Implicit Regularization of Sub-Gradient Method in Robust Matrix Recovery: Don't be Afraid of Outliers

Jianhao Ma and Salar Fattahi

Industrial and Operations Engineering, University of Michigan

February 4, 2021

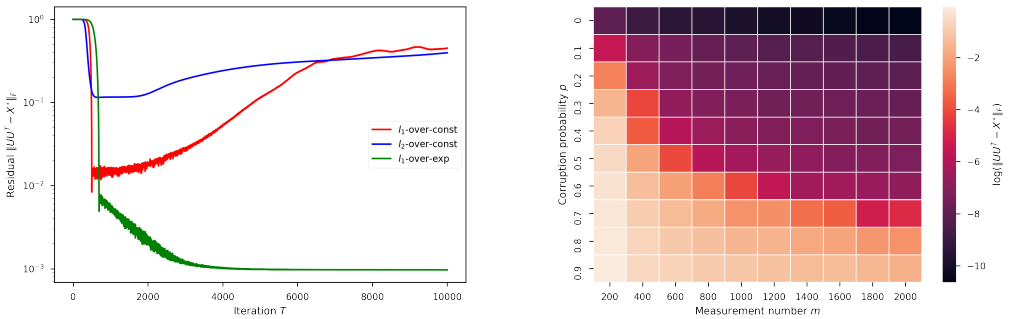
Abstract

It is well-known that simple short-sighted algorithms, such as gradient descent, generalize well in the over-parameterized learning tasks, due to their *implicit regularization*. However, it is unknown whether the implicit regularization of these algorithms can be extended to *robust* learning tasks, where a subset of samples may be grossly corrupted with noise. In this work, we provide a positive answer to this question in the context of robust matrix recovery problem. In particular, we consider the problem of recovering a low-rank matrix from a number of linear measurements, where a subset of measurements are corrupted with large noise. We show that a simple sub-gradient method converges to the true low-rank solution efficiently, when it is applied to the over-parameterized ℓ_1-loss function without any explicit regularization or rank constraint. Moreover, by building upon a new notion of restricted isometry property, called *sign-RIP*, we prove the robustness of the sub-gradient method against outliers in the over-parameterized regime. In particular, we show that, with Gaussian measurements, the sub-gradient method is guaranteed to converge to the true low-rank solution, even if an *arbitrary* fraction of the measurements are grossly corrupted with noise.

1 Introduction

Bolstered by the success of deep learning and their desirable generalization properties, over-parameterized models in modern machine learning have gained special attention in recent years [Neyshabur et al., 2014, Soudry et al., 2018, Oymak and Soltanolkotabi, 2019, Zhang et al., 2016, You et al., 2020]. While classical results in statistical learning suggest that increasing the number of parameters beyond the true dimension of the problem would lead to *overfitting*, a growing body of work shows that, for a large class of learning problems, simple "short-sighted" algorithms, such as gradient descent (GD) or stochastic gradient descent (SGD), lead to surprisingly good generalization bounds, due to their *implicit regularization* property.

The implicit regularization of GD and SGD by now has a very well-established theory: it is well-known that GD converges to a solution with *max-margin* property in logistic regression [Soudry et al., 2018]; it leads to solutions with desirable generalization properties in over-parameterized linear regression [Bartlett et al., 2020]; and, when combined with a proper initialization and step size, it recovers the true low-rank solution in the over-parameterized nonconvex matrix sensing problem [Gunasekar et al., 2018, Li et al., 2018]. However, the current theory behind the success of GD in over-parameterized learning problems hinges heavily upon the smoothness of the loss



(a) (b)

Figure 1: (a) The iterations of SubGD and GD for the over-parameterized ℓ_1- and ℓ_2-loss functions. ℓ_1-`over-const`, ℓ_1-`over-exp`, and ℓ_2-`over-const` respectively correspond to SubGD with constant step size for over-parameterized ℓ_1-loss, SubGD with geometric step size for over-parameterized ℓ_1-loss, and GD with constant step size for over-parameterized ℓ_2-loss. (b) The performance of SubGD with geometric step size for different corruption probability p and number of measurements m. The dimension is set to $d = 20$.

function, rendering them impractical in the *robust* learning tasks, where the loss function is often non-differentiable.

Inspired by this observation, in this work we study the robust matrix recovery problem, where the goal is to recover a rank-r positive semidefnite 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 $\mathbf{y} = [y_1, y_2, \ldots, y_m]^\top$ is the vector of measurements, $\mathcal{A}$ is a linear operator defined as $\mathcal{A}(X^*) = [\langle A_1, X^*\rangle, \langle A_2, X^*\rangle, \ldots, \langle A_m, X^*\rangle]^\top$, and $\mathbf{s}$ is a sparse-and-large noise vector. In particular, we consider the following empirical risk minimization (ERM) problem

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

where r' is chosen as an upper bound for the rank of the true solution, the ℓ_1-loss function is used to promote sparsity in the estimated noise values, and $UU^\top$ is used in lieu of X to ensure the positive semidefiniteness of the solution. Evidently, the above optimization problem is over-parameterized for $r' \gg r$, since the unknown variable is not restricted to the set of low-rank matrices, and consequently, its globally optimal solution need *not* be low-rank. To circumvent this challenge, the existing methods either resort to *convex relaxation* techniques by indirectly penalizing the rank of the solution via trace-norm regularizers [Candès et al., 2011, Chandrasekaran et al., 2011], or directly impose the rank constraint by restricting the feasible region to the set of rank-r matrices [Josz et al., 2018, Fattahi and Sojoudi, 2020, Li et al., 2020b], (also known as Burer-Monteiro approach [Burer and Monteiro, 2003]). However, these methods suffer from several fundamental issues that preclude their practical use. On the one hand, convex relaxation techniques suffer from notoriously high computational cost. On the other hand, methods based on Burer-Monteiro approach impose restrictive structures on the measurements to avoid getting trapped in sub-optimal local solutions, or require prior knowledge on the problem-dependent parameters, such as the rank or approximate value of the true solution.

As a first attempt to address these fundamental challenges, we study the convergence of a simple and unregularized sub-gradient descent (SubGD) algorithm, when it is directly applied to (1) in the over-parameterized regime, where $r' \geq r$ (e.g., $r' = d$). In particular, we aim to shed light on the following questions:

Question 1: *Does SubGD converge to the true low-rank solution $X^* = U^* U^{*\top}$ without any explicit regularization or rank constraint?*
Question 2: *How much noise can be tolerated in the model, without jeopardizing the convergence of SubGD?*

2 Summary of Contributions

In this paper, we address the aforementioned questions for the over-parameterized robust matrix recovery problem, where the rank of the true solution is one. In particular, we show that, under some conditions:

SubGD converges to the true rank-1 solution in the over-parameterized robust matrix recovery problem, when the measurements are subject to arbitrarily large noise values.

Before delving into the details, we shall illustrate the performance of SubGD on a simple toy example[1]. In particular, suppose that the dimension and the number of measurements are chosen as $d = 50$ and $m = 500$, respectively. Moreover, the elements of the measurement matrices $\{A_i\}_{i=1}^m$ are chosen according to a standard Gaussian distribution, and each measurement is corrupted with a large noise value with probability $p = 0.1$. Figure 1a illustrates the performance of SubGD with constant and geometrically decaying step sizes in the over-parameterized robust matrix recovery problem (1) with $r' = d$, together with the performance of GD applied to the over-parameterized ℓ_2-loss function. It can be seen that neither GD nor SubGD with constant step sizes converge to the true low-rank matrix. On the contrary, SubGD with geometrically decaying step sizes converges to the true low-rank solution. This gives rise to the following observation:

Observation 1: *SubGD with geometrically decaying step sizes converges to the true solution of the over-parameterized robust matrix with sparse-and-large noise values.*

To better illustrate the effect of noise, we evaluate the performance of SubGD with respect to the corruption probability p and the number of measurements m in Figure 1b. Intriguingly, it can be seen that SubGD converges to the true solution, even if more than half of the measurements are grossly corrupted with noise. This leads to another observation:

Observation 2: *SubGD is robust against large corruption probabilities, provided that the number of measurements is sufficiently large.*

Inspired by the promising performance of SubGD in the over-parameterized robust matrix recovery problem, we prove the following main result, whose formal version can be found in Section 5.2.

Theorem 1 (Convergence of SubGD, Informal)**.** *Consider the problem (1) with arbitrary $r' \geq 1$. Suppose that the measurements satisfy the sign restricted isometry property (sign-RIP) condition delineated in Section 4 with parameters (r, δ) and an appropriate choice of scaling function φ_t. Moreover, suppose that the initial point is chosen as $U_0 = \alpha B$, where $B \in \mathbb{R}^{d \times r'}$ is an arbitrary*

[1] We provide a more extensive numerical study in the supplementary file.

orthonormal matrix and $\alpha = \mathcal{O}\left(\sqrt{\frac{1}{r'}}\delta\right)$. *Then, upon choosing a geometrically decaying step size* $\eta_t = \frac{\eta_0}{\varphi_t}\rho^t$, *where* $\eta_0 = \mathcal{O}(\delta)$ *and* $\rho = 1 - \Theta\left(\frac{\eta_0}{\log(1/\alpha)}\right)$, *we have*

$$\left\|U_T U_T^\top - X^\star\right\|_F^2 \lesssim \delta^2 \log^2\left(r'/\delta\right). \tag{2}$$

after $T = \Theta\left(\frac{\log(r'/\delta)}{\delta}\right)$ *iterations.*

Our theorem relies on a new notion of sign-RIP, which is in parallel with the classical notions of ℓ_2-RIP [Recht et al., 2010] and ℓ_1/ℓ_2-RIP [Li et al., 2020b]. Roughly speaking, the sign-RIP condition entails that the sub-differentials of the ℓ_1-loss are δ-away from an ideal isotropic linear operator. One of the main contributions of this work is to show that the sign-RIP condition is not more restrictive than its commonly-known counterparts, i.e., ℓ_2- and ℓ_1/ℓ_2-RIP conditions. In particular, we show that, with an overwhelming probability, the sub-differentials of (1) are concentrated around an isotropic linear operator for Gaussian measurement matrices, provided that the number of measurements scales linearly with the number of variables.

A number of observations can be made based on Theorem 1. First, it implies that the corruption probability p can be arbitrarily close to 1, provided that the sign-RIP condition is satisfied. This improves upon a recent result in [Li et al., 2020b], which shows that SubGD converges to the true solution in the exact formulation of the robust matrix recovery problem with $r' = r$, provided that $p < 0.5$. As will be shown later, the sign-RIP condition holds for Gaussian measurements with an arbitrarily large corruption probability p, provided that the number of measurements scales as $(1-p)^{-4}$. Moreover, unlike GD for ℓ_2-loss function, SubGD requires a geometric step size that is inverse proportional to an scaling parameter φ_t, capturing the effect of the corruption probability. In particular, we show that φ_t scales as $\sqrt{\frac{2}{\pi}}(1-p)$ for Gaussian measurements. This implies that SubGD must take "more aggressive" steps with an increasing corruption probability. The main intuition behind this dependency can be traced back to the expected behavior of the sub-differential of the objective: roughly speaking, we show that the norm of this expected value is proportional to $1-p$. Therefore, a large value of p leads to a smaller expected sub-differential, which in turn should be compensated by a larger step size to make meaningful progress. Finally, our result holds under the so-called *early stopping* regime, where the number of iterations of SubGD is both upper and lower bounded by problem-specific parameters. Similar requirement are also imposed in other similar over-parameterized problems [Gunasekar et al., 2018, Li et al., 2018].

It is crucial to characterize the class of measurement matrices that satisfy the sign-RIP condition required in Theorem 1. In our next corollary, we show that this condition holds with high probability for Gaussian measurements, provided that the number of measurements exceeds a threshold.

Proposition 1 (sign-RIP, Informal)**.** *Suppose that the measurement matrices* $\{A_i\}_{i=1}^m$ *have i.i.d. standard Gaussian entries. Then, with an overwhelming probability, the sign-RIP condition 1 holds with parameters* (r, δ) *and an appropriate scaling function, provided that* $m = \Omega\left(\frac{dr}{\delta^4(1-p)^4}\right)$ *(modulo logarithmic factors).*

Proposition 1 combined with Theorem 1 gives rise to the sample complexity of SubGD when applied to the over-parameterized robust matrix recovery with Gaussian measurements.

	Exact	Over-parameterized
ℓ_2-loss	[Ge et al., 2017]	[Li et al., 2018]
ℓ_1-loss	[Li et al., 2020b]	This Work

Table 1: *Exact* and *over-parameterized* refer to $r' = r$ and $r' > r$, respectively.

3 Related Work

Low-Rank matrix recovery Earlier methods for solving low-rank matrix recovery problem were based on convex relaxation techniques, where the problem is *convexified* by lifting to a higher dimension, and the rank constraint is imposed implicitly via a convex regularizer. However, such convexification methods often lead to a computationally expensive conic optimization [Recht et al., 2010, Candès et al., 2011, Chandrasekaran et al., 2011]. To alleviate this issue, a recent line of works focuses on matrix factorization techniques, where the goal is to recover the true low-rank matrix by solving the following optimization problem

$$\min_{U \in \mathbb{R}^{d \times r'}} \left\| \mathbf{y} - \mathcal{A}\left(UU^\top\right) \right\|_{\ell_q}^q . \tag{3}$$

When $r' = r$ and $q = 2$, the optimization problem (3) is devoid of spurious local minima [Ge et al., 2017, 2016, Zhang et al., 2019], and hence, simple local search methods converge to the ground truth. Moreover, in the over-parameterized regime, where $r' = d$, a simple GD method converges to the ground truth of (3) [Li et al., 2018]. However, it is known that ℓ_2-loss is vulnerable to outliers. To address this issue, Li et al. [2020b] demonstrate that under the so-called ℓ_1/ℓ_2-RIP, SubGD method converges to the ground truth with $q = 1$, provided that the true rank of the solution is known and the initial point is sufficiently close to the ground truth. Moreover, Fattahi and Sojoudi [2020] and Josz et al. [2018] prove that (3) with $r' = r = 1$ and $q = 1$ has no spurious local solution, provided that the measurement matrices correspond to element-wise projection operators. In the over-parameterized regime, You et al. [2020] propose to circumvent the nonsmoothness of the robust matrix recovery problem by resorting to a smooth doubly over-parameterized model, and then solving it via GD. However, this method is not directly applicable to the nonsmooth ℓ_1-loss that is considered in this paper. In conclusion, our work fills in the gap in Table 1.

As for the convergence of different local search algorithms, it is known that GD enjoys a polynomial convergence rate when it is applied to (3) with $q = 2$ [Bhojanapalli et al., 2016, Ge et al., 2017]. More recent works achieve (nearly) linear rate via Riemannian gradient descent [Hou et al., 2020, Zhang and Yang, 2018], accelerated gradient descent [Ajayi et al., 2018], and other variants of gradient descent [Yi et al., 2016, Chen et al., 2019, Li et al., 2019].

Uniform convergence of gradients Uniform convergence of gradients is at the core of the convergence analysis of gradient-based algorithms in the empirical risk minimization problems. Recently, Mei et al. [2018] and Foster et al. [2018] provide uniform gradient bounds for nonconvex and smooth ERM problems based on covering arguments. However, these works heavily rely on the smoothness of the loss function. Within the realm of nonsmooth optimization, Bai et al. [2018] study robust dictionary learning problem with ℓ_1-loss function, and derive a uniform bound on its sub-differential with respect to Hausdorff distance. Moreover, Davis and Drusvyatskiy [2018] obtain a dimension-dependent bound for a similar problem via a smoothing technique, analyzing the uniform convergence of the gradient of proximal map in the Hausdorff distance.

Robustness of over-parameterized model It is known that over-parameterized models can overfit to noisy values, but still *generalize* well [Bartlett et al., 2020, Chatterji and Long, 2020]. The main reason behind this seemingly contradictory phenomenon is that the noise can be absorbed in the "unimportant" dimensions of the solution [Bartlett et al., 2020]. However, it is more challenging to identify and reject outliers in over-parameterized models. Li et al. [2020a] show that GD with early stopping is robust against label noise in classification problems, since the model would first fit the clean data and then overfit the outliers. Similar results were also observed by Dong et al. [2019]. However, none of the existing works have addressed the effect of over-parameterization in the robust matrix recovery with ℓ_1-loss function; a topic that is at the core of this paper.

Notations For two matrices X and Y of the same size, their inner product is defined as $\langle X, Y\rangle = \mathrm{Tr}(X^\top Y)$. 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 $\mathbb{S}_r = \{X \in \mathbb{R}^{d\times d} : \|X\|_F = 1, \mathrm{rank}(X) \le r\}$. For simplicity, we denote $\mathbb{S} = \mathbb{S}_d$. The ℓ_q norm of a vector x is defined as $\|x\|_{\ell_q} = (\sum |x_i|^q)^{1/q}$. For simplicity of notation, we write $\|x\|_1 = \|x\|_{\ell_1}$ and $\|x\| = \|x\|_{\ell_2}$. Given two sequences $f(n)$ and $g(n)$, the notation $f(n) \lesssim g(n)$ implies that there exists a constant $C < \infty$ that satisfies $f(n) \le Cg(n)$. Moreover, the notation $f(n) \asymp g(n)$ implies that $f(n) \lesssim g(n)$ and $g(n) \lesssim f(n)$. The sign function $\mathrm{Sign}(\cdot)$ is defined as $\mathrm{Sign}(x) = x/|x|$ if $x \neq 0$, and $\mathrm{Sign}(0) = [-1, 1]$. The letters C and c are reserved to denote universal constants whose values may change throughout the paper.

4 Proposed Algorithm and Preliminaries

Algorithm 1 illustrates our proposed SubGD method: at every iteration, the algorithm selects an arbitrary direction D_t from the sub-differential of the ℓ_1-loss function at the current solution, and then updates the solution by moving towards $-D_t$ with a step size η_t. Evidently, to study the convergence of the iterates $\{U_t\}_{t=0}^T$, it is essential to analyse the behavior of the sub-differential $\partial f_{\ell_1}(U_t)$, which can be written as

$$\partial f_{\ell_1}(U_t) = \frac{1}{m}\sum_{i=1}^m \mathrm{Sign}(\langle A_i, U_tU_t^\top - X^\star\rangle - s_i)A_iU_t.$$

Before analyzing the behavior of $\partial f_{\ell_1}(U_t)$, it is worthwhile to revisit the behavior of GD for the ℓ_2-loss function in the noiseless setting. Given the loss function $f_{\ell_2}(U) = \frac{1}{2m}\|\mathbf{y} - \mathcal{A}(UU^\top)\|^2$, its

Algorithm 1 Sub-Gradient Descent

Input: measurement matrices $\{A_i\}_{i=1}^m$, measurement vector $\mathbf{y} = [y_1, \cdots, y_m]^\top$, number of iterations T, the initialization parameter α, an upper bound on the rank r', and an orthonormal matrix $B \in \mathbb{R}^{d\times r'}$;
Output: Solution $\hat{X}_T = U_TU_T^\top$ to (1);
Initialize $U_0 = \alpha B$.
for $t \le T$ **do**
  Compute a sub-gradient $D_t \in \partial f_{\ell_1}(U_t)$;
  Select the step size η_t;
  Set $U_{t+1} \leftarrow U_t - \eta_t D_t$;
end for

gradient can be written as $\nabla f_{\ell_2}(U) = M(UU^\top - X^*)U$, where $M(X) = \frac{1}{m}\sum_{i=1}^m \langle A_i, X\rangle A_i$. Based on this derivation, Li et al. [2018] show that GD with constant step size converges to a solution that satisfies $U_T U_T^\top \approx X^*$, provided that $\mathcal{A}(X)$ is approximately norm-preserving, which in turn implies that $M(UU^\top - X^*)U \approx (UU^\top - X^*)U$.

Motivated by this observation, our aim is to establish a similar guarantee on the sub-differentials of the ℓ_1-loss function. In particular, upon defining $\mathcal{M}(X) = \frac{1}{m}\sum_{i=1}^m \mathrm{Sign}(\langle A_i, X\rangle - s_i)A_i$, our goal is to guarantee that, for *any* $D \in \mathcal{M}(UU^\top - X^*)$, the sub-gradient DU of the ℓ_1-loss function points *approximately* to the direction $(UU^\top - X^*)U$. This is formalized through the notion of *sign-RIP*, which is defined below.

Definition 1 (sign-RIP)**.** *The measurements are said to satisfy sign-RIP with parameters* (δ, r) *and a scaling function* $\varphi : \mathbb{R}^{d\times d} \to \mathbb{R}$ *if, for every rank-*r *matrix* $X \in \mathbb{R}^{d\times d}$ *and every* $D \in \mathcal{M}(X)$*, we have* $\left\| D - \varphi(X)\frac{X}{\|X\|_F}\right\|_F \le \varphi(X)\delta$.

At the first glance, one may speculate that sign-RIP is extremely restrictive: roughly speaking, it requires the uniform concentration of random set-valued function $\mathcal{M}(X)$, for *every* rank-r matrix X. However, as we will show later, with appropriate choices of parameters and scaling function, the sign-RIP condition is guaranteed to be satisfied with high probability for Gaussian measurement matrices, even if an *arbitrary* fraction of the measurements are corrupted with large noise values.

5 Main Results

In this section, we present the main results of this paper. Due to space restrictions, all proofs are deferred to the supplementary file.

5.1 Noiseless Case

As a base case, we assume that the measurements are noiseless, i.e., $\mathbf{s} = 0$ and $\mathbf{y} = \mathcal{A}(X^*)$. As will be shown later, our analysis in the noiseless setting will be a building block of our subsequent analysis in the noisy scenario. Before delving into the analysis of SubGD, we first show that the sign-RIP condition is satisfied for the Gaussian measurement matrices in the noiseless setting.

Proposition 2. *Assume that* $\mathbf{s} = 0$ *and the measurement matrices* $\{A_i\}_{i=1}^m$ *defining the linear operator* $\mathcal{A}(\cdot)$ *have i.i.d. standard Gaussian entries. Then, the sign-RIP condition holds with parameters* (r, δ) *and a constant scaling function* $\varphi(X) = \sqrt{\frac{2}{\pi}}$ *with probability of at least* $1 - Ce^{-cm\delta^4}$*, provided that* $m \gtrsim \frac{dr\log(1/\delta)}{\delta^4}$.

Proposition 2 is a special case of a more general theorem on the sign-RIP condition with noisy measurements, which will be discussed in details in the next section. Recalling the definition of $\mathcal{M}(X)$, the above proposition implies the uniform convergence of the random sets $\mathrm{Sign}(\langle A_i, X\rangle)A_i$ to their expected value $\sqrt{\frac{2}{\pi}}\frac{X}{\|X\|_F}$. Similar guarantees have also been derived for other variants of RIP condition, such as ℓ_2- and ℓ_1/ℓ_2-RIP. However, our result is fundamentally different, since it does not rely on the usual covering arguments based on the Lipschitzness of the gradients.

Intuition behind our analysis. Before providing the formal convergence analysis of SubGD, we first provide the intuition behind its behavior. Suppose that Proposition 2 holds with sufficiently

small δ. Then, we have $D_t \approx \sqrt{\frac{2}{\pi}} \frac{(U_t U_t - X^*)U_t}{\|U_t U_t - X^*\|_F}$ for every $D_t \in \partial f_{\ell_1}(U_t)$, and the iterations of SubGD can be approximated as

$$U_{t+1} \approx U_t - \eta_t \cdot \sqrt{\frac{2}{\pi}} \frac{(U_t U_t - X^*)U_t}{\|U_t U_t - X^*\|_F}. \tag{4}$$

Consequently, with the choice of $\eta_t = \eta_0 \sqrt{\frac{\pi}{2}} \|U_t U_t - X^*\|_F$, the iterations of SubGD reduce to

$$U_{t+1} \approx U_t - \eta_0 \cdot (U_t U_t - X^*)U_t \tag{5}$$

which are precisely the iterations of GD with a constant step size η_0, applied to the ℓ_2-loss function $\|UU^\top - X^*\|_F^2$. This implies that, under the sign-RIP condition, SubGD enjoys a similar implicit regularization property to GD with a constant step size. A caveat of this analysis is that the proposed step size is in terms of $\|U_t U_t - X^*\|_F$, which is not known *a priori*. To address this issue, we again invoke the sign-RIP condition, and show that η_t can be accurately estimated as

$$\eta_t = \frac{\pi}{2} \eta_0 \frac{1}{m} \|\mathbf{y} - \mathcal{A}(U_t U_t)\|_1 \approx \eta_0 \sqrt{\frac{\pi}{2}} \|U_t U_t - X^*\|_F$$

Inspired by this intuition, we proceed with the formal convergence analysis of SubGD.

Formal analysis. Suppose that $X^* = u^* u^{*\top}$ for $u^* \in \mathbb{R}^{d \times 1}$. Moreover, without loss of generality, we assume that $\|u^*\| = 1$. Inspired by Li et al. [2018], we decompose the solution U_t as

$$U_t = u^* u^{*\top} U_t + \left(1 - u^* u^{*\top}\right) U_t := u^* r_t^\top + E_t, \tag{6}$$

where $r_t = U_t^\top u^*$ is called *signal term*, and $E_t = (1 - u^* u^{*\top}) U_t$ is referred to as *error term*, which is the projection of U_t onto the orthogonal complement of the subspace spanned by u^*. Evidently, we have $U_t U_t^\top = X^\star$ if and only if $\|r_t\| = 1$ and $\|E_t\|_F = 0$. More generally, our next lemma shows that the error $\|U_t U_t - X^*\|_F$ can be precisely controlled in terms of $\|E_t\|_F$ and $\|r_t\|$.

Lemma 1. *The following inequality holds:*

$$\left\| U_t U_t^\top - X^\star \right\|_F^2 \leq \left(1 - \|r_t\|^2\right)^2 + 2\|E_t\|^2 \|r_t\|^2 + \|E_t\|_F^4. \tag{7}$$

Based on Lemma 1, we provide a high-level idea of our proof technique:

1. Since the algorithm is initialized at $U_0 = \alpha B$ with sufficiently small scalar α and an orthonormal matrix B, both the signal and error terms $\|r_0\|$ and $\|E_0\|_F$ are guaranteed to be small.

2. It is shown that the signal term $\|r_t\|^2$ approaches 1 at a geometric rate. Therefore, $1 - \|r_t\|^2$ converges to zero at a same rate.

3. Moreover, it is proven that the error term $\|E_t\|_F$ grows *linearly*, and its growth rate is significantly slower than that of the signal term.

4. This discrepancy in the growth rates of the signal and error terms ensures that after a certain number of iterations T, the signal term $\|r_t\|$ is sufficiently close to 1, while the error term $\|E_t\|_F$ remains small. Combined with Lemma 1, this establishes the convergence of SubGD with early stopping of the algorithm.

More precisely, the following propositions characterize the dynamics of the signal and error terms.

Proposition 3 (Error Dynamics). *Assume that the measurements are noiseless and satisfy the sign-RIP with parameters (r', δ), and constant scaling function $\varphi(X) = \sqrt{\frac{2}{\pi}}$. Moreover, suppose that $\|E_t\|_F \leq 1, \|r_t\| \leq 2, \delta \leq \frac{1}{2}$, and the step size η_t is chosen as (5) with $\eta_0 \leq \frac{2}{45}$. Then, we have*

$$\|E_{t+1}\|_F \leq \|E_t\|_F + 22\delta\eta_0. \tag{8}$$

$$\|E_{t+1}\| \leq \|E_t\| + 15\delta\eta_0. \tag{9}$$

Proposition 4 (Signal Dynamics). *Under the conditions of Proposition 3, we have*

$$\left\| r_{t+1} - \left(1 + \eta_0(1 - \|r_t\|^2)\right) r_t \right\| \leq 10\eta_0\delta(\|E_t\| + \|r_t\|) + 2\eta_0 \|E_t\|^2 \|r_t\|. \tag{10}$$

Equipped with these propositions, we are ready to establish the convergence of SubGD in the noiseless setting.

Theorem 2. *Assume that the measurements are noiseless and satisfy the sign-RIP condition with parameters (r', δ), $\delta \lesssim 1$, and constant scaling function $\varphi(X) = \sqrt{\frac{2}{\pi}}$. Suppose that $\alpha \asymp \sqrt{\frac{1}{r'}}\delta$ and the step size η_t is chosen as (5) with $\eta_0 \lesssim 1$. Then, after $T \asymp \frac{\log(r'/\delta)}{\eta_0}$ iterations, we have*

$$\left\| U_T U_T^\top - X^\star \right\|_F^2 \lesssim \delta^2 \log^2\left(\frac{r'}{\delta}\right). \tag{11}$$

The above theorem implies that, for any $r' \geq 1$ (including $r' = d$), SubGD converges to the true low-rank solution at a (nearly) *linear* and *dimension-free* rate without any explicit regularization or rank constraint, provided that the measurements are noiseless and satisfy the sign-RIP condition. Interestingly, our convergence rate matches that of GD in similar problems [Hou et al., 2020, Chen et al., 2019]. In the next subsection, we extend this result to the noisy case, where the measurements are grossly corrupted with noise.

5.2 Noisy Case

Motivated by the desirable performance of SubGD in the noiseless setting, we study its behavior when a subset of measurements are contaminated with noise. In particular, suppose that $\mathbf{y} = \mathcal{A}(X^*) + \mathbf{s}$, where the noise vector is generated according to the following model:

Assumption 1 (Noise model). *Given a corruption probability p, the noise vector $\mathbf{s} \in \mathbb{R}^m$ is generated as follows: first, a subset $\mathcal{S} \subset \{1, 2, \ldots, m\}$ with cordiality pm is selected uniformly at random. Then, for every $i \in \mathcal{S}$, the element s_i is randomly drawn from a zero mean distribution, i.e., $s_i \overset{i.i.d.}{\sim} \mathbb{P}$ and $\mathbb{E}[s_i] = 0$. Finally, $s_i = 0$ for every $i \notin \mathcal{S}$.*

Remark 1. *Our results hold under an alternative (and equivalent) noise model studied in [Bai et al., 2018], where $s_i \overset{i.i.d.}{\sim} \mathbb{P}$ with probability p, and $s_i = 0$ with probability $1 - p$.*

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.

Next, we show that, with sufficiently large number of Gaussian measurements, the sign-RIP condition is satisfied with an appropriate choice of scaling function, even if a constant fraction of the measurements are corrupted with noise.

Proposition 5. *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* $\mathbf{s}$ *satisfies Assumption 1. Then, the sign-RIP condition holds with parameters* (r, δ) *and a scaling function* $\varphi(X) = \sqrt{\frac{2}{\pi}}\left(1 - p + p\mathbb{E}\left[e^{-s_i^2/(2\|X\|_F)}\right]\right)$ *with probability of at least* $1 - Ce^{-cm\delta^4}$*, provided that* $m \gtrsim \frac{dr\left(\log\left(\frac{1}{(1-p)\delta}\right)\vee 1\right)}{\delta^4(1-p)^4}$.

A number of observations can be made based on Proposition 5. First, our result does not impose any restriction on the corruption probability p, which improves upon the assumption $p < 1/2$ made in [Li et al., 2020b] for the robust matrix recovery problem. Moreover, it shows that the sign-RIP condition holds *irrespective* of the magnitude of the noise values. As will be shown later, this property does not hold for the ℓ_2-RIP condition, implying that GD is vulnerable to large noise values in the over-parameterized matrix recovery (see Figure 1a). Our proof technique for Proposition 5 is based on a novel *localization* method that eschews the need for Lipschitzness of the gradients. Moreover, it is worth noting that our result on the sign-RIP can be extended to more general sub-Gaussian measurement matrices.

Intuition behind our analysis. Similar to the noiseless case, first we present the intuition behind our analysis. Assuming that the measurements satisfy sign-RIP condition with parameters (r, δ) and an arbitrary scaling function $\varphi(X)$, every element of the sub-differential $\partial f_{\ell_1}(U_t)$ is concentrated around its expected value, i.e.,

$$D_t \approx \varphi(U_tU_t - X^*)\frac{(U_tU_t - X^*)U_t}{\|U_tU_t - X^*\|_F}$$

for every $D_t \in \partial f_{\ell_1}(U_t)$. Inspired by our analysis in the noiseless case, a possible choice of step size would be $\eta_t = \eta_0\varphi(U_tU_t - X^*)^{-1}\|U_tU_t - X^*\|_F$, which ensures that SubGD behaves as GD applied to $\|UU^\top - X^*\|_F^2$. However, this choice of η_t relies on the explicit value of $\|U_tU_t - X^*\|_F$, which is unknown. Moreover, unlike the noiseless case, the value of $\|U_tU_t - X^*\|_F$ *cannot* be estimated via $\|\mathbf{y} - \mathcal{A}(U_tU_t^\top)\|_1$ because of two reasons: (i) due to the presence of noise, the ℓ_1-loss function $\|\mathbf{y} - \mathcal{A}(U_tU_t^\top)\|_1$ is no longer an unbiased estimator of $\sqrt{\frac{2}{\pi}}\|U_tU_t^\top - X^*\|_F$; and (ii) the value of $\|\mathbf{y} - \mathcal{A}(U_tU_t^\top)\|_1$ is sensitive to the magnitude of the noise, i.e., its variance grows with the variance of the noise. To alleviate this issue, we propose the following alternative choice of step size:

$$\eta_t = \frac{\eta_0}{\|D\|_F}\rho^t \tag{12}$$

where $D \in \mathcal{M}(U_tU_t - X^*)$, and $\rho < 1$ is a predefined decay rate. Due to sign-RIP condition, we have $\|D\|_F \approx \varphi(U_tU_t^\top - X^*)$, which implies

$$U_{t+1} \approx U_t - \eta_0\rho^t\frac{\left(U_tU_t^\top - X^\star\right)U_t}{\left\|U_tU_t^\top - X^\star\right\|_F}. \tag{13}$$

A closer look at (13) reveals that, at every iteration, SubGD points to the negative gradient of $\|U_tU_t^\top - X^*\|_F^2$, while the geometrically decaying step size $\eta_t = \eta_0\rho^t$ guarantees the convergence of the algorithm.

Formal analysis. Inspired by this intuition, we present the counterparts of Propositions 3 and 4 in the noisy setting, characterizing the error and signal dynamics.

Proposition 6 (Signal Dynamics). *Assume that the measurements are noiseless and satisfy the sign-RIP with parameters (r', δ) and a strictly positive and uniformly bounded scaling function $\varphi(X)$. Moreover, suppose that $\|E_t\|_F \leq 1, \|r_t\| \leq 2, \delta \leq \frac{1}{2}$, and the step size η_t is chosen as (12). Then, we have*

$$\begin{aligned}\left\| r_{t+1} - \left(1 + \frac{\eta_0 \rho^t (1 - \|r_t\|^2)}{\left\|U_t U_t^\top - X^\star\right\|_F}\right) r_t \right\| &\leq 2\delta\eta_0\rho^t(\|E_t\| + \|r_t\|) + \frac{2\eta_0\rho^t}{\left\|U_t U_t^\top - X^\star\right\|_F} \|E_t\|^2 \|r_t\| \\ &\quad + \frac{2\delta\eta_0\rho^t}{\left\|U_t U_t^\top - X^\star\right\|_F}(1 - \|r_t\|^2)\|r_t\|\,.\end{aligned} \tag{14}$$

Proposition 7 (Error Dynamics). *Suppose that the conditions of Proposition 6 are satisfied and $\eta_0 \lesssim \delta \lesssim \left\|U_t U_t^\top - X^*\right\|_F$. Then, we have*

$$\|E_{t+1}\|_F \leq \|E_t\|_F + 10\delta\eta_0\rho^t, \tag{15}$$

$$\|E_{t+1}\| \leq \|E_t\| + 2\delta\eta_0\rho^t(\|r_t\| + \|E_t\|). \tag{16}$$

Based on the signal and error dynamics, we present our main theorem on the convergence of SubGD in the noisy case.

Theorem 3. *Assume that the measurements are noiseless and satisfy the sign-RIP condition with parameters (r', δ), $\delta \lesssim 1$, and a strictly positive and uniformly bounded scaling function $\varphi(X)$. Moreover, suppose that $\alpha \asymp \sqrt{\frac{1}{r'}}\delta$, and the step size η_t is chosen as (12) with $\eta_0 \lesssim \delta$ and $\rho \asymp 1 - \eta_0/\log\frac{1}{\alpha}$. Then, after $T \asymp \log\left(\frac{r'}{\delta}\right)/\eta_0$ iterations, we have*

$$\left\|U_T U_T^\top - X^\star\right\|_F^2 \lesssim \delta^2 \log^2\left(\frac{r'}{\delta}\right). \tag{17}$$

Remark 2. *Comparing Theorems 2 and 3, it can be seen that SubGD is slower in the noisy case by a factor of $1/\delta$. This is due to a smaller value of η_0, which is implied by a different choice of step size. Verifying whether SubGD can be accelerated in the noisy setting to achieve the same convergence rate as in the noiseless case is considered as an enticing challenge for future research.*

The effect of scaling function. It can be seen that both signal and error dynamics are independent of the scaling function $\varphi(X)$. This is due to the special choice of the step size: roughly speaking, the sign-RIP condition implies that the chosen step size is proportional to $\varphi(U_t U_t^\top - X^*)^{-1}$, thereby cancelling the effect of the scaling function in the dynamics. This implies that the step size changes with the corruption probability, whose effect is captured via the scaling function. To see this, recall that for the Gaussian measurements, the scaling function takes the form $\varphi(U_t U_t - X^*) = \sqrt{\frac{2}{\pi}}(1-p) + \sqrt{\frac{2}{\pi}} p \mathbb{E}\left[e^{-s_i^2/(2\|U_t U_t - X^*\|_F)}\right]$. It is easy to see that, for small values of $\|U_t U_t - X^*\|_F$ (or alternatively, large values of noise), the scaling function can be well-approximated as $\varphi(U_t U_t - X^*) \approx \sqrt{\frac{2}{\pi}}(1-p)$. This implies that SubGD automatically takes more aggressive steps with increasing corruption probability.

Combining Theorem 3 and Proposition 5 leads to an end-to-end sample complexity guarantee for SubGD with Gaussian measurements.

Corollary 1. *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* $\mathbf{s}$ *satisfies Assumption 1. Moreover, suppose that* α, η_t, *and* T *are chosen according to Theorem 3. Then, SubGD satisfies the error bound (17) with an overwhelming probability, provided that* $m \gtrsim \frac{dr \log\left(\frac{1}{(1-p)\delta}\right)}{\delta^4(1-p)^4}$.

5.3 Comparison to ℓ_2-Loss

As mentioned before, another line of works focuses on characterizing the implicit regularization of GD in the over-parameterized matrix recovery with ℓ_2-loss function $f_{\ell_2}(U) = \frac{1}{2m}\|\mathbf{y} - \mathcal{A}(UU^\top)\|^2$. Therefore, it is important to study the behavior of GD in the noisy setting, and compare its performance to SubGD. It is easy to verify that the gradient of $f_{\ell_2}(U)$ can be written as $\nabla f_{\ell_2}(U) = Q(UU^\top - X^*)U$, where

$$Q(UU^\top - X^*) = \frac{1}{m}\sum_{i=1}^m \Big(\langle A_i, UU^\top - X^*\rangle + s_i\Big) A_i.$$

Recently, Li et al. [2018] showed that GD with constant step size converges to a solution U_T that satisfies $U_T {U_T}^\top \approx X^*$, provided that $Q(X)$ is close to X for every rank-r matrix X. In the noiseless setting, it is known that $Q(X) \approx X$ under the so-called ℓ_2-RIP condition. However, in the noisy setting, our next proposition reveals that $Q(X)$ and X may be far apart, especially when the noise has high variance.

Proposition 8. *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* $\mathbf{s}$ *satisfies Assumption 1 with* $s_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. *Then, we have*

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

The above proposition implies that, in order to guarantee the convergence of GD to the true low-rank solution, the number of measurements should satisfy $m \gtrsim (1+p\sigma^2)d^2$. Comparing this bound with Proposition 5 reveals the crucial advantage of SubGD over GD: if the measurements are corrupted with large outliers, GD requires significantly larger number of samples to converge, since its sample complexity increases with the noise variance.

6 Conclusion and Future Work

Contrary to the conventional wisdom in statistical learning, It is by now a well-known fact that over-parameterization in some modern learning tasks may avoid overfitting and lead to better generalization. At the heart of this seemingly contradictory phenomenon lies the *implicit regularization* of simple local search algorithms, such as gradient descent (GD). However, it is unknown whether the implicit regularization of these algorithms can be carried over to *robust* learning tasks, where the available data may be grossly corrupted with noise. To demystify this question, we study the robust matrix recovery problem, where the goal is to recover a low-rank matrix, given a limited number of linear measurements that are potentially corrupted with noise. In particular, we show that a

simple sub-gradient method (SubGD) can recover the true low-rank matrix in the noisy setting, when it is applied to the over-parameterized ℓ_1-loss function without any explicit regularization or rank constraint. Moreover, by introducing a new notion of restricted isometry property, called *sign-RIP*, we show that SubGD is robust against large outliers. In particular, we show that, with Gaussian measurements, SubGD recovers the true low-rank solution, even if an arbitrary fraction of the measurements are contaminated with large noise values.

Lastly, we believe that our results can be improved in a variety of directions. First, based on our extensive simulation results (see supplementary file), we conjecture that early stopping may *not* be necessary for the convergence of SubGD in the over-parameterized regime, due to the geometrically decaying nature of the step size. Second, we intend to extend our results to the settings where the true solution has a general rank r. Finally, we will investigate whether it is possible to obtain a near-linear convergence rate for SubGD in the noisy variant of the problem.

References

Tayo Ajayi, David Mildebrath, Anastasios Kyrillidis, Shashanka Ubaru, Georgios Kollias, and Kristofer Bouchard. Provably convergent acceleration in factored gradient descent with applications in matrix sensing. *arXiv preprint arXiv:1806.00534*, 2018.

Yu Bai, Qijia Jiang, and Ju Sun. Subgradient descent learns orthogonal dictionaries. *arXiv preprint arXiv:1810.10702*, 2018.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016.

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 Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *arXiv preprint arXiv:2004.12019*, 2020.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1):5–37, 2019.

Damek Davis and Dmitriy Drusvyatskiy. Graphical convergence of subgradients in nonconvex optimization and learning. *arXiv preprint arXiv:1810.07590*, 2018.

Bin Dong, Jikai Hou, Yiping Lu, and Zhihua Zhang. Distillation $\approx$ early stopping? harvesting dark knowledge utilizing anisotropic information retrieval for overparameterized neural network. *arXiv preprint arXiv:1910.01255*, 2019.

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*, 21(59):1–51, 2020.

Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, pages 8745–8756, 2018.

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.

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.

Thomas Y Hou, Zhenzhen Li, and Ziyun Zhang. Fast global convergence for low-rank matrix recovery via riemannian gradient descent with random initialization. *arXiv preprint arXiv:2012.15467*, 2020.

Cedric Josz, Yi Ouyang, Richard Zhang, Javad Lavaei, and Somayeh Sojoudi. A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization. In *Advances in neural information processing systems*, pages 2441–2449, 2018.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 4313–4324. PMLR, 2020a.

Xiao Li, Zhihui Zhu, Anthony Mancho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. *Siam Journal on Optimization*, 30(1):660–686, 2020b.

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. pages 2–47, 2018.

Song Mei, Yu Bai, Andrea Montanari, et al. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 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.

Bodhisattva Sen. A gentle introduction to empirical process theory and applications. 2018.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Roman Vershynin. High-dimensional probability, 2019.

Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust pca via gradient descent. *arXiv preprint arXiv:1605.07784*, 2016.

Chong You, Zhihui Zhu, Qing Qu, and Yi Ma. Robust recovery via implicit bias of discrepant learning rates for double over-parameterization. *arXiv preprint arXiv:2006.08857*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

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

Teng Zhang and Yi Yang. Robust pca by manifold optimization. *The Journal of Machine Learning Research*, 19(1):3101–3139, 2018.

A Numerical Experiments

In this section, we provide extensive numerical experiments to verify our theoretical guarantees, and to shed light on possible future directions.

A.1 Relationship between dimension and measurement number

In this experiment, we verify the dependency between the number of measurements m and dimension d. Our theoretical result suggests that $m \gtrsim dr'$ is enough to ensure the convergence of SubGD. In this experiment, we choose d from 10 to 100 and set $r' = d$. Moreover, we set the corruption probability to $p = 0.1$. Moreover, each element of the noise is generated according to a standard Gaussian distribution. The step sizes are selected as $\eta_t = \eta_0 \rho^t$, where $\eta_0 = 0.4$ and $\rho = 0.98$. For each group of parameters, we run 5 independent trials and plot the average log-residual for the last iteration in Figure 2. It can be seen that, in order to ensure the same value for the error, the number of measurements should grow almost linearly with the dimension.

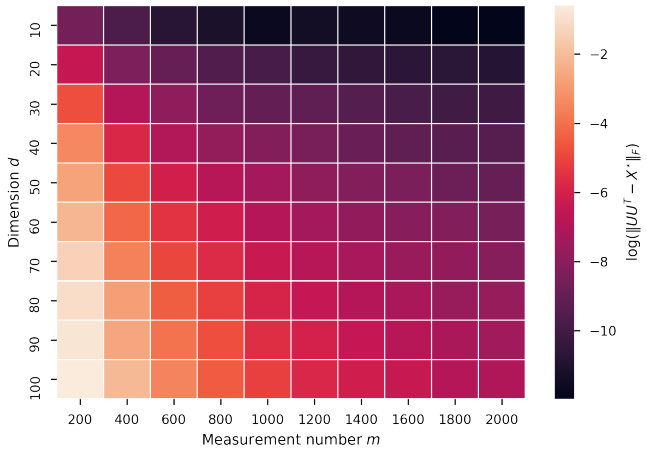


Figure 2: The error with respect to the number of measurements and dimension

A.2 Effect of Different Noise Magnitudes

In this experiment, we certify the robustness of SubGD against large noise values. Our theoretical result suggests that the convergence of SubGD is independent of the noise magnitude. To verify this, we set the dimension and the number of measurements to $d = 50$ and $m = 500$. Moreover, we set the corruption probability to $p = 0.1$, and select each element of the noise according to a Gaussian distribution $s_i \sim \mathcal{N}(0, \sigma^2)$ with varying variance σ^2. Finally, we set the step size to $\eta_t = \eta_0 \rho^t$, where $\eta_0 = 0.25$, and $\rho = 0.99$. Based on Figure 3, it can be seen that increasing variance slightly deteriorates the error. However, beyond a certain threshold, increasing variance does not have any effect on the error.

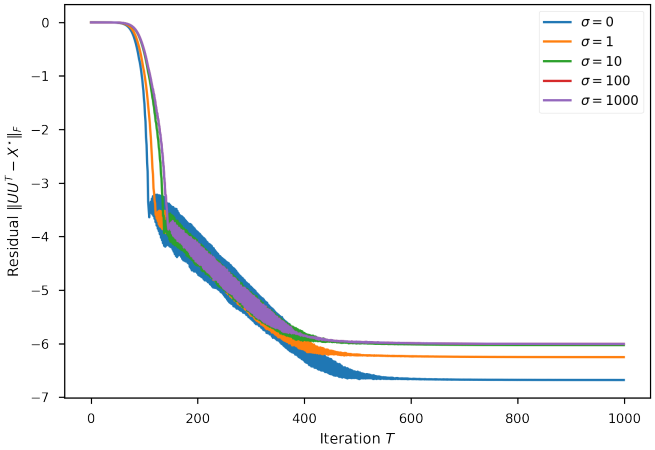


Figure 3: Effect of noise variance.

A.3 Effect of Different Types of Noise

In this experiment, we study the effect of different types of noise on the performance of SubGD. In particular, we choose five different types of distribution for the noise: Gaussian, uniform, Laplace, Cauchy, and Rademacher. The experiments are designed under the same settings as Subsection A.2. Moreover, for all types of noise (except for the Cauchy distribution), we set the variance to 100. As can be seen in Figure 4, SubGD is insensitive to the particular choice of noise.

A.4 Effect of the upper bound on the rank

Next, we investigate the effect of different values of r' on the performance of SubGD. The experiments are designed under the same settings as Subsection A.2. Figure 5 shows the behavior of SubGD for different values of r'. It can be seen that SubGD with $r' = 1$ has a significantly slower convergence rate at the earlier iterations. However, after the first phase, it convergences to an infinitesimal error at a linear rate. On the other hand, SubGD with $r' > 1$ has a faster convergence rate at the earlier iterations, and then platuaus at a higher error level.

A.5 Effect of different step size regimes

Finally, we explore the effect of different step sizes in both noiseless and noisy case under the same settings as Section A.2. In the noiseless case, we compare four different types of step sizes: 1) $\eta_t = \eta_0 \rho^t$ with $\eta_0 = 0.25, \rho = 0.99$; 2) $\eta_t = \frac{\eta_0}{t}$ with $\eta_0 = 2.0$; 3) $\eta_t = \frac{\eta_0}{\sqrt{t}}$ with $\eta_0 = 0.3$; and 4) our proposed choice $\eta_t = \eta_0 \frac{1}{m} \left\| \mathbf{y} - \mathcal{A}(U_t U_t^\top) \right\|_1$ where $\eta_0 = 0.25$. As can be seen in Figure 6a, SubGD converges to the true solution with all of the aforementioned step sizes. However, our proposed step size leads to the fastest convergence rate. In the noisy case, we compare the performance of five different step sizes: 1) $\eta_t = \eta_0 \rho^t$ with $\eta_0 = 0.45, \rho = 0.98$; 2) $\eta_t = \frac{\eta_0}{t}$ with $\eta_0 = 2.0$; 3) $\eta_t = \frac{\eta_0}{\sqrt{t}}$ with $\eta_0 = 0.3$; 4) $\eta_t = \eta_0 \frac{1}{m} \left\| \mathbf{y} - \mathcal{A}(U_t U_t^\top) \right\|_1$ with $\eta_0 = 0.25$; 5) our proposed choice $\eta_t = \frac{\eta_0}{\|D_t\|_F} \rho^t$,

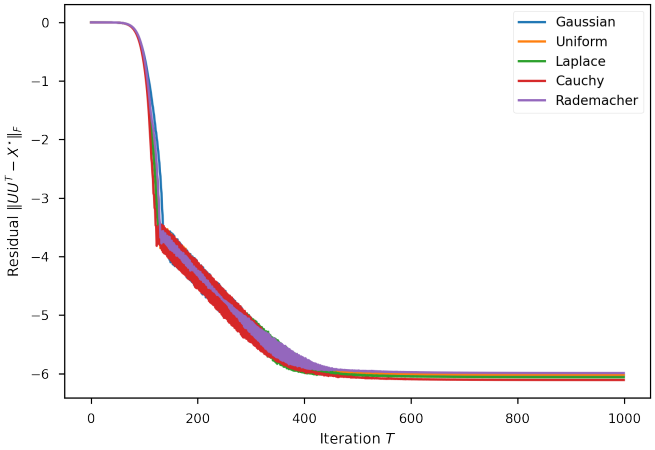


Figure 4: Effect of different types of noise.

where $D_t \in \mathcal{M}(U_t U_t - X^*)$, $\eta_0 = 0.4$, and $\rho = 0.99$. From Figure 6b, it is evident that the step size $\eta_t = \eta_0 \frac{1}{m} \left\| \mathbf{y} - \mathcal{A}(U_t U_t^\top) \right\|_1$, which was the best choice in the noiseless case, does not result in the convergence of SubGD to the true solution in the noisy case. As mentioned before, this is due to the sensitivity of $\frac{1}{m} \left\| \mathbf{y} - \mathcal{A}(U_t U_t^\top) \right\|_1$ to outliers. Moreover, our proposed step size outperforms its vanilla counterpart. Finally, the polynomially decaying step size $\eta_t \propto \frac{1}{t}$ performs slightly better than our proposed step size. Motivated by this interesting observation, we will study the performance of SubGD with polynomially decaying step sizes in the future.

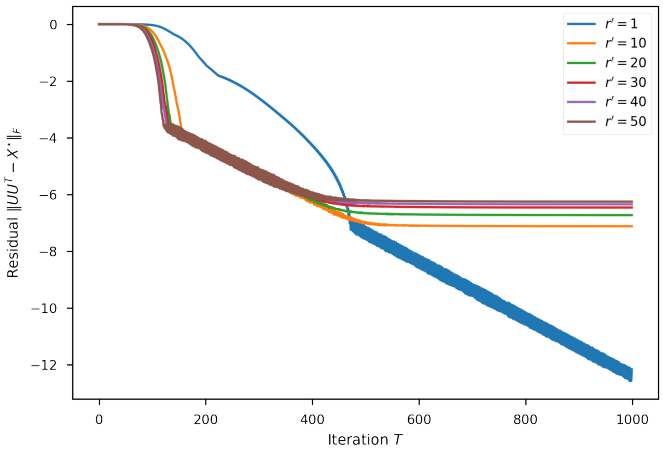
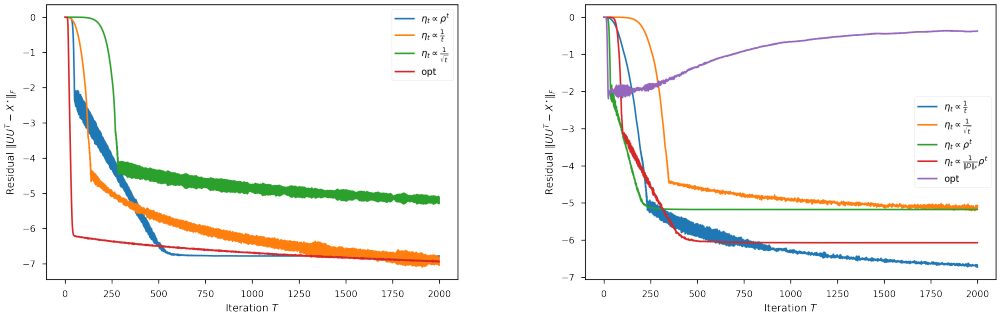


Figure 5: Effect of r'.



(a) noiseless (b) noisy

Figure 6: Effect of different step size regimes.

B Proofs for Sign-RIP

In this section, we provide the proofs for Propositions 2 and 5. As a first step, we start with the noiseless case, and show that a weaker version of Proposition 2 can be obtained directly from the so-called ℓ_1/ℓ_2-RIP condition.

Lemma 2 (l_1/l_2-RIP, Proposition 1 in [Li et al., 2020b]). *Let $r \geq 1$ be given, suppose measurements $\{A_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries with $m \gtrsim dr$. Then for any $0 < \delta < \sqrt{\frac{2}{\pi}}$, there exists a universal constant $c > 0$, such that with probability of at least $1 - e^{-cm\delta^2}$, we have*

$$\sup_{X \in \mathbb{S}_r} \left| \frac{1}{m} \sum_{i=1}^m |\langle A_i, X \rangle| - \sqrt{\frac{2}{\pi}} \|X\|_F \right| \leq \sqrt{\frac{2}{\pi}} \delta. \tag{18}$$

Based on the above ℓ_1/ℓ_2-RIP condition, we proceed to prove a weaker version of Proposition 2:

Proposition 9. *Assume that the measurement matrices $\{A_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries, and that $\mathbf{s} = 0$. Then, the sign-RIP condition holds with parameters $(r, \delta), \delta \leq \sqrt{\frac{2}{\pi}}$ and a constant scaling function $\varphi(X) = \sqrt{\frac{2}{\pi}}$ with probability of at least $1 - Ce^{-cm\delta^4}$, provided that $m \gtrsim d^2$.*

Proof. Without loss of generality, we assume that $\|X\|_F = 1$. For any given $0 < \delta \leq \sqrt{\frac{2}{\pi}}$ and any $D(X) \in \frac{1}{m} \sum_{i=1}^m \mathrm{Sign}(\langle A_i, X \rangle) A_i$, we have

$$\begin{aligned}
&\sup_{X \in \mathbb{S}_r} \left\| D(X) - \sqrt{\frac{2}{\pi}} X \right\|_F^2 \\
\overset{(a)}{=} &\sup_{X \in \mathbb{S}_r} \|D(X)\|_F^2 - \sqrt{\frac{8}{\pi}} \frac{1}{m} \sum_{i=1}^m |\langle A_i, X \rangle| + \frac{2}{\pi} \\
\leq &\sup_{X \in \mathbb{S}_r} \|D(X)\|_F^2 - \sqrt{\frac{8}{\pi}} \inf_{X \in \mathbb{S}_r} \frac{1}{m} \sum_{i=1}^m |\langle A_i, X \rangle| + \frac{2}{\pi} \\
\overset{(b)}{\leq} &\sup_{X \in \mathbb{S}_r} \|D(X)\|_F^2 + \sqrt{\frac{8}{\pi}} \sqrt{\frac{2}{\pi}} \delta - \sqrt{\frac{8}{\pi}} \sqrt{\frac{2}{\pi}} + \frac{2}{\pi} \\
= &\sup_{X \in \mathbb{S}_r} \|D(X)\|_F^2 + \frac{4}{\pi} \delta - \frac{2}{\pi}
\end{aligned} \tag{19}$$

with probability of at least $1 - Ce^{-cm\delta^2}$. Here, (a) follows from $\langle D(X), X \rangle = \frac{1}{m} \sum_{i=1}^m |\langle A_i, X \rangle|$, and (b) uses ℓ_1/ℓ_2-RIP condition from Lemma 2. Then, recall that for an arbitrary $M \in \mathbb{R}^{d \times d}$, we have

$$\|M\|_F = \sup_{Y \in \mathbb{S}} \langle M, Y \rangle . \tag{20}$$

This implies

$$\begin{aligned}\sup_{X\in\mathbb{S}_r}\|D(X)\|_F^2 &\le \sup_{X,Y\in\mathbb{S}}\left(\langle D(X),Y\rangle\right)^2\\ &\overset{(c)}{=}\sup_{Y\in\mathbb{S}}\left(\frac{1}{m}\sum_{i=1}^m|\langle A_i,Y\rangle|\right)^2\\ &\overset{(d)}{\le}\frac{2}{\pi}(1+\delta)^2\\ &\overset{(e)}{\le}\frac{2}{\pi}+\frac{6}{\pi}\delta\end{aligned}\tag{21}$$

with high probability $1-Ce^{-cm\delta^2}$. Here, (c) uses the fact that for a fixed Y, the supremum over X is taken exactly at $X=Y$, (d) uses the ℓ_1/ℓ_2-RIP condition, and (e) uses the assumption $\delta\le 1$.

Combining these two parts, we obtain

$$\sup_{X\in\mathbb{S}}\left\|D(X)-\sqrt{\frac{2}{\pi}}X\right\|_F^2\le\frac{10}{\pi}\delta\tag{22}$$

with probability of at least $1-Ce^{-c'm\delta^2}$. Therefore, upon choosing $\delta'=\sqrt{5\delta}$, we obtain

$$\sup_{X\in\mathbb{S}}\left\|D(X)-\sqrt{\frac{2}{\pi}}X\right\|_F=\sup_{X,Y\in\mathbb{S}}\left\langle D(X)-\sqrt{\frac{2}{\pi}}X,Y\right\rangle\le\sqrt{\frac{2}{\pi}}\delta'\tag{23}$$

with probability of at least $1-Ce^{-cm\delta'^4}$. □

Despite its simplicity, the above analysis has two major drawbacks: (1) its sample complexity scales with d^2, as opposed to dr in Proposition 2, (2) it is not clear how to extend this analysis to the noisey case. To address these issues and prove Propositions 2 and 5, we need a more in-depth analysis of the sign-RIP condition. First, we provide an intermediate lemma.

Lemma 3. *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 $\mathbf{s}$ satisfies Assumption 1. Then, for every $D\in\mathcal{M}(X)$, we have*

$$\mathbb{E}[D]=\sqrt{\frac{2}{\pi}}\left(1-p+p\mathbb{E}\left[e^{-s_i^2/(2\|X\|_F)}\right]\right)\frac{X}{\|X\|_F}\tag{24}$$

where the expectation is taken with respect to both $\mathbf{s}$ and $\{A_i\}_{i=1}^m$.

Proof. To prove this lemma, it is enough to show that for any $X,Y\in\mathbb{R}^{d\times d}$, we have

$$\mathbb{E}\left[\mathrm{Sign}\left(s+\langle A,X\rangle\right)\langle A,Y\rangle\right]=\sqrt{\frac{2}{\pi}}\mathbb{E}\left[e^{-s^2/2\|X\|_F^2}\right]\left\langle\frac{X}{\|X\|_F},Y\right\rangle.\tag{25}$$

provided that A is Gaussian and s has zero mean. Without loss of generality. suppose that

$\|X\|_F = \|Y\|_F = 1$. Let us denote $u := \langle A, X\rangle, v := \langle A, Y\rangle, \rho := \mathrm{Cov}(u, v) = \langle X, Y\rangle$. Then

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Sign}\left(s + \langle A, X\rangle\right)\langle A, Y\rangle\right] &= \mathbb{E}\left[\mathrm{Sign}\left(s+u\right)v\right] \\
&\overset{(a)}{=} \rho\mathbb{E}\left[\mathrm{Sign}(u+s)u\right] \\
&= \rho\mathbb{E}_s\left[\int_{-s}^{\infty} u\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du - \int_{-\infty}^{-s} u\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du\right] \\
&= \rho\mathbb{E}_s\left[\int_{-s}^{\infty} u\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du + \int_{s}^{\infty} u\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du\right] \\
&= 2\rho\mathbb{E}_s\left[\int_{|s_i|}^{\infty} u\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du\right] \\
&= \sqrt{\frac{2}{\pi}}\langle X, Y\rangle\,\mathbb{E}_s\left[\int_{|s|}^{\infty} d\left(-e^{-u^2/2}\right)\right] \\
&= \sqrt{\frac{2}{\pi}}\langle X, Y\rangle\,\mathbb{E}_s\left[e^{-s^2/2}\right].
\end{aligned} \tag{26}$$

Here (a) uses the fact that $v|u, s \sim \mathcal{N}(\rho u, 1-\rho^2)$ since $s \perp\!\!\!\perp u, v$. This together with the variational form of the Frobenius norm implies

$$\mathbb{E}\left[\mathrm{Sign}(s + \langle A, X\rangle)A\right] = \sqrt{\frac{2}{\pi}}\mathbb{E}\left[e^{-s^2/2\|X\|_F^2}\right]\frac{X}{\|X\|_F}, \tag{27}$$

for any $X \in \mathbb{R}^{d\times d}$. On the other hand, it is easy to verify that $\mathbb{E}\left[\mathrm{Sign}\left(\langle A, X\rangle\right)A\right] = \sqrt{\frac{2}{\pi}}\frac{X}{\|X\|_F}$. The proof is completed by noting that the size of the noisy measurements is equal to pm. □

B.1 Proof of Proposition 5

For the sake of simplicity, we assume that pm is an integer. Moreover, we abuse the notation and use $\mathrm{Sign}(\cdot)$ as a regular function taking an arbitrary value $\mathrm{Sign}(0) \in [-1, 1]$. To prove Proposition 5, we first present an intermediate lemma, which holds for any fixed $Y, Y \in \mathbb{S}$.

Lemma 4. *There exists a universal constant c, for any $\delta > 0$, we have*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}(\langle A_i, X\rangle + s_i)\langle A_i, Y\rangle - \sqrt{\frac{2}{\pi}}\left(1 - p + p\mathbb{E}\left[e^{-s_i^2/2}\right]\right)\langle X, Y\rangle\right| \geq \delta\right) \leq 2e^{-cm\delta^2}. \tag{28}$$

Proof of Lemma 4. We first show that $\mathrm{Sign}(\langle A_i, X\rangle + s_i)\langle A_i, Y\rangle$ is a sub-Gaussian random variable. First notice that $\langle A_i, Y\rangle \sim \mathcal{N}(0, 1)$ since $\|Y\|_F = 1$. Moreover, notice that $\|\mathrm{Sign}(\langle A_i, X\rangle + s_i)\langle A_i, Y\rangle\|_{\ell^{2k}} \leq \|\langle A_i, Y\rangle\|_{\ell^{2k}}$ for $\forall k \in \mathbb{N}_+$, where $\|M\|_{\ell^{2k}}$ is defined as $\left(\mathbb{E}\left[|M|^p\right]\right)^{1/p}$. Therefore, based on equivalent definition of sub-Gaussian random variables (see Definition 2), we obtain that $\mathrm{Sign}(\langle A_i, X\rangle + s_i)\langle A_i, Y\rangle$ is also $O(1)$-sub-Gaussian. Moreover, according to the proof of Lemma 3, we have $\mathbb{E}\left[\frac{1}{m}\sum_{i\in S}\mathrm{Sign}(\langle A_i, X\rangle)\langle A_i, Y\rangle\right] = \sqrt{\frac{2}{\pi}}(1-p)\langle X, Y\rangle$ and $\mathbb{E}\left[\frac{1}{m}\sum_{i\in S}\mathrm{Sign}(\langle A_i, X\rangle + s_i)\langle A_i, Y\rangle\right] = \sqrt{\frac{2}{\pi}}p\mathbb{E}\left[e^{-s_i^2/2}\right]\langle X, Y\rangle$. This together with the standard concentration bound on sub-Gaussian random variables leads to

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i\notin S}\text{Sign}(\langle A_i,X\rangle)\langle A_i,Y\rangle-\sqrt{\frac{2}{\pi}}(1-p)\langle X,Y\rangle\right|\geq\frac{1}{2}\delta\right)\leq 2e^{-cm\delta^2/(1-p)}, \tag{29}$$

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i\in S}\text{Sign}(\langle A_i,X\rangle+s_i)\langle A_i,Y\rangle-\sqrt{\frac{2}{\pi}}p\mathbb{E}\left[e^{-s_i^2/2}\right]\langle X,Y\rangle\right|\geq\frac{1}{2}\delta\right)\leq 2e^{-cm\delta^2/p}. \tag{30}$$

which implies

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\text{Sign}(\langle A_i,X\rangle+s_i)\langle A_i,Y\rangle-\sqrt{\frac{2}{\pi}}\left(1-p+p\mathbb{E}\left[e^{-s_i^2/2}\right]\right)\langle X,Y\rangle\right|\geq\delta\right)\\ \leq&4e^{-cm\delta^2\min\left\{\frac{1}{p},\frac{1}{1-p}\right\}}\leq 2e^{-c'm\delta^2}.\end{aligned} \tag{31}$$

□

Consider an ϵ-covering $\mathbb{S}_{\epsilon,r}\subseteq\mathbb{S}_r$ with a property that for every $X\in\mathbb{S}_r$, there exists $\bar{X}\in\mathbb{S}_{\epsilon,r}$ that satisfies $\|X-\bar{X}\|_F\leq\epsilon$. According to Lemma 13, there exists an ϵ-covering that satisfies $|\mathbb{S}_{\epsilon,r}|\leq\left(\frac{9}{\epsilon}\right)^{(2d+1)r}$. For any $\bar{X}\in\mathbb{S}_{\epsilon,r}$, define $B_r(\bar{X},\epsilon)=\{X\in\mathbb{S}_r:\left\|X-\bar{X}\right\|_F\leq\epsilon\}$. Then, for any $\bar{X},\bar{Y}$ and $X,Y\in B_r(\bar{X},\epsilon)\times B_r(\bar{Y},\epsilon)$, we have

$$\left|\langle X,Y\rangle-\left\langle\bar{X},\bar{Y}\right\rangle\right|\leq\left|\left\langle X-\bar{X},\bar{Y}\right\rangle\right|+\left|\left\langle X,Y-\bar{Y}\right\rangle\right|\leq 2\epsilon. \tag{32}$$

Based on the defined ϵ-covering, one can write

$$\begin{aligned}&\sup_{X,Y\in\mathbb{S}_r}\left|\frac{1}{m}\sum_{i=1}^{m}\text{Sign}(\langle A_i,X\rangle+s_i)\langle A_i,Y\rangle-\sqrt{\frac{2}{\pi}}\left(1-p+p\mathbb{E}\left[e^{-s_i^2/2}\right]\right)\langle X,Y\rangle\right|\\ =&\sup_{\bar{X},\bar{Y}\in\mathbb{S}_{\epsilon,r}}\sup_{\substack{X\in B_r(\bar{X},\epsilon)\\ Y\in B_r(\bar{Y},\epsilon)}}\left|\frac{1}{m}\sum_{i=1}^{m}\text{Sign}(\langle A_i,X\rangle+s_i)\langle A_i,Y\rangle-\sqrt{\frac{2}{\pi}}\left(1-p+p\mathbb{E}\left[e^{-s_i^2/2}\right]\right)\langle X,Y\rangle\right|\\ \leq&\underbrace{\sup_{\bar{X},\bar{Y}\in\mathbb{S}_{\epsilon,r}}\left|\frac{1}{m}\sum_{i=1}^{m}\text{Sign}(\left\langle A_i,\bar{X}\right\rangle+s_i)\left\langle A_i,\bar{Y}\right\rangle-\sqrt{\frac{2}{\pi}}\left(1-p+p\mathbb{E}\left[e^{-s_i^2/2}\right]\right)\left\langle\bar{X},\bar{Y}\right\rangle\right|}_{\text{(A)}}\\ &+\underbrace{\sup_{\bar{X},\bar{Y}\in\mathbb{S}_{\epsilon,r}}\sup_{Y\in B_r(\bar{Y},\epsilon)}\left|\frac{1}{m}\sum_{i=1}^{m}\text{Sign}(\left\langle A_i,\bar{X}\right\rangle+s_i)\langle A_i,Y\rangle-\text{Sign}(\left\langle A_i,\bar{X}\right\rangle+s_i)\left\langle A_i,\bar{Y}\right\rangle\right|}_{\text{(B)}}\\ &+\underbrace{\sup_{\bar{X}\in\mathbb{S}_{\epsilon,r},Y\in\mathbb{S}_r}\sup_{X\in B(\bar{X},\epsilon)}\left|\frac{1}{m}\sum_{i=1}^{m}\text{Sign}(\left\langle A_i,\bar{X}\right\rangle+s_i)\langle A_i,Y\rangle-\text{Sign}(\langle A_i,X\rangle+s_i)\langle A_i,Y\rangle\right|}_{\text{(C)}}\\ &+\underbrace{\sup_{\bar{X},\bar{Y}\in\mathbb{S}_{\epsilon,r}}\sup_{\substack{X\in B_r(\bar{X},\epsilon)\\ Y\in B_r(\bar{Y},\epsilon)}}\sqrt{\frac{2}{\pi}}\left(1-p+p\mathbb{E}\left[e^{-s_i^2/2}\right]\right)\left|\langle X,Y\rangle-\left\langle\bar{X},\bar{Y}\right\rangle\right|}_{\leq\sqrt{\frac{8}{\pi}}\epsilon\text{ by (32)}}.\end{aligned} \tag{33}$$

We control the first three terms separately. Based on a union bound and Lemma 4, we have

$$\text{(A)} \le \delta_1 \quad \text{with probability of at least } 1 - 2\left|\mathbb{S}_{\epsilon,r}\right|^2 e^{-cm\delta_1^2}. \tag{34}$$

Moreover, one can write

$$\begin{aligned}
\text{(B)} &\le \sup_{\bar{Y}\in\mathbb{S}_{\epsilon,r}} \sup_{Y\in B_r(\bar{Y},\epsilon)} \frac{1}{m}\sum_{i=1}^m \left|\left\langle A_i, Y-\bar{Y}\right\rangle\right| \\
&\overset{(a)}{\le} \epsilon \sup_{Z\in\mathbb{S}_{2r}} \frac{1}{m}\sum_{i=1}^m \left|\left\langle A_i, Z\right\rangle\right| \\
&\le \sqrt{\frac{2}{\pi}}\epsilon(1+\delta_2)
\end{aligned} \tag{35}$$

with probability of at least $1 - Ce^{c_1 dr\log\frac{1}{\delta_2} - c_2 m\delta_2^2}$. Here we used ℓ_1/ℓ_2 condition from Lemma 2, and the fact for X, Y with ranks at most r, we have $\operatorname{rank}(X-Y) \le \operatorname{rank}(X) + \operatorname{rank}(Y) \le 2r$. Next, we provide an upper bound for (C). First by Cauchy-Schwartz inequality, we have

$$\text{(C)} \le \sup_{\bar{X}\in\mathbb{S}_{\epsilon,r}} \sup_{X\in B_r(\bar{X},\epsilon)} \left(\underbrace{\frac{1}{m}\sum_{i=1}^m \left(\operatorname{Sign}(\left\langle A_i, \bar{X}\right\rangle + s_i) - \operatorname{Sign}(\langle A_i, X\rangle + s_i)\right)^2}_{\text{(C1)}}\right)^{\frac{1}{2}} \sup_{Y\in\mathbb{S}_r}\left(\frac{1}{m}\sum_{i=1}^m \langle A_i, Y\rangle^2\right)^{\frac{1}{2}}. \tag{36}$$

The second term in the above inequality can be readily controlled via ℓ_2-RIP (see Lemma 14):

$$\sup_{Y\in\mathbb{S}_r} \frac{1}{m}\sum_{i=1}^m \langle A_i, Y\rangle^2 \le 1+\delta_3 \tag{37}$$

which holds with probability of at least $1 - Ce^{c_1 dr\log\frac{1}{\delta_3} - c_2 m\delta_3^2}$ for any $0 < \delta_3 < 1$. For the remaining part (C1), first note that if $\left|\left\langle A_i, X-\bar{X}\right\rangle\right| \le \left|\left\langle A_i, \bar{X}+s_i\right\rangle\right|$, then $\operatorname{Sign}(\left\langle A_i, \bar{X}\right\rangle + s_i) = \operatorname{Sign}(\langle A_i, X\rangle + s_i)$. This leads to

$$\begin{aligned}
\sup_{\bar{X}\in\mathbb{S}_{\epsilon,r}} \sup_{X\in B_r(\bar{X},\epsilon)} \text{(C1)} &\le \sup_{\bar{X},X} \frac{4}{m}\sum_{i=1}^m \mathbb{1}\left(\left|\left\langle A_i, X-\bar{X}\right\rangle\right| \ge \left|\left\langle A_i, \bar{X}\right\rangle + s_i\right|\right) \\
&\overset{(a)}{\le} \sup_{X,\bar{X}} \frac{4}{m}\sum_{i=1}^m \mathbb{1}\left(\left|\left\langle A_i, X-\bar{X}\right\rangle\right| \ge t\right) + \mathbb{1}\left(\left|\left\langle A_i, \bar{X}\right\rangle + s_i\right| \le t\right) \\
&\le \sup_{Z\in\epsilon\mathbb{S}_{2r}} \frac{4}{m}\sum_{i=1}^m \mathbb{1}\left(\left|\left\langle A_i, Z\right\rangle\right| \ge t\right) + \sup_{\bar{X}} \frac{4}{m}\mathbb{1}\left(\left|\left\langle A_i, \bar{X}\right\rangle + s_i\right| \le t\right) \\
&\overset{(b)}{\le} \sup_{Z\in\epsilon\mathbb{S}_{2r}} \frac{4}{m}\sum_{i=1}^m \mathbb{1}\left(\left|\left\langle A_i, Z\right\rangle\right| \ge t\right) + 4\mathbb{E}\left[\mathbb{1}\left(\left|\left\langle A_i, \bar{X}\right\rangle + s_i\right| \le t\right)\right] + \delta_4 \\
&\overset{(c)}{\le} \sup_{Z\in\epsilon\mathbb{S}_{2r}} \frac{4}{m}\sum_{i=1}^m \mathbb{1}\left(\left|\left\langle A_i, Z\right\rangle\right| \ge t\right) + 4t + \delta_4
\end{aligned} \tag{38}$$

which holds with probability of at least $1 - C|\mathbb{S}_{\epsilon,r}|e^{-cm\delta_4^2}$, where $t > 0$ is an arbitrary scalar. Here, in (a) we use a simple fact that for two arbitrary random variables A, B and a scalar $t \in \mathbb{R}$, the event $\{A \geq B\}$ is included in $\{A \geq t\} \cup \{B \leq t\}$. Moreover, in (b) we use a union bound and Hoeffding's inequality. Finally, in (c) we use the anti-concentration inequality conditioned on s_i. For the first term in the above inequality, we have the following lemma.

Lemma 5. *We have*

$$\mathbb{E}\left[\sup_{Z\in\epsilon\mathbb{S}_{2r}} \frac{4}{m}\sum_{i=1}^{m} \mathbb{1}\left(|\langle A_i, Z\rangle| \geq t\right)\right] \lesssim e^{-t^2/4\epsilon^2}\sqrt{\frac{dr}{m}} \vee \frac{dr}{m}, \tag{39}$$

moreover, for fixed $0 < \delta < 1$, we have the following tail bound

$$\mathbb{P}\left(\left|\sup_{Z\in\epsilon\mathbb{S}_{2r}} \frac{4}{m}\sum_{i=1}^{m} \mathbb{1}\left(|\langle A_i, Z\rangle| \geq t\right) - \mathbb{E}\left[\sup_{Z\in\epsilon\mathbb{S}_{2r}} \frac{4}{m}\sum_{i=1}^{m} \mathbb{1}\left(|\langle A_i, Z\rangle| \geq t\right)\right]\right| > \delta\right) \leq 2e^{-cm\delta^2}. \tag{40}$$

Proof. The tail bound directly follows from Theorem 8.5 in [Sen, 2018]. Here we only give the proof sketch for the expectation bound. To apply Theorem 8.7 in [Sen, 2018], we only need to upper bound

$$\sigma^2 := \sup_{Z\in\epsilon\mathbb{S}_{2r}} \mathrm{Var}(\mathbb{1}(|\langle A, Z\rangle| > t)). \tag{41}$$

Note that

$$\begin{aligned}
\sigma^2 \leq & \sup_{Z\in\epsilon\mathbb{S}_{2r}} \mathrm{Var}(\mathbb{1}(|\langle A, Z\rangle| > t)) \\
\leq & \sup_{Z\in\epsilon\mathbb{S}_{2r}} \mathbb{E}[\mathbb{1}(|\langle A, Z\rangle| > t)] \\
\leq & \sup_{W\in\mathbb{S}} \mathbb{P}(|\langle A, W\rangle| > t/\epsilon) \\
\leq & 2e^{-t^2/2\epsilon^2},
\end{aligned} \tag{42}$$

where in the last inequality, we used the tail bound for Gaussian random variables. Therefore, by Theorem 8.7 in [Sen, 2018], we have

$$\mathbb{E}\left[\sup_{Z\in\epsilon\mathbb{S}_{2r}} \frac{4}{m}\sum_{i=1}^{m} \mathbb{1}\left(|\langle A_i, Z\rangle| \geq t\right)\right] \lesssim \sigma\sqrt{\frac{dr}{m}\log\frac{1}{\sigma}} \vee \frac{dr}{m}\log\frac{1}{\sigma} \lesssim e^{-t^2/4\epsilon^2}\sqrt{\frac{dr}{m}} \vee \frac{dr}{m}. \tag{43}$$

This completes the proof. □

Based on Lemma 5, we have

$$\sup_{\bar{X},X} (\mathrm{C1}) \lesssim e^{-t^2/4\epsilon^2}\sqrt{\frac{dr}{m}} \vee \frac{dr}{m} + 4t + \delta_4 + \delta_5 \tag{44}$$

with probability of at least $1 - C|\mathbb{S}_{\epsilon,r}|e^{-cm\delta_4^2} - Ce^{-cm\delta_5^2}$. Combining all derived bounds, we have

$$\begin{aligned}
&\sup_{X,Y\in\mathbb{S}_r} \left|\frac{1}{m}\sum_{i=1}^{m} \mathrm{Sign}(\langle A_i, X\rangle + s_i)\langle A_i, Y\rangle - \left(1 - p + p\mathbb{E}\left[e^{-s_i^2/2}\right]\right)\langle X, Y\rangle\right| \\
\leq & \delta_1 + C\epsilon(1+\delta_2) + \sqrt{1+\delta_3}\sqrt{e^{-t^2/4\epsilon^2}\sqrt{\frac{dr}{m}} + \frac{dr}{m} + 4t + \delta_4 + \delta_5}
\end{aligned} \tag{45}$$

with probability of at least $1-2\left|\mathbb{S}_{\epsilon,r}\right|^2 e^{-cm\delta_1^2}-Ce^{-cm\delta_2^2}-Ce^{c_1 dr\log\frac{1}{\delta_3}-c_2 m\delta_3^2}-C\left|\mathbb{S}_{\epsilon,r}\right| e^{-cm\delta_4^2}-Ce^{-cm\delta_5^2}$. Upon choosing $\delta_2=\delta_3=\frac{1}{2}$, $\delta_4=\delta_5=\delta^2$, $\delta_1=\delta$, $\epsilon=\delta$, $t=\delta^2$, and $m\gtrsim dr(\log\frac{1}{\delta}\vee 1)/\delta^4$, we have

$$\sup_{X,Y\in\mathbb{S}}\left|\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}(\langle A_i,X\rangle+s_i)\langle A_i,Y\rangle-\sqrt{\frac{2}{\pi}}\left(1-p+p\mathbb{E}\left[e^{-s_i^2/2}\right]\right)\langle X,Y\rangle\right|\lesssim\delta, \tag{46}$$

with probability of at least $1-Ce^{-cm\delta^4}$. This leads to

$$\sup_{X\in\mathbb{S},D\in\mathcal{M}(X)}\left\|D-\sqrt{\frac{2}{\pi}}\left(1-p+p\mathbb{E}\left[e^{-s_i^2/2}\right]\right)X\right\|_F\lesssim\delta \tag{47}$$

Finally, note that

$$\sqrt{\frac{2}{\pi}}\left(1-p+p\mathbb{E}\left[e^{-s_i^2/2}\right]\right)\geq\sqrt{\frac{2}{\pi}}(1-p). \tag{48}$$

Therefore, we have

$$\sup_{X\in\mathbb{S},D\in\mathcal{M}(X)}\left\|D-\sqrt{\frac{2}{\pi}}\left(1-p+p\mathbb{E}\left[e^{-s_i^2/2}\right]\right)X\right\|_F\lesssim\sqrt{\frac{2}{\pi}}\left(1-p+p\mathbb{E}\left[e^{-s_i^2/2}\right]\right)\delta, \tag{49}$$

with probability of at least $1-Ce^{-cm\delta^4}$, given $m\gtrsim\frac{dr\left(\log(\frac{1}{(1-p)\delta})\vee 1\right)}{(1-p)^4\delta^4}$. □

C Proofs for the Noiseless Case

C.1 Proof of Lemma 1

Due to (6), one can write

$$\begin{aligned}
&\left\|U_tU_t^\top-X^\star\right\|_F^2\\
=&\left\|(u^\star r_t^\top+E_t)(r_tu^{\star\top}+E_t^\top)-u^\star u^{\star\top}\right\|_F^2\\
=&\left\|\underbrace{(\|r_t\|^2-1)u^\star u^{\star\top}}_{\text{(A)}}+\underbrace{E_tE_t^\top}_{\text{(B)}}+\underbrace{E_tr_tu^{\star\top}}_{\text{(C)}}+\underbrace{u^\star r_t^\top E_t^\top}_{\text{(D)}}\right\|_F^2.
\end{aligned} \tag{50}$$

Now, note that $\|\mathrm{A}\|_F^2=\left(1-\|r_t\|^2\right)^2$, and

$$\|\mathrm{C}\|_F=\|\mathrm{D}\|_F=\|E_tr_t\|\,\|u^\star\|=\|E_tr_t\|. \tag{51}$$

On the other hand, $u^{\star\top}E_t=u^{\star\top}(I-u^\star u^{\star\top})U_t=0$, and therefore, $\langle\mathrm{A},\mathrm{B}\rangle=0$. Similarly, $\langle\mathrm{A},\mathrm{C}\rangle=(\|r_t\|^2-1)\mathrm{Tr}\left(u^\star u^{\star\top}E_tr_tu^{\star\top}\right)=0$, and $\langle\mathrm{A},\mathrm{D}\rangle=0$ since $\mathrm{Tr}\left(u^\star u^{\star\top}u^\star r_t^\top E_t^\top\right)=\mathrm{Tr}\left(u^\star u^{\star\top}U_tU_t^\top(I-u^\star u^{\star\top})\right)=\mathrm{Tr}\left(U_tU_t^\top(I-u^\star u^{\star\top})u^\star u^{\star\top}\right)=0$. Similarly, we have $\langle\mathrm{B},\mathrm{C}\rangle=0,\langle\mathrm{B},\mathrm{D}\rangle=0,\langle\mathrm{C},\mathrm{D}\rangle=0$. This leads to

$$\begin{aligned}\left\|U_tU_t^\top - X^\star\right\|_F^2 &= \left(1-\|r_t\|^2\right)^2 + 2\|E_tr_t\|^2 + \left\|E_tE_t^\top\right\|_F^2 \\ &\le \left(1-\|r_t\|^2\right)^2 + 2\|E_t\|^2\|r_t\|^2 + \|E_t\|_F^4.\end{aligned} \tag{52}$$

□

C.2 Proof of Proposition 3

For simplicity of notation, we define $\Delta_t = U_tU_t^\top - X^*$ throughout the proof. First, we provide a useful fact, which will be widely used in our subsequent arguments.

Fact 1. *Suppose $\|E_t\|_F \le 1, \|r_t\| \le 2$, then by Lemma 1, we have $\|\Delta_t\|_F^2 \le 1 + 2\times 4 + 1 = 10$.*

We first prove the error dynamics under a general learning rate η_t.

Lemma 6. *Suppose that $\|E_t\|_F \le 1, \|r_t\| \le 2$, $\delta \le \frac{1}{2}$, then, the following inequalities hold*

$$\|E_{t+1}\|_F^2 \le \|E_t\|_F^2 + \sqrt{\frac{8}{\pi}}\eta_t\left(-\frac{\left\|E_tU_t^\top\right\|_F^2}{\|\Delta_t\|_F} + \delta\left\|E_tU_t^\top\right\|_F\right) + \frac{20}{\pi}\eta_t^2\left(\delta^2 + \left\|E_tU_t^\top\right\|_F^2/\|\Delta_t\|_F^2\right), \tag{53}$$

$$\|E_{t+1}\| \le \left\|I - \frac{\eta_tU_t^\top U_t}{\sqrt{\frac{\pi}{2}}\|\Delta_t\|_F}\right\|\cdot\|E_t\| + \sqrt{\frac{2}{\pi}}\delta\eta_t(\|r_t\| + \|E_t\|). \tag{54}$$

Proof. For simplicity, we denote $M_t \in \frac{1}{m}\sum_{i=1}^m \mathrm{Sign}(\langle A_i, U_tU_t^\top - X^\star\rangle)A_i$, and $\bar{M}_t = \sqrt{\frac{2}{\pi}}\frac{\Delta_t}{\|\Delta_t\|_F}$. It is easy to verify that

$$E_{t+1} = E_t - \eta_t(I - u^\star u^{\star\top})M_tU_t, \tag{55}$$

Based on the above equation, one can write

$$\|E_{t+1}\|_F^2 = \|E_t\|_F^2 - 2\eta_t\left\langle E_t, (I-u^\star u^{\star\top})M_tU_t\right\rangle + \eta_t^2\left\|(I-u^\star u^{\star\top})M_tU_t\right\|_F^2. \tag{56}$$

Next, we will provide separate upper bounds for the second and third terms in the above equation. First, note that

$$\begin{aligned}\langle E_t, M_tU_t\rangle &= \left\langle M_t, E_tU_t^\top\right\rangle \\ &\overset{(a)}{\ge} \frac{1}{\sqrt{\frac{\pi}{2}}\|\Delta_t\|_F}\left\langle\Delta_t, E_tU_t^\top\right\rangle - \sqrt{\frac{2}{\pi}}\delta\left\|E_tU_t^\top\right\|_F \\ &\overset{(b)}{=} \sqrt{\frac{2}{\pi}}\frac{\left\|E_tU_t^\top\right\|_F^2}{\|\Delta_t\|_F} - \sqrt{\frac{2}{\pi}}\delta\left\|E_tU_t^\top\right\|_F.\end{aligned} \tag{57}$$

where we used sign-RIP condition in (a), and (b) follows from $\left\langle U_tU_t^\top - X^\star, E_tU_t^\top\right\rangle = \left\langle E_tU_t^\top, E_tU_t^\top\right\rangle = \left\|E_tU_t^\top\right\|_F^2$. On the other hand, we have

$$\left\langle E_t, u^\star u^{\star\top}M_tU_t\right\rangle = \mathrm{Tr}\left(E_t^\top u^\star u^{\star\top}M_tU_t\right) = \mathrm{Tr}\left(U_t^\top(I-u^\star u^{\star\top})u^\star u^{\star\top}M_tU_t\right) = 0. \tag{58}$$

Combining (57) and (58) leads to

$$-2\eta_t \left\langle E_t, (I - u^\star u^{\star\top}) M_t U_t \right\rangle \le -2\eta_t \left(\sqrt{\frac{2}{\pi}} \frac{\left\| E_t U_t^\top \right\|_F^2}{\|\Delta_t\|_F} - \sqrt{\frac{2}{\pi}} \delta \left\| E_t U_t^\top \right\|_F \right) \tag{59}$$

Now, we provide an upper bound for the third term in (56). One can write

$$\begin{aligned} \left\| (I - u^\star u^{\star\top}) M_t U_t \right\|_F^2 &\le 2 \left\| (I - u^\star u^{\star\top})(M_t - \bar{M}_t) U_t \right\|_F^2 + 2 \left\| (I - u^\star u^{\star\top}) \bar{M}_t U_t \right\|_F^2 \\ &\overset{(a)}{\le} 2 \left\| (M_t - \bar{M}_t) U_t \right\|_F^2 + \frac{4}{\pi} \frac{\left\| E_t U_t^\top U_t \right\|_F^2}{\|\Delta_t\|_F^2} \\ &\overset{(b)}{\le} \frac{4}{\pi} \delta^2 \|U_t\|_F^2 + \frac{4}{\pi} \left\| E_t U_t^\top \right\|_F^2 \|U_t\|_F^2 / \|\Delta_t\|_F^2 \\ &\overset{(c)}{\le} \frac{20}{\pi} \delta^2 + \frac{20}{\pi} \left\| E_t U_t^\top \right\|_F^2 / \|\Delta_t\|_F^2 . \end{aligned} \tag{60}$$

where we used the contraction of projection and $(I - u^\star u^{\star\top})(U_t U_t^\top - X^\star) U_t = E_t U_t^\top U_t$ in (a). Moreover, (b) directly follows from the sign-RIP condition. Finally, we used the following fact in (c).

Fact 2. *Assuming* $\|E_t\|_F \le 1, \|r_t\| \le 2$, *we have* $\|U_t\|_F^2 = \|r_t\|^2 + \|E_t\|_F^2 \le 1^2 + 2^2 = 5$.

Finally, combining all the three terms, we finally have:

$$\|E_{t+1}\|_F^2 \le \|E_t\|_F^2 + \sqrt{\frac{8}{\pi}} \eta_t \left(- \frac{\left\| E_t U_t^\top \right\|_F^2}{\|\Delta_t\|_F} + \delta \left\| E_t U_t^\top \right\|_F \right) + \frac{20}{\pi} \eta_t^2 \left(\delta^2 + \frac{\left\| E_t U_t^\top \right\|_F^2}{\|\Delta_t\|_F^2} \right). \tag{61}$$

Now, we turn to control the spectral norm. First, notice that

$$\begin{aligned} &\left\| (I - u^\star u^{\star\top}) M_t U_t - (I - u^\star u^{\star\top}) \frac{\Delta_t U_t}{\sqrt{\frac{\pi}{2}} \|\Delta_t\|_F} \right\| \\ \overset{(a)}{\le} &\left\| M_t U_t - \frac{\Delta_t U_t}{\sqrt{\frac{\pi}{2}} \|\Delta_t\|_F} \right\| \\ \le &\left\| M_t - \bar{M}_t \right\| \cdot \|U_t\| \\ \le &\sqrt{\frac{2}{\pi}} \delta \|U_t\| \\ \le &\sqrt{\frac{2}{\pi}} \delta (\|E_t\| + \|r_t\|). \end{aligned} \tag{62}$$

Here in (a) we used the contraction of projection. On the other hand, observing that $(I -$

$u^\star u^{\star\top})(U_tU_t^\top - X^\star)U_t = E_tU_t^\top U_t$, we have

$$\begin{aligned}
\|E_{t+1}\| &= \left\|E_t - \eta_t(I - u^\star u^{\star\top})M_tU_t\right\| \\
&\le \left\|E_t - \eta_t(I - u^\star u^{\star\top})\bar{M}_tU_t\right\| + \eta_t\left\|(I - u^\star u^{\star\top})(M_t - \bar{M}_t)U_t\right\| \\
&\le \left\|E_t\left(I - \frac{\eta_tU_t^\top U_t}{\sqrt{\frac{\pi}{2}}\|\Delta_t\|_F}\right)\right\| + \sqrt{\frac{2}{\pi}}\delta\eta_t(\|r_t\| + \|E_t\|) \\
&\le \left\|I - \frac{\eta_tU_t^\top U_t}{\sqrt{\frac{\pi}{2}}\|\Delta_t\|_F}\right\| \cdot \|E_t\| + \sqrt{\frac{2}{\pi}}\delta\eta_t(\|r_t\| + \|E_t\|),
\end{aligned} \tag{63}$$

which completes the proof. □

Before presenting the proof of Proposition 3, we need the following intermediate result

Lemma 7. *Suppose that the measurements satisfy sign-RIP with parameters (δ, r) and a constant scaling function $\varphi(X) = \sqrt{\frac{2}{\pi}}$. Then, for every $X \in \mathbb{S}_r$, we have*

$$\left|\frac{1}{m}\|\mathcal{A}(X)\|_1 - \sqrt{\frac{2}{\pi}}\|X\|_F\right| \le \sqrt{\frac{2}{\pi}}\delta \tag{64}$$

Proof. Due to the sign-RIP condition, we have $\left\|D - \sqrt{\frac{2}{\pi}}X\right\|_F \le \sqrt{\frac{2}{\pi}}\delta$ for every $X \in \mathbb{S}_r$ and $D \in \frac{1}{m}\sum_{i=1}^m \mathrm{Sign}(\langle A_i, X\rangle)A_i$. This implies that

$$\sqrt{\frac{2}{\pi}}\delta = \sup_{Y\in\mathbb{S}}\left\langle D - \sqrt{\frac{2}{\pi}}X, Y\right\rangle \ge \left\langle D - \sqrt{\frac{2}{\pi}}X, X\right\rangle \ge \frac{1}{m}\|\mathcal{A}(X)\|_1 - \sqrt{\frac{2}{\pi}}\|X\|_F \tag{65}$$

Similarly, it can be shown that $\sqrt{\frac{2}{\pi}}\delta \ge -\frac{1}{m}\|\mathcal{A}(X)\|_1 + \sqrt{\frac{2}{\pi}}\|X\|_F$. This completes the proof. □

Based on the above lemma, the sign-RIP condition implies the ℓ_1/ℓ_2-RIP condition. Given Lemmas 6 and 7, we are ready to present the proof of Proposition 3.

Proof of Proposition 3. It is enough to show that the choice of $\eta_t = \frac{\pi}{2}\eta_0 \cdot \frac{1}{m}\sum\left|\langle A_i, U_tU_t^\top - X^\star\rangle\right|$ in Lemma 6 leads to the desired bounds. Based on Lemma 7 and $\delta \le \frac{1}{2}$, we have

$$\eta_t \le \sqrt{\frac{\pi}{2}}\eta_0\|\Delta_t\|_F(1+\delta) \le \sqrt{\frac{9\pi}{8}}\eta_0\|\Delta_t\|_F. \tag{66}$$

Similarly, we have the lower bound $\eta_t \ge \sqrt{\frac{\pi}{8}}\eta_0\|\Delta_t\|_F$. Substituting these inequalities in Lemma 6 leads to

$$\begin{aligned}
\|E_{t+1}\|_F^2 &\le \|E_t\|_F^2 - \eta_0\left\|E_tU_t^\top\right\|_F^2 + 3\delta\eta_0\|\Delta_t\|_F\left\|E_tU_t^\top\right\|_F + \frac{45}{2}\eta_0^2\left(\delta^2\|\Delta_t\|_F^2 + \left\|E_tU_t^\top\right\|_F^2\right) \\
&\overset{(a)}{\le} \|E_t\|_F^2 + 3\delta\eta_0\|\Delta_t\|_F\left\|E_tU_t^\top\right\|_F + \frac{45}{2}\delta^2\eta_0^2\|\Delta_t\|_F^2 \\
&\overset{(b)}{\le} \|E_t\|_F^2 + 3\sqrt{10}\delta\eta_0\|E_t\|_F\|U_t\|_F + 225\delta^2\eta_0^2 \\
&\overset{(c)}{\le} \|E_t\|_F^2 + 22\delta\eta_0\|E_t\|_F + 225\delta^2\eta_0^2 \\
&= (\|E_t\|_F + 11\delta\eta_0)^2 + 104\delta^2\eta_0^2.
\end{aligned} \tag{67}$$

where (a) follows from the assumption $\eta_0 \le \frac{2}{45}$, (b) follows form Fact 1, and (c) follows from Fact 2. Therefore, we have

$$\|E_{t+1}\|_F \le \|E_t\|_F + 11\delta\eta_0 + 11\delta\eta_0 = \|E_t\|_F + 22\delta\eta_0. \tag{68}$$

Similarly,

$$\|E_{t+1}\| \le \left\| I - \frac{\eta_t U_t^\top U_t}{\sqrt{\frac{\pi}{2}}\|\Delta_t\|_F} \right\| \cdot \|E_t\| + \sqrt{\frac{2}{\pi}}\delta\eta_t(\|r_t\| + \|E_t\|). \tag{69}$$

Note that

$$\left\|U_t^\top U_t\right\| \le \|U_t\|^2 \le (\|E_t\| + \left\|u^\star r_t^\top\right\|)^2 \le (\|E_t\| + \|r_t\|)^2 \le 9. \tag{70}$$

which, together with $\eta_t \le \sqrt{\frac{9\pi}{8}}\eta_0$, implies

$$\left\| \frac{\eta_t U_t^\top U_t}{\sqrt{\frac{\pi}{2}}\|\Delta_t\|_F} \right\| < 1 \tag{71}$$

Therefore, the first term in the right hand side of the above inequality is upper bounded by $\|E_t\|$. On the other hand, the second term in (69) can be bounded as

$$\sqrt{\frac{2}{\pi}}\delta\eta_t(\|r_t\| + \|E_t\|) \le \frac{3}{2}\delta\eta_0\|\Delta_t\|_F(\|r_t\| + \|E_t\|) \le 15\delta\eta_0 \tag{72}$$

This completes the proof. □

C.3 Proof of Proposition 4

Similarly, we first prove the following lemma which holds for a general choice of η_t.

Lemma 8. *Assuming $\|E_t\|_F \le 1, \|r_t\| \le 2$, we have*

$$\begin{aligned}\left\| r_{t+1} - \left(1 + \frac{\eta_t(1-\|r_t\|^2)}{\sqrt{\frac{\pi}{2}}\left\|U_tU_t^\top - X^\star\right\|_F}\right) r_t \right\| \le{}& \sqrt{\frac{2}{\pi}}\eta_t\delta(\|E_t\| + \|r_t\|) \\ &+ \frac{\eta_t}{\sqrt{\frac{\pi}{2}}\left\|U_tU_t^\top - X^\star\right\|_F}\|E_t\|^2\|r_t\|.\end{aligned} \tag{73}$$

Proof. Recalling the notations $M_t \in \frac{1}{m}\sum_{i=1}^m \mathrm{Sign}(\langle A_i, \Delta_t\rangle)A_i$ and $\bar{M}_t = \sqrt{\frac{2}{\pi}}\frac{\Delta_t}{\|\Delta_t\|_F}$, we have

$$r_{t+1} = r_t - \eta_t U_t^\top M_t^\top u^\star. \tag{74}$$

Therefore,

$$\begin{aligned}\left\| r_{t+1} - r_t + \eta_t \frac{U_t^\top \Delta_t u^\star}{\sqrt{\frac{\pi}{2}}\|\Delta_t\|_F} \right\| &\le \eta_t \left\| U_t^\top (M_t - \bar{M}_t)^\top u^\star \right\| \\ &\le \eta_t \|U_t\| \cdot \left\| M_t - \bar{M}_t \right\| \cdot \|u^\star\| \\ &\le \sqrt{\frac{2}{\pi}}\eta_t\delta(\|E_t\| + \|r_t\|),\end{aligned} \tag{75}$$

where the last inequality follows from the sign-RIP condition. On the other hand, since $U_t^\top \left(U_t U_t^\top - X^\star\right) u^\star = U_t^\top U_t r_t - r_t = \left(r_t r_t^\top + E_t^\top E_t\right) r_t - r_t = \left(\|r_t\|^2 - 1\right) r_t - E_t^\top E_t r_t$, one can write

$$\begin{aligned}
&\left\| r_{t+1} - \left(1 + \frac{\eta_t(1-\|r_t\|^2)}{\sqrt{\frac{\pi}{2}}\|\Delta_t\|_F}\right) r_t \right\| \\
\leq &\sqrt{\frac{2}{\pi}}\eta_t\delta(\|E_t\| + \|r_t\|) + \frac{\eta_t}{\sqrt{\frac{\pi}{2}}\|\Delta_t\|_F}\left\|E_t^\top E_t r_t\right\| \\
\leq &\sqrt{\frac{2}{\pi}}\eta_t\delta(\|E_t\| + \|r_t\|) + \frac{\eta_t}{\sqrt{\frac{\pi}{2}}\|\Delta_t\|_F}\|E_t\|^2\|r_t\|.
\end{aligned} \tag{76}$$

□

Equipped with this lemma and (66), we write

$$\begin{aligned}
\left\| r_{t+1} - \left(1 + \eta_0(1-\|r_t\|^2)\right) r_t \right\| \leq\ & \eta_0\delta(1-\|r_t\|^2)\|r_t\| \\
&+ 2\eta_0\|\Delta_t\|_F\,\delta(\|E_t\| + \|r_t\|) \\
&+ 2\eta_0\|E_t\|^2\|r_t\|.
\end{aligned} \tag{77}$$

Note that the first term is dominated by the second term since $\|\Delta_t\|_F \geq 1 - \|r_t\|^2$ due to (52) and the fact that $\|\Delta_t\|_F \leq \sqrt{10}$. We finally have

$$\left\| r_{t+1} - \left(1 + \eta_0(1-\|r_t\|^2)\right) r_t \right\| \leq 10\eta_0\delta(\|E_t\| + \|r_t\|) + 2\eta_0\|E_t\|^2\|r_t\|. \tag{78}$$

C.4 Proof of Theorem 2

Due to the choice of initial point $U_0 = \alpha B$, we have $\|E_0\| = \|r_0\| = \alpha \leq 1$, and

$$\|E_0\|_F^2 = \alpha^2\|(I - u^* u^{*\top})B\|_F^2 \leq r'\alpha^2 < 1 \tag{79}$$

Hence, the assumptions in Propositions 3 and 4 are valid. Next, we control the Frobenius norm of the error term. From Proposition 3, we have

$$\|E_{t+1}\|_F \leq \|E_t\|_F + 22\delta\eta_0 \tag{80}$$

which implies

$$\|E_T\|_F = \|E_0\|_F + \sum_{t=1}^{T}\left(\|E_t\|_F - \|E_{t-1}\|_F\right) \leq \|E_0\|_F + 22\delta\eta_0 T. \tag{81}$$

Therefore, since $T \asymp \log\left(\frac{r'}{\delta}\right)/\eta_0$ and $\alpha \asymp \sqrt{\frac{1}{r'}}\delta$, we have $\|E_T\|_F \lesssim \sqrt{r'}\alpha + \delta\log\frac{r'}{\alpha} \lesssim \delta + \delta\log\frac{r'}{\alpha} \lesssim \delta\log\frac{r'}{\delta}$. Similarly, due to Proposition 3, we have

$$\|E_{t+1}\| \leq \|E_t\| + 15\delta\eta_0, \tag{82}$$

and $\|E_0\| = \alpha$. Therefore, one can write

$$\|E_T\| = \|E_0\| + \sum_{t=1}^{T} (\|E_t\| - \|E_{t-1}\|) \leq \alpha + 15\delta\eta_0 T \lesssim \delta \log \frac{r'}{\delta}. \tag{83}$$

This shows that the error term remains small throughout the iterations of SubGD. Without loss of generality and to simplify our subsequent analysis, we assume that $\|E_T\| \leq \delta \log \frac{r'}{\delta}$, which can be ensured with sufficiently small η_0. Next, we control the signal term. Due to Proposition 4, we have

$$\|r_{t+1}\| \geq (1 + \eta_0(1 - \|r_t\|^2)) \|r_t\| - 10\eta_0\delta(\|E_t\| + \|r_t\|) - 2\eta_0 \|E_t\|^2 \|r_t\|. \tag{84}$$

Now, we separate our analysis into two stages. In the first stage, we show that the signal grows at a linear rate, provided that $\|r_t\| \leq 1/2$. To show this, we first prove that during the whole training process, the signal term is always larger than the error term.

Lemma 9. *Suppose that $\delta \leq 1/50$. Then, for any $0 \leq t \leq T = \Theta\left(\log \frac{1}{\alpha}/\eta_0\right)$, we have*

$$\|E_t\| \leq \|r_t\|. \tag{85}$$

Proof. We prove this lemma by induction. For the base case, (85) holds since we have $\|E_0\| = \|r_0\| = \alpha$. Now, suppose that (85) holds at time t. Based on (69) and (72), we have

$$\|E_{t+1}\| \leq (1 + 5\eta_0\delta) \|E_t\| + 5\eta_0\delta \|r_t\| \leq (1 + 10\eta_0\delta) \|r_t\|. \tag{86}$$

On the other hand, due to (84), we have

$$\begin{aligned} \|r_{t+1}\| &\geq (1 + \eta_0(1 - \|r_t\|^2)) \|r_t\| - 10\eta_0\delta(\|E_t\| + \|r_t\|) - 2\eta_0 \|E_t\|^2 \|r_t\| \\ &\geq (1 + \eta_0(1 - 3\|r_t\|^2)) \|r_t\| - 20\delta\eta_0 \|r_t\| \\ &\geq \left(1 + \frac{1}{5}\eta_0\right) \|r_t\|. \end{aligned} \tag{87}$$

Here we used the induction hypothesis $\|E_t\| \leq \|r_t\|$. The above two inequalities, together with $\delta \leq 1/50$, imply that $\|E_{t+1}\| \leq \|r_{t+1}\|$. □

During the proof of the above lemma, we showed that

$$\|r_{t+1}\| \geq (1 + \eta_0/5) \|r_t\|. \tag{88}$$

provided that $\delta \leq 1/50$. Now, assuming that $T_1 \gtrsim \log \frac{1}{\alpha}/\eta_0$, we have

$$\|r_{T_1}\| \geq \alpha(1 + \eta_0/5)^{T_1} \geq \frac{1}{2}, \tag{89}$$

This implies that, after T_1 iterations, the signal term will have a norm of at least $1/2$. In the second stage, we assume that $1 \geq \|r_t\| \geq 1/2$. One can write

$$\begin{aligned} \|r_{t+1}\| &\geq (1 + \eta_0(1 - \|r_t\|^2)) \|r_t\| - 10\eta_0\delta(\|E_t\| + \|r_t\|) - 2\eta_0 \|E_t\|^2 \|r_t\| \\ &\geq (1 + \eta_0(1 - \|r_t\|)) \|r_t\| - 20\eta_0\delta \|r_t\| - 4\eta_0\delta^2 \log^2 \frac{r'}{\delta}, \end{aligned} \tag{90}$$

where we used $1-\|r_t\|^2 \geq 1-\|r_t\|$ given $\|r_t\| \leq 1$, and Lemma 9.

For the sake of simplicity, we define $x_t = 1-\|r_{t+T_1}\|$. Hence, (90) can be simplified as

$$\begin{aligned} x_{t+1} &\leq 1-(1-20\eta_0\delta+\eta_0 x_t)(1-x_t)+4\eta_0\delta^2\log^2\frac{r'}{\delta} \\ &\leq (1-\eta_0+20\eta_0\delta)x_t+\eta_0 x_t^2+20\eta_0\delta+4\eta_0\delta^2\log^2\frac{r'}{\delta} \\ &\leq (1-\frac{3}{4}\eta_0)x_t+\frac{1}{2}\eta_0 x_t+20\eta_0\delta\left(1+\delta\log^2\frac{r'}{\delta}\right) \\ &\leq (1-\eta_0/4)x_t+20\eta_0\delta\left(1+\delta\log^2\frac{r'}{\delta}\right), \end{aligned} \tag{91}$$

Here, we used $x_t \leq 1/2$ and $\delta \leq 1/80$. Then, we have

$$x_{t+1}-80\delta\left(1+\delta\log^2\frac{r'}{\delta}\right) \leq \left(1-\frac{\eta_0}{4}\right)\left(x_t-80\delta\left(1+\delta\log^2\frac{r'}{\delta}\right)\right), \tag{92}$$

which implies

$$x_{T_2} \leq 80\delta\left(1+\delta\log^2\frac{r'}{\delta}\right)+\frac{1}{2}\left(1-\eta_0/4\right)^{T_2}. \tag{93}$$

Upon choosing $T_2 \gtrsim \log\frac{1}{\delta}/\eta_0$, we have $x_{T_2} \lesssim \delta \vee \delta^2\log^2\frac{r'}{\delta}$, which is equivalent to $\|r_{T_1+T_2}\| \geq 1-O\left(\delta+\delta^2\log^2\frac{r'}{\delta}\right)$.

This completes the proof under the assumption $\|r_{T_1+T_2}\| \leq 1$. Now, it remains to show that the error bound holds even if $\|r_{T_1+T_2}\| > 1$. To this goal, first we show that $T_3 = \Omega\left(\log\frac{r'}{\delta}/\eta_0\right)$ is necessary to guarantee the convergence of SubGD. In particular, we prove that we need at least $T_3 = \Omega\left(\log\frac{r'}{\delta}/\eta_0\right)$ to ensure $\|r_{T_3}\| \geq \frac{1}{2}$. To this goal, suppose that $\|r_t\| \leq 1/2$ for every $t \leq T$. Due to Proposition 4, we have

$$\begin{aligned} \|r_{t+1}\| &\leq \left(1+\eta_0(1-\|r_t\|^2)\right)\|r_t\|+10\eta_0\delta(\|E_t\|+\|r_t\|)+2\eta_0\|E_t\|^2\|r_t\|. \\ &\overset{(a)}{\leq} \left(1+20\eta_0\delta+\eta_0(1-\|r_t\|^2)\right)\|r_t\|+2\eta_0\|r_t\|^2\cdot\|r_t\| \\ &\overset{(b)}{\leq} \left(1+20\eta_0\delta+\eta_0+\frac{1}{2}\eta_0\right)\|r_t\| \\ &\leq (1+2\eta_0)\|r_t\|. \end{aligned} \tag{94}$$

Here we used Lemma 9 and $\|r_t\| \leq \frac{1}{2}$ in (a) and (b), respectively. Therefore,

$$\|r_T\| \leq \alpha(1+2\eta_0)^T. \tag{95}$$

This shows that we need at least $T_3 = \Omega\left(\log\frac{1}{\alpha}/\eta_0\right) = \Omega\left(\log\frac{r'}{\delta}/\eta_0\right)$ iterations to guarantee $\|r_t\| \geq 1/2$. Now, suppose $\|r_{T_1+T_2}\| > 1$. Without loss of generality, we assume that $\|r_{T_1+T_2-1}\| \leq 1 < \|r_{T_1+T_2}\|$ (since T_3 and T_1+T_2 have the same order). Under this assumption, we show that $\|r_{T_1+T_2}\| \leq 1+O\left(\delta+\delta^2\log^2\frac{r'}{\delta}\right)$. By Proposition 3, we have

$$\begin{aligned} \|r_{t+1}\|-\|r_t\| &\leq \eta_0(1-\|r_t\|^2)\|r_t\|+10\eta_0\delta(\|E_t\|+\|r_t\|)+2\eta_0\|E_t\|^2\|r_t\| \\ &\leq 6\eta_0(1-\|r_t\|)+40\eta_0\delta+4\eta_0\delta^2\log^2\frac{r'}{\delta}. \end{aligned} \tag{96}$$

where we used the Lemma 9 and $\|r_t\| \leq 2$. Then, by our choice of $\|r_{T_1+T_2-1}\|$ and $\|r_{T_1+T_2}\|$, we have

$$\begin{aligned}\|r_{T_1+T_2}\| - \|r_{T_1+T_2-1}\| &\leq 6\eta_0(1 - \|r_{T_1+T_2-1}\|) + 40\eta_0\delta + 4\eta_0\delta^2 \log^2 \frac{r'}{\delta} \\ &\leq 6\eta_0 \left(\|r_{T_1+T_2}\| - \|r_{T_1+T_2-1}\|\right) + 40\eta_0\delta + 4\eta_0\delta^2 \log^2 \frac{r'}{\delta}.\end{aligned} \tag{97}$$

Then, since $\eta_0 \lesssim 1$, we have

$$\|r_{T_1+T_2}\| - 1 \leq \|r_{T_1+T_2}\| - \|r_{T_1+T_2-1}\| \lesssim \delta \vee \delta^2 \log^2 \frac{r'}{\delta}. \tag{98}$$

This implies that $|1 - \|r_{T_1+T_2}\|\,| \lesssim \delta \vee \delta^2 \log^2 \frac{r'}{\delta}$.

Finally, these two stages characterize the behavior of r_t and its convergence to the true solution. In particular, with the choice of $T = T_1 + T_2 = O(\log \frac{r'}{\delta}/\eta_0)$, and according to Lemma 1, we have

$$\begin{aligned}\left\|U_T U_T^\top - X^\star\right\|_F^2 &\leq (1 - \|r_T\|^2)^2 + 2\|E_T\|^2 \|r_T\|^2 + \|E_T\|_F^4 \\ &\lesssim \delta^2 + \delta^4 \log^4 \frac{r'}{\delta} + \delta^2 \log^2 \frac{r'}{\delta} + r'^2\alpha^4 + \delta^4 \log^4 \frac{r'}{\delta} \\ &\lesssim \delta^2 \log^2 \frac{r'}{\delta}\end{aligned} \tag{99}$$

which completes the proof. □

D Proofs for the Noisy Case

For simplicity of notation, we denote $\varphi_t = \varphi(\Delta_t)$, where $\Delta_t = U_t U_t^\top - X^*$.

D.1 Proof of Proposition 7

Analogous to the proof of Proposition 3, first we provide a general result which holds for arbitrary learning rates.

Lemma 10. *Suppose that $\|E_t\|_F \leq 1, \|r_t\| \leq 2$, $\delta \leq \frac{1}{2}$, then, the following inequalities hold*

$$\|E_{t+1}\|_F^2 \leq \|E_t\|_F^2 + 2\eta_t\varphi_t \left(-\frac{\left\|E_t U_t^\top\right\|_F^2}{\|\Delta_t\|_F} + \delta \left\|E_t U_t^\top\right\|_F\right) + 10\eta_t^2\varphi_t^2 \left(\delta^2 + \frac{\left\|E_t U_t^\top\right\|_F^2}{\|\Delta_t\|_F^2}\right), \tag{100}$$

$$\|E_{t+1}\| \leq \left\|I - \frac{\eta_t\varphi_t U_t^\top U_t}{\|\Delta_t\|_F}\right\| \cdot \|E_t\| + \delta\eta_t\varphi_t(\|r_t\| + \|E_t\|). \tag{101}$$

Proof. The proof is similar to that of Lemma 6. The details are omitted for brevity. □

Now, we are ready to present the proof of Proposition 7.

Proof of Proposition 7. Based on the sign-RIP condition, the step sizes satisfy

$$\eta_t = \frac{\eta_0\rho^t}{\|D\|_F} \leq \frac{\eta_0\rho^t}{\varphi_t(1-\delta)} \leq \frac{2\eta_0\rho^t}{\varphi_t}, \tag{102}$$

where $D \in \mathcal{M}(U_t U_t - X^*)$. For the Frobenius norm, we have

$$\begin{aligned}\|E_{t+1}\|_F^2 &\le \|E_t\|_F^2 + 2\eta_t\varphi_t\left(-\frac{\left\|E_tU_t^\top\right\|_F^2}{\|\Delta_t\|_F} + \delta\left\|E_tU_t^\top\right\|_F\right) + 10\eta_t^2\varphi_t^2\left(\delta^2 + \frac{\left\|E_tU_t^\top\right\|_F^2}{\|\Delta_t\|_F^2}\right)\\ &\le \|E_t\|_F^2 + 2\eta_t\varphi_t\delta\left\|E_tU_t^\top\right\|_F + 10\delta^2\eta_t^2\varphi_t^2\\ &\le \|E_t\|_F^2 + 4\delta\eta_0\rho^t\left\|E_tU_t^\top\right\|_F + 20\delta^2\eta_0^2\rho^{2t}.\end{aligned} \tag{103}$$

where in the second inequality, we used the assumption $\eta_0 \lesssim \delta \lesssim \|\Delta_t\|_F$, which implies

$$-2\eta_t\varphi_t\frac{\left\|E_tU_t^\top\right\|_F^2}{\|\Delta_t\|_F} + 10\eta_t^2\varphi_t^2\frac{\left\|E_tU_t^\top\right\|_F^2}{\|\Delta_t\|_F^2} \le 0 \tag{104}$$

Furthermore, note that

$$\begin{aligned}\left\|E_tU_t^\top\right\|_F^2 &= \left\|E_tE_t^\top\right\|_F^2 + \left\|E_tr_tu^{\star\top}\right\|_F^2\\ &\le \|E_t\|_F^4 + \|E_t\|_F^2\|r_t\|^2\\ &\le (1+4)\|E_t\|_F^2,\end{aligned} \tag{105}$$

which implies

$$\begin{aligned}\|E_{t+1}\|_F^2 &\le \|E_t\|_F^2 + 20\delta\eta_0\rho^t\|E_t\|_F + 20\delta^2\eta_0^2\rho^{2t}\\ &\le \left(\|E_t\|_F + 10\delta\eta_0\rho^t\right)^2.\end{aligned} \tag{106}$$

This leads to $\|E_{t+1}\|_F \le \|E_t\|_F + 10\delta\eta_0\rho^t$.

For the spectral norm, since we suppose $\eta_0 \lesssim \delta \lesssim \|\Delta_t\|_F$, we have $\left\|I - \frac{\eta_t\varphi_tU_t^\top U_t}{\|\Delta_t\|_F}\right\| \le 1$. Combined with Lemma 10, this implies that

$$\|E_{t+1}\| \le \|E_t\| + \delta\eta_t\varphi_t(\|r_t\| + \|E_t\|) \le \|E_t\| + 2\delta\eta_0\rho^t(\|r_t\| + \|E_t\|). \tag{107}$$

thereby completing the proof. □

D.2 Proof of Proposition 6

Similar to the proof of Proposition 4, we first present a general result which holds for arbitrary learning rates.

Lemma 11. *For any learning rate η_t, if $\|E_t\|_F \le 1, \|r_t\| \le 2$, then we have*

$$\begin{aligned}&\left\|r_{t+1} - \left(1 + \frac{\varphi_t\eta_t(1-\|r_t\|^2)}{\left\|U_tU_t^\top - X^\star\right\|_F}\right)r_t\right\|\\ \le&\delta\eta_t\varphi_t(\|E_t\| + \|r_t\|) + \frac{\varphi_t\eta_t}{\left\|U_tU_t^\top - X^\star\right\|_F}\|E_t\|^2\|r_t\|.\end{aligned} \tag{108}$$

Proof. The proof is similar to that of Lemma 8. The details are omitted for brevity. □

Proof of Proposition 6. Assuming that $\delta \le \frac{1}{2}$, we have

$$\left|\eta_t - \frac{\eta_0}{\varphi_t}\rho^t\right| = \left|\frac{\eta_0}{\|D\|_F}\rho^t - \frac{\eta_0}{\varphi_t}\rho^t\right| \le \frac{\delta\varphi_t\eta_0\rho^t}{(1-\delta)\varphi_t^2} \le \frac{2\delta\eta_0\rho^t}{\varphi_t}. \tag{109}$$

Combined with Lemma 11, this implies that

$$\begin{aligned}\left\|r_{t+1} - \left(1 + \frac{\eta_0\rho^t(1-\|r_t\|^2)}{\left\|U_tU_t^\top - X^\star\right\|_F}\right)r_t\right\| &\le 2\delta\eta_0\rho^t(\|E_t\| + \|r_t\|) + \frac{2\eta_0\rho^t}{\left\|U_tU_t^\top - X^\star\right\|_F}\|E_t\|^2\|r_t\| \\ &\quad + \frac{2\delta\eta_0\rho^t}{\left\|U_tU_t^\top - X^\star\right\|_F}(1-\|r_t\|^2)\|r_t\|.\end{aligned} \tag{110}$$

which completes the proof. □

D.3 Proof of Theorem 3

Recall that $T = \Theta(\log\frac{1}{\alpha}/\eta_0)$. First, we show that the error term remains small during the iterations of SubGD. Due to the choice of initial point, we have $\|E_0\| = \|r_0\| = \alpha$, $\|E_0\|_F = \sqrt{r'}\alpha$. Hence, Proposition 7 leads to

$$\|E_t\| \le \|E_t\|_F = \|E_0\|_F + \sum_{t=1}^{t}(\|E_t\|_F - \|E_{t-1}\|_F) \le \sqrt{r'}\alpha + 10\delta\eta_0 t \le \sqrt{r'}\alpha + 10\delta\eta_0 T \lesssim \delta\log\frac{r'}{\delta}. \tag{111}$$

provided that $\|\Delta_t\|_F \ge 1 - \|r_t\|^2 \gtrsim \delta$. To verify this assumption, we show that $\|\Delta_t\|_F \ge 1 - \|r_t\|^2 \gtrsim \delta\log\frac{r'}{\delta}$ for every $t \le \bar{T}$, where $\bar{T} \gtrsim \log\frac{1}{\alpha}/\eta_0$. To this goal, first we present a preliminary claim

Claim 1. *For every $0 \le t \le T$, we have $\|E_t\| \le \|r_t\|$.*

Proof. It follows an argument analogous to the proof of Lemma 9. The details are omitted for brevity. □

Based on this claim, we are ready to show that $\|\Delta_t\|_F \ge 1 - \|r_t\|^2 \gtrsim \delta\log\frac{r'}{\delta}$ for every $t \le \bar{T}$, where $\bar{T} \gtrsim \log\frac{1}{\alpha}/\eta_0$.

Claim 2. *Suppose that $\delta \le 1/6$. Then, for every $0 \le t \lesssim \log\frac{1}{\alpha}/\eta_0$, we have $\|r_t\| \le \frac{1}{2}$.*

Proof. The statement holds for $t = 0$ since $\|r_0\| = \alpha$. Now, suppose that $\|r_t\| \le \frac{1}{2}$. Then, we have

$$\begin{aligned}\|r_{t+1}\| &\le \left(1 + \frac{4}{3}\frac{\eta_0\rho^t}{\|\Delta_t\|_F}(1-\|r_t\|^2)\right)\|r_t\| + 2\delta\eta_0\rho^t(\|E_t\| + \|r_t\|) + \frac{2\eta_0\rho^t}{\|\Delta_t\|_F}\|E_t\|^2\|r_t\| \\ &\overset{(a)}{\le} \left(1 + O(1)\eta_0\rho^t\right)\|r_t\| + 2\delta\eta_0\rho^t(\|E_t\| + \|r_t\|) \\ &\overset{(b)}{\le} \left(1 + O(1)\eta_0\rho^t\right)\|r_t\|.\end{aligned} \tag{112}$$

where in (a) we use the fact that $\|\Delta_t\|_F \ge 1 - \|r_t\|^2 \ge 3/4$ and $\|E_t\| \lesssim 1$; and in (b) we use Claim 1. Without loss of generality, we assume that $\|r_{t+1}\| \le \left(1 + \eta_0\rho^t\right)\|r_t\|$. Hence, it suffices to show that

$$\|r_0\|\prod_{s=1}^{t}(1 + \eta_0\rho^s) = \alpha\prod_{s=1}^{t}(1 + \eta_0\rho^s) \le \frac{1}{2}, \tag{113}$$

for every $0 \leq t \lesssim \log \frac{1}{\alpha}/\eta_0$. This is equivalent to

$$\sum_{s=1}^{t} \log\left(1+\eta_0\rho^s\right) \leq \log \frac{1}{2\alpha}. \tag{114}$$

On the other hand, note that

$$\sum_{s=1}^{t} \log\left(1+\eta_0\rho^s\right) \leq \sum_{s=1}^{t} \eta_0\rho^s \leq \eta_0 \frac{1-\rho^t}{1-\rho} \leq C \log\frac{1}{\alpha}\left(1-\rho^t\right). \tag{115}$$

Therefore, to finish the proof, we need to show that $C\log\frac{1}{\alpha}\left(1-\rho^t\right) \leq \log\frac{1}{2\alpha}$, which implies $1-\frac{1}{C}+\frac{\log 2}{C\log\frac{1}{\alpha}} \leq \rho^T$. This can be easily verified for every $t \lesssim \log\frac{1}{\alpha}/\eta_0$, by noting that $\rho = 1-\Theta(\eta_0/\log\frac{1}{\alpha})$. □

Based on the above claim and upon choosing $\bar{T} \asymp \log\frac{1}{\alpha}/\eta_0$, the error term is bounded as (111) for every $t \leq \bar{T}$. Now, note that the proof is completed if $\|\Delta_t\|_F \lesssim \delta \log\frac{r'}{\delta}$ for some $\bar{T} \leq t \leq T$. Therefore, suppose that $\|\Delta_t\|_F \gtrsim \delta\log\frac{r'}{\delta}$ for every $\bar{T} \leq t \leq T$. This implies that the error bound (111) holds for every $\bar{T} \leq t \leq T$. Moreover, we assume that $1-\|r_t\|^2 \geq 3\|E_t\|_F$, since otherwise, we have $1-\|r_t\|^2 \lesssim \delta\log(r'/\delta)$, and the proof is completed together with $\|E_t\|_t \leq \delta\log(r'/\delta)$ and Lemma 1. This leads to

$$1-\|r_t\|^2 \leq \|\Delta_t\|_F \leq 1-\|r_t\|^2 + \|E_t\|\|r_t\| + \|E_t\|_F^2 \leq \frac{13}{9}(1-\|r_t\|^2). \tag{116}$$

assuming that $\|r_t\| \leq 1$. Then, according to Proposition 6, we have

$$\begin{aligned}\|r_{t+1}\| &\geq \left(1+\frac{2}{3}\frac{\eta_0\rho^t}{\|\Delta_t\|_F}(1-\|r_t\|^2)\right)\|r_t\| - 2\delta\eta_0\rho^t(\|E_t\|+\|r_t\|) - \frac{2\eta_0\rho^t}{\|\Delta_t\|_F}\|E_t\|^2\|r_t\| \\ &\overset{(a)}{\geq} \left(1+\Omega(1)\eta_0\rho^t\right)\|r_t\| - 2\delta\eta_0\rho^t\|E_t\|.\end{aligned} \tag{117}$$

where in (a) we used $\|E_t\|^2 \leq (1-\|r_t\|^2)/9$, inequality (116), and $\delta \lesssim 1$. To proceed, note that $\|E_t\| \leq \|r_t\|$ due to Claim 1. Hence, we have

$$\|r_{t+1}\| \geq \left(1+\Omega(1)\eta_0\rho^t\right)\|r_t\|. \tag{118}$$

for every $0 \leq t \leq T$. Now, it remains to show that after $T = O\left(\log\left(\frac{1}{\alpha}\right)/\eta_0\right)$ iterations, the signal term approaches 1. Without loss of generality, we assume that $\|r_{t+1}\| \geq (1+\eta_0\rho^t)\|r_t\|$, which implies $\|r_T\| \geq \alpha\prod_{t=1}^{T}\left(1+\eta_0\rho^t\right)$. Taking the logarithm of the right hand side leads to

$$\sum_{t=1}^{T}\log\left(1+\eta_0\rho^t\right) \geq \sum_{t=1}^{T}\frac{\eta_0\rho^t}{1+\eta_0\rho^t} \geq \frac{\eta_0}{2}\frac{1-\rho^T}{1-\rho}. \tag{119}$$

where we used the lower bound $\log(1+x) \geq \frac{x}{1+x}$ for $x \geq -1$. Now, upon defining $\gamma = 1-\rho$, we have

$$\begin{aligned}\frac{\eta_0}{2}\frac{1-\rho^T}{1-\rho} &= \frac{\eta_0}{2}\frac{1-(1-\gamma)^T}{\gamma} \\ &\geq \frac{\eta_0}{2\gamma}\left(1-\left(1-\frac{\gamma T}{1+(T-1)\gamma}\right)\right) \\ &\geq \frac{\eta_0}{2\gamma}\frac{\gamma T}{2}.\end{aligned} \tag{120}$$

where we used the basic inequality $(1-x)^r \le 1 - \frac{rx}{1+(r-1)x}$ for $x \in [0,1], r > 1$. Now, recalling $T = \Theta\left(\log \frac{1}{\alpha}/\eta_0\right)$ and $\gamma = \Theta\left(\eta_0/\log \frac{1}{\alpha}\right)$, we have $\frac{\eta_0}{2\gamma}\frac{\gamma T}{2} \ge \log(1/\alpha)$, which implies that after $T = \Theta(\log \frac{1}{\alpha}/\eta_0)$ iterations, the signal term satisfies $\|r_T\| \ge 1$. So, the only remaining part is to show that $\|r_T\| = 1 \pm O(\delta \log \frac{r'}{\delta})$. Recall that, based on the definition of $\bar{T}$, we have $\|r_{\bar{T}}\| < 1$. Now, we assume that $\|r_{T-1}\| < 1$, and $\|r_T\| \ge 1$. Note that this assumption is without loss of generality, since $\bar{T}$ and T have the same order. Then we have the following claim.

Claim 3. *Either* $1 - \delta \log \frac{r'}{\delta} \lesssim \|r_{T-1}\|^2$, *or* $\|r_T\| \lesssim 1 + \delta^2 \log \frac{r'}{\delta}$.

Proof. Assume that $\|\Delta_{T-1}\|_F \ge 1 - \|r_{T-1}\|^2 \gtrsim \delta \log \frac{r'}{\delta}$. Then, by Proposition 6, we have

$$\begin{aligned}
\|r_T\| - \|r_{T-1}\| &\le \frac{4}{3}\frac{\eta_0\rho^{T-1}(1-\|r_{T-1}\|^2)}{\|\Delta_{T-1}\|_F}\|r_{T-1}\| + \frac{2\eta_0\rho^{T-1}\|E_{T-1}\|^2}{\|\Delta_{T-1}\|_F}\|r_{T-1}\| + O(\delta\eta_0\rho^T) \\
&\lesssim \frac{1}{\log\frac{r'}{\delta}}(1-\|r_{T-1}\|) + \delta^2\log\frac{r'}{\delta} \\
&\lesssim \frac{1}{\log\frac{r'}{\delta}}(\|r_T\| - \|r_{T-1}\|) + \delta^2\log\frac{r'}{\delta}
\end{aligned} \tag{121}$$

This implies that, for sufficiently small δ, we have $\|r_T\| - \|r_{T-1}\| = O(\delta^2 \log \frac{r'}{\delta})$, thereby completing the proof. □

In summary, we showed that $1 - \delta \log \frac{r'}{\delta} \lesssim \|r_{T-1}\|^2 \le 1$, or $1 \le \|r_T\| \lesssim 1 + \delta^2 \log \frac{r'}{\delta}$. On the other hand, we know that $\|E_t\| \lesssim \delta \log \frac{r'}{\delta}$ for every $t \le T$. This together with Lemma 1 completes the proof. □

E Proof for GD

E.1 Proof of Proposition 8

We divide our analysis into two cases. In the first case, we assume $p\sigma^2 = \Omega(1)$. We have

$$\begin{aligned}
\sup_{X\in\mathbb{S}} \|Q(X) - X\|_F &= \sup_{X,Y\in\mathbb{S}} \left| \frac{1}{m}\sum_{i=1}^m \langle A_i, X\rangle \langle A_i, Y\rangle + \frac{1}{m}\sum_{i\in S} s_i \langle A_i, Y\rangle - \langle X, Y\rangle \right| \\
&\overset{(a)}{\ge} \sup_{Y\in\mathbb{S}} \left| \frac{1}{m}\sum_{i=1}^m \langle A_i, Y\rangle^2 + \frac{1}{m}\sum_{i\in S} s_i \langle A_i, Y\rangle - 1 \right| \\
&\ge \sup_{Y\in\mathbb{S}} \left| \frac{1}{m}\sum_{i\in S} s_i \langle A_i, Y\rangle \right| - \sup_{Y\in\mathbb{S}} \left| \frac{1}{m}\sum_{i=1}^m \langle A_i, Y\rangle^2 - 1 \right| \\
&\overset{(b)}{=} \left\| \frac{1}{m}\sum_{i\in S} s_i A_i \right\|_F - \sup_{Y\in\mathbb{S}} \left| \frac{1}{m}\sum_{i=1}^m \langle A_i, Y\rangle^2 - 1 \right|.
\end{aligned} \tag{122}$$

where in (a) we add a constraint $X = Y$ to the supremum; and in (b) we use the Cauchy-Schwartz inequality and the variational form of the Frobenius norm. By the ℓ_2-RIP for Gaussian measurements

(Lemma 14), we have

$$\sup_{X,Y\in\mathbb{S}}\left|\frac{1}{m}\sum_{i=1}^{m}\langle A_i,X\rangle\langle A_i,Y\rangle+\frac{1}{m}\sum_{i\in S}s_i\langle A_i,Y\rangle-\langle X,Y\rangle\right|\geq\left\|\frac{1}{m}\sum_{i\in S}s_iA_i\right\|_F-\delta_1 \tag{123}$$

with probability of at least $1-Ce^{-cm\delta_1^2}$, given $m\gtrsim d^2$. The expectation and tail bound of $\left\|\frac{1}{m}\sum_{i\in S}s_iA_i\right\|_F$ is provided in the following lemma.

Lemma 12. *For any $0<t<1$, we have*

$$\mathbb{P}\left(\left|\left\|\frac{1}{m}\sum_{i\in S}s_iA_i\right\|_F-\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i\in S}s_iA_i\right\|_F\right]\right|\geq t\right)\leq 2e^{-\frac{Cmt^2}{p\sigma^2d^2}}, \tag{124}$$

where C is a universal constant. Moreover, the expectation is lower bounded as

$$\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i\in S}s_iA_i\right\|_F\right]\gtrsim\sqrt{\frac{p\sigma^2d^2}{m}}. \tag{125}$$

Before providing the proof of Lemma 12, we complete the proof of Proposition 8. Based on the above lemma and (123), we have

$$\sup_{X,Y\in\mathbb{S}}\left|\frac{1}{m}\sum_{i=1}^{m}\langle A_i,X\rangle\langle A_i,Y\rangle+\frac{1}{m}\sum_{i\in S}s_i\langle A_i,Y\rangle-1\right|\geq C\sqrt{\frac{p\sigma^2d^2}{m}}-\delta_1-\delta_2, \tag{126}$$

with probability of at least $1-Ce^{-c_1m\delta_1^2}-e^{-c_2\frac{m\delta_2^2}{p\sigma^2d^2}}$. Hence, with the proper choice of δ_1,δ_2, we have

$$\mathbb{P}\left(\sup_{X\in\mathbb{S}}\|Q(X)-X\|_F\geq C\sqrt{\frac{p\sigma^2d^2}{m}}\right)\geq\frac{1}{2}. \tag{127}$$

Since $p\sigma^2=\Omega(1)$, we can choose C' such that

$$\mathbb{P}\left(\sup_{X\in\mathbb{S}}\|Q(X)-X\|_F\geq C'\sqrt{\frac{(1+p\sigma^2)d^2}{m}}\right)\geq\frac{1}{2}. \tag{128}$$

In the second case, we assume that $p\sigma^2=O(1)$. Making a similar argument, we can show that there exists a universal constant C such that

$$\mathbb{P}\left(\sup_{X\in\mathbb{S}}\|Q(X)-X\|_F\geq C\sqrt{\frac{d^2}{m}}\right)\geq\frac{1}{2}. \tag{129}$$

Combining the two cases, the following inequality holds for an arbitrary $\sigma>0$

$$\mathbb{P}\left(\sup_{X\in\mathbb{S}}\left\|M_2(X)-\bar{M}_2(X)\right\|_F\geq C'\sqrt{\frac{(1+p\sigma^2)d^2}{m}}\right)\geq\frac{1}{2}. \tag{130}$$

Which completes the proof of Proposition 8. □

Now, we present the proof for Lemma 12.

Proof of Lemma 12. For simplicity, we denote $B = \frac{1}{m}\sum_{i\in S} s_i A_i$. First, we prove the lower bound on the expectation. Note that, conditioned on s_i, we have $B_{j,k} = \frac{1}{m}\sum_{i\in S} s_i A^i_{j,k} \sim N\left(0, \frac{1}{m^2}\sum_{i\in S} s_i^2\right)$. Then, by invoking Theorem 3.1.1. in [Vershynin, 2019], we have

$$\begin{aligned}
\mathbb{E}\left[\|B\|_F\right] &= \mathbb{E}\left[\mathbb{E}\left[\|B\|_F\right] | s_i, i \in S\right] \\
&\gtrsim \mathbb{E}\left[\frac{d}{m}\sqrt{\sum_{i\in S} s_i^2}\right] \\
&\gtrsim \frac{\sigma d}{m}\sqrt{pm} = \sqrt{\frac{p\sigma^2 d^2}{m}}.
\end{aligned} \tag{131}$$

Now, we show that $\|B\|_F$ is a sub-exponential random variable. First, for arbitrary indices i, j, k, the random variable $s_i A^i_{j,k}$ is sub-exponential according to Lemma 17 since $\left\|s_i A^i_{j,k}\right\|_{\psi_1} \le \|s_i\|_{\psi_2}\left\|A^i_{j,k}\right\|_{\psi_2} = \Theta(\sigma)$. This implies that $\|B_{j,k}\|_{\psi_1} = \Theta\left(\sqrt{\frac{p\sigma^2}{m}}\right)$. Finally, we have

$$\begin{aligned}
\left\|\|B\|_F\right\|_{\ell^{2k}} &= \left(\left\|\sum_{j,k} B^2_{j,k}\right\|_{\ell^k}\right)^{1/2} \\
&\overset{(a)}{\le} \left(\sum_{j,k}\left\|B^2_{j,k}\right\|_{\ell^k}\right)^{1/2} \\
&= d\left\|B_{j,k}\right\|_{\ell^{2k}} \lesssim \sqrt{\frac{p\sigma^2 d^2}{m}}k.
\end{aligned} \tag{132}$$

which implies that $\|B\|_F$ is sub-exponential with sub-exponential norm $O\left(\sqrt{\frac{p\sigma^2 d^2}{m}}\right)$ due to the equivalent definition of sub-exponential random variable (see Definition 3). Note that in (a) we used the Minkowski inequality. Given the lower bound on the expected value, the tail bound directly follows from the tail of sub-exponential distribution. □

F Auxiliary Lemmas

F.1 Restricted Isometry Property

Lemma 13. *Let $\mathbb{S}_r = \{X \in \mathbb{R}^{d\times d} : \operatorname{rank}(X) \le r, \|X\|_F = 1\}$. Then, there exists an ϵ-covering $\mathbb{S}_{\epsilon,r}$ with respect to the Frobenius norm satisfying $|\mathbb{S}_\epsilon| \le \left(\frac{9}{\epsilon}\right)^{(2d+1)r}$.*

Lemma 14 (ℓ_2-RIP, Theorem 4.2 in [Recht et al., 2010])**.** *Fix $0 < \delta < 1$, suppose that the measurement matrices $\{A_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries. Then, we have*

$$\sup_{X\in\mathbb{S}_r}\left|\frac{1}{m}\sum_{i=1}^m \langle A_i, X\rangle^2 - \|X\|_F^2\right| \le \delta. \tag{133}$$

with probability of at least $1 - Ce^{c_1 dr\log\frac{1}{\delta} - c_2 m\delta^2}$.

F.2 Basic Probability

Lemma 15 (Conditional Gaussian Variable in Bivariate Case)**.** *For two Gaussian random variables $X \sim \mathcal{N}(\mu_1, \sigma_1^2), Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ with correlation coefficient ρ, we have*

$$X|Y = a \sim \mathcal{N}\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho\left(a - \mu_2\right), \left(1 - \rho^2\right)\sigma_1^2\right). \tag{134}$$

Definition 2 (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 $\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}$. This definition is equivalent to the following statements*

- *(Tail bound) For any $t > 0$, we have $\mathbb{P}(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$.*
- *(Moment bound) For any positive integer p, we have $\|X\|_{\ell^p} = \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \lesssim \sigma\sqrt{p}$.*

Moreover, the sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} := \sup_{p\geq 1}\left\{p^{-1/2}\|X\|_{\ell^p}\right\}$.

For sum of independent sub-Gaussian random variables, their sub-Gaussian norm can be bounded via the following lemma.

Lemma 16 (Proposition 2.6.1 in [Vershynin, 2019])**.** *Let $X_1, \cdots, X_m$ be a series independent zero-mean sub-Gaussian variables, then $\sum_{i=1}^m$ is sub-Gaussian and*

$$\left\|\sum_{i=1}^m X_i\right\|_{\psi_2}^2 \lesssim \sum_{i=1}^m \|X_i\|_{\psi_2}^2. \tag{135}$$

Definition 3 (Sub-exponential random variable)**.** *A random variable X with expectation μ is sub-exponential if there exists (μ, α), such that $\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2\nu^2}{2}}$ for all $|\lambda| \leq \alpha$. This definition is equivalent to the following statements:*

- *(tail bound) There exists a universal constant C, for any $t > 0$, we have $\mathbb{P}(|X-\mu| \geq t) \leq 2e^{-Ct}$.*
- *(moment bound) For any positive integer p, we have $\|X\|_{\ell^p} = \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \lesssim p$.*

Moreover, the sub-exponential norm of X is defined as $\|X\|_{\psi_1} := \sup_{p\geq 1}\left\{p^{-1}\|X\|_{L^p}\right\}$.

For sub-Gaussian and sub-exponential random variables, we have the following lemma to illustrate their relations.

Lemma 17. *The following statements hold*

- *(Lemma 2.7.6 in [Vershynin, 2019]) A random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover, $\|X\|_{\psi_2}^2 = \left\|X^2\right\|_{\psi_1}$.*
- *(Lemma 2.7.7 in [Vershynin, 2019]) Let X and Y be sub-Gaussian random variables. Then XY is sub-exponential. Moreover, $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$.*

F.3 Basic Inequalities

Lemma 18 (Bernoulli inequality)**.**

$$(1+x)^r \leq 1 + \frac{rx}{1-(r-1)x}, \quad \text{for } x \in \left[-1, \frac{1}{r-1}\right), r \geq 1. \tag{136}$$