right)^{-1} \right\| \\
 & \leq \|R_s\| \left\| \left(I - \bar{\eta}_t \left(U_s U_s^{\top} - X^{\star} \right) + R_s \right)^{-1} \right\| \left\| U_{s+1} S_{s+1}^{\top} \left(S_{s+1} S_{s+1}^{\top} \right)^{-1} \right\| \\
 & \lesssim \bar{\eta}_t \delta \|\Delta_s\|_F \left\| U_{s+1} S_{s+1}^{\top} \left(S_{s+1} S_{s+1}^{\top} \right)^{-1} \right\|.
 \end{aligned}$$

To proceed, we provide an upper bound for $\left\| U_{s+1} S_{s+1}^{\top} \left(S_{s+1} S_{s+1}^{\top} \right)^{-1} \right\|$.

$$\begin{aligned}
 & \left\| U_{s+1} S_{s+1}^{\top} \left(S_{s+1} S_{s+1}^{\top} \right)^{-1} \right\| \\
 & \leq \left\| V S_{s+1}^{\top} S_{s+1}^{\top} \left(S_{s+1} S_{s+1}^{\top} \right)^{-1} + V_{\perp} E_{s+1} S_{s+1}^{\top} \left(S_{s+1} S_{s+1}^{\top} \right)^{-1} \right\| \\
 & \leq 1 + \|H_{s+1}\|.
 \end{aligned}$$

Similarly, one can show that

$$\left\| V_{\perp}^{\top} R_s U_s U_s^{\top} R_s^{\top} V \left(S_{s+1} S_{s+1}^{\top} \right)^{-1} \right\| \lesssim \bar{\eta}_t^2 \delta^2 \bar{\varphi}^4 \|\Delta_s\|_F^2 (1 + \|H_s\|) \lesssim \bar{\eta}_t \delta \|\Delta_s\|_F.$$

Combining the derived bounds leads to

$$\|B_s\| \lesssim \bar{\eta}_t \left\| S_s E_s^{\top} \right\| \|H_s\| + \bar{\eta}_t \delta \|\Delta_s\|_F (1 + \|H_{s+1}\|). \quad (115)$$

In a similar way, one can show that

$$\|C_s\| \lesssim \bar{\eta}_t \left\| E_s E_s^{\top} \right\| \|H_s\| + \bar{\eta}_t \delta \|\Delta_s\|_F (1 + \|H_{s+1}\|). \quad (116)$$

Substituting (114), (115), and (116) in (112) yields

$$\begin{aligned}
 & (1 - c_1 \bar{\eta}_t \delta \|\Delta_s\|_F) \|H_{s+1}\| \leq \left(1 - 0.46 \bar{\eta}_t \sigma_r + c_2 \bar{\eta}_t \left(\left\| S_s E_s^{\top} \right\| + \left\| E_s E_s^{\top} \right\| \right) \right. \\
 & \quad \left. + c_3 \bar{\eta}_t \delta \|\Delta_s\|_F \right) \|H_s\| + c_4 \bar{\eta}_t \delta \|\Delta_s\|_F \\
 \implies & \|H_{s+1}\| \leq \left(\frac{1 - 0.46 \bar{\eta}_t \sigma_r + c_2 \bar{\eta}_t \left(\left\| S_s E_s^{\top} \right\| + \left\| E_s E_s^{\top} \right\| \right) + c_3 \bar{\eta}_t \delta \|\Delta_s\|_F}{1 - c_1 \bar{\eta}_t \delta \|\Delta_s\|_F} \right) \|H_s\| \\
 & \quad + \frac{c_4 \bar{\eta}_t \delta \|\Delta_s\|_F}{1 - c_1 \bar{\eta}_t \delta \|\Delta_s\|_F} \\
 \implies & \|H_{s+1}\| \leq (1 - c_5 \bar{\eta}_t \sigma_r) \|H_s\| + c_6 \sqrt{r} \sigma_1 \bar{\eta}_t \delta,
 \end{aligned}$$

where the last inequality follows from the assumed upper bound on δ , as well as (37)-(39). This completes the proof. \square

F.5 Proof of Lemma 37

We first prove the upper bound. One can write

$$\log \left(\prod_{t=0}^T (1 + \alpha \rho^t) \right) = \sum_{t=0}^T \log (1 + \alpha \rho^t) \leq \sum_{t=0}^{\infty} \alpha \rho^t = \frac{\alpha}{1 - \rho}.$$

Hence, we have $\prod_{t=0}^T (1 + \alpha \rho^t) \leq \exp \left(\frac{\alpha}{1 - \rho} \right)$. For the lower bound, we have

$$\begin{aligned} \log \left(\prod_{t=0}^{\infty} (1 + \alpha \rho^t) \right) &= \sum_{t=0}^T \log (1 + \alpha \rho^t) \\ &\geq \sum_{t=0}^T \frac{\alpha \rho^t}{1 + \alpha \rho^t} \\ &\geq \sum_{t=0}^T \frac{\alpha \rho^t}{1 + \alpha} \\ &\geq \frac{\alpha}{1 + \alpha} T \rho^T. \end{aligned}$$

Hence, we have $\prod_{t=0}^T (1 + \alpha \rho^t) \geq \exp \left(\frac{\alpha}{1 + \alpha} T \rho^T \right)$. \square

F.6 Proof of Lemma 38

We start with the proof of the first statement. We consider two cases:

- Suppose that $\|\Delta_t\| \leq 0.01\sigma_r\rho^t$. Recall that $U_{t+1} = U_t - \frac{\eta_0\rho^t}{\|\Delta_t\|}\Delta_t U_t + R_t U_t$. Hence, we have

$$\Delta_{t+1} = \left(U_t - \frac{\eta_0\rho^t}{\|\Delta_t\|}\Delta_t U_t + R_t U_t \right) \left(U_t^\top - \frac{\eta_0\rho^t}{\|\Delta_t\|}U_t^\top \Delta_t^\top + U_t^\top R_t^\top \right) - X^*.$$

The above equality leads to

$$\begin{aligned} \|\Delta_{t+1}\| &\leq \|\Delta_t\| + \eta\rho^t \left\| U_t U_t^\top \right\| (2 + 2\|R_t\|) \\ &\quad + 2 \left\| U_t U_t^\top \right\| \|R_t\| + \left\| U_t U_t^\top \right\| \|R_t\|^2 + \eta^2 \rho^{2t} \left\| U_t U_t^\top \right\| \\ &\stackrel{(a)}{\leq} \|\Delta_t\| + 4\sigma_1\eta\rho^t + 4\sigma_1\eta\delta\rho^t \|\Delta_t\|_F + 2\sigma_1\eta^2\rho^{2t} \\ &\leq 0.02\sigma_r\rho^t. \end{aligned}$$

On the other hand, we know that $\gamma_{t+1} \leq 5\|\Delta_{t+1}\|$, which, together with the above inequality, implies that $\gamma_{t+1} \leq 0.1\sigma_r$.

- Suppose that $\|\Delta_t\| \geq 0.01\sigma_r\rho^t$. Therefore, we have $\frac{\eta_0\rho^t}{\|\Delta_t\|} \lesssim \frac{1}{\sigma_1}$. On the other hand, since $\gamma_t \leq 0.1\sigma_r$, we have

$$\lambda_{\min} \left(S_t S_t^\top \right) \geq 0.9\sigma_r, \quad \left\| S_t S_t^\top \right\| \leq 1.1\sigma_1,$$

and

$$\left\| E_t E_t^\top \right\| \leq 0.1 \sigma_r, \quad \left\| E_t S_t^\top \left(S_t S_t^\top \right)^{-1} \right\| \leq 0.2.$$

This implies that the assumptions of Propositions 24, 25, and 26 are satisfied at iteration t , and we have

$$\begin{aligned} \gamma_{t+1} &\leq \left\| \Sigma - S_{t+1} S_{t+1}^\top \right\| + 2 \left\| S_{t+1} E_{t+1}^\top \right\| + \|F_{t+1}\|^2 + \|G_{t+1}\|^2 \\ &\stackrel{(a)}{\leq} \left(1 - \Omega(1) \frac{\sigma_r \eta_0 \rho^t}{\gamma_t} \right) \gamma_t + \mathcal{O}(1) \sqrt{r} \sigma_1 \eta_0 \delta \rho^t + \mathcal{O}(1) \frac{\sigma_r \eta_0 \rho^t}{\|\Delta_t\|} \|G_t\|^2 \\ &\leq \gamma_t - \Omega(\sigma_r \eta_0 \rho^t) + \mathcal{O}(1) \sigma_r \eta_0 \rho^t \frac{\|G_t\|^2}{\|\Delta_t\|} \\ &\stackrel{(b)}{\leq} \gamma_t - \Omega(\sigma_r \eta_0 \rho^t) + \mathcal{O}\left(\sigma_r \eta_0 \rho^t / \sqrt{d}\right) \\ &\leq \gamma_t - \Omega(\sigma_r \eta_0 \rho^t), \end{aligned}$$

where (a) follows from the one-step dynamics of the signal, cross, and residual terms derived in Propositions 24, 25, and 26. Moreover, (b) follows from $\|\Delta_t\| \geq \sqrt{d} \|G_t\|^2$. Therefore, we have $\gamma_{t+1} \leq \gamma_t \leq 0.1 \sigma_r$.

To complete the proof of this lemma, it suffices to show that if $\|\Delta_{t+1}\| \leq \sqrt{d} \|G_{t+1}\|^2$, then $\|\Delta_{t+1}\| \leq \sqrt{d} \alpha^{1-\mathcal{O}(\sqrt{r\kappa\delta})}$. Note that based on our assumption and Phase 1 of the proof of Theorem 27, the one-step dynamic of G_s holds for every $0 \leq s \leq t+1$. Therefore, an analysis similar to Lemma (36) leads to $\|\Delta_{t+1}\| \leq \sqrt{d} \alpha^{1-\mathcal{O}(\sqrt{r\kappa\delta})}$. \square

F.7 Proof of Lemma 39

Since $\|\Delta_{t_0+T_3-1}\| \leq 0.02 \sigma_r \rho^{t_0-1}$, an argument similar to the proof of Lemma 38 can be invoked to show that $\|\Delta_{t_0+T_3}\| \leq 0.03 \sigma_r \rho^{t_0}$ and $\gamma_{t_0+T_3} \leq 0.15 \sigma_r \rho^{t_0}$. Let Δt be the first time that $\|\Delta_{t_0+T_3+\Delta t}\| \leq 0.02 \sigma_r \rho^{t_0+\Delta t}$. Note that since $\|\Delta_{t_0+T_3}\| > 0.02 \sigma_r \rho^{t_0}$, we have $\Delta t \geq 1$. This implies that, for every $0 \leq s \leq \Delta t - 1$, we have $\|\Delta_{t_0+T_3+s}\| > 0.02 \sigma_r \rho^{t_0+s}$. Therefore, (117) implies that $\gamma_{t_0+T_3+s+1} \leq \gamma_{t_0+T_3+s} - \Omega(\sigma_r \eta \rho^{t_0+s})$. This in turn leads to

$$\begin{aligned} \gamma_{t_0+T_3+\Delta t} &\leq \gamma_{t_0+T_3} - \Omega \left(\sum_{s=0}^{\Delta t-1} \sigma_r \eta \rho^{t_0+s} \right) \\ &\leq 0.15 \sigma_r \rho^{t_0} - \Omega \left(\sum_{s=0}^{\Delta t-1} \sigma_r \eta \rho^{t_0+s} \right) \\ &= \sigma_r \rho^{t_0} \left(0.15 - \Omega \left(\sum_{s=0}^{\Delta t-1} \eta \rho^s \right) \right). \end{aligned}$$

Let us assume that $\Delta t \lesssim (\kappa/\eta) \log(1/\alpha)$. Under this assumption, we have $\rho^s = \Omega(1)$ for every $s \leq \Delta t$. This implies that $\Omega \left(\sum_{s=0}^{\Delta t-1} \eta \rho^s \right) = \Omega(\eta \Delta t)$. Therefore, upon choosing $\Delta t = \Omega(1/\eta)$, we have $\gamma_{t_0+T_3+\Delta t} \leq 0.15 - \Omega \left(\sum_{s=0}^{\Delta t-1} \eta \rho^s \right) \leq 0$. This implies that, there must exist $\tilde{t} \leq \Delta t$ such that $\gamma_{t_0+T_3+\tilde{t}} \leq 0.02 \sigma_r \rho^{t_0+\tilde{t}}$. This completes the proof. \square