NONCONVEX ROBUST LOW-RANK MATRIX RECOVERY*

XIAO LI[†], ZHIHUI ZHU [‡], ANTHONY MAN-CHO SO [§], AND RENÉ VIDAL[‡]

Abstract. In this paper we study the problem of recovering a low-rank matrix from a number of random linear measurements that are corrupted by outliers taking arbitrary values. We consider a nonsmooth nonconvex formulation of the problem, in which we explicitly enforce the low-rank property of the solution by using a factored representation of the matrix variable and employ an ℓ_1-loss function to robustify the solution against outliers. We show that even when a constant fraction (which can be up to almost half) of the measurements are arbitrarily corrupted, as long as certain measurement operators arising from the measurement model satisfy the so-called ℓ_1/ℓ_2-restricted isometry property, the ground-truth matrix can be exactly recovered from any global minimum of the resulting optimization problem. Furthermore, we show that the objective function of the optimization problem is sharp and weakly convex. Consequently, a subgradient Method (SubGM) with geometrically diminishing step sizes will converge linearly to the ground-truth matrix when suitably initialized. We demonstrate the efficacy of the SubGM for the nonconvex robust low-rank matrix recovery problem with various numerical experiments.

Key words. robust low-rank matrix recovery, sharpness, weak convexity, subgradient method, robust PCA

AMS subject classifications. 65K10, 90C26, 68Q25, 68W40, 62B10.

1. Introduction.

Low-rank matrices are ubiquitous in computer vision [8, 23], machine learning [40], and signal processing [13] applications. One fundamental computational task is to recover a low-rank matrix $\boldsymbol{X}^\star \in \mathbb{R}^{n_1 \times n_2}$ from a small number of linear measurements

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}^\star), \tag{1.1}$$

where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is a known linear operator. Such a task arises in quantum tomography [1], face recognition [8], linear system identification [18], collaborative filtering [10], etc. We refer the interested reader to [13,53] for more detailed discussions.

Although in many interesting scenarios the number of linear measurements m is much smaller than $n_1 n_2$, the low-rank property of $\boldsymbol{X}^\star$ suggests that its degrees of freedom can also be much smaller than $n_1 n_2$, thus making the task of recovering $\boldsymbol{X}^\star$ possible. This has been demonstrated in, e.g., [10], where a nuclear norm minimization appproach for recovering a low-rank matrix from random linear measurements is studied. Despite the strong theoretical guarantees of such approach (see also [21]), most existing methods for solving the nuclear norm minimization problem do not scale well with the problem size (i.e., n_1, n_2, and m). To overcome this computational bottleneck, one approach is to enforce the low-rank property explicitly by using a factored representation of the matrix variable in the optimization formulation. Such

---

*Submitted to the editors December 12, 2019. The first and second authors contributed equally to this paper.

Funding: X. Li was partially supported by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) Project CUHK 14210617. Z. Zhu and R. Vidal were partially supported by NSF Grant 1704458. A. M.-C. So was partially supported by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) Project CUHK 14208117.

[†]Department of Electronic Engineering, The Chinese University of Hong Kong. (xli@ee.cuhk.edu.hk, http://www.ee.cuhk.edu.hk/~xli/).

[‡]Center for Imaging Science, Mathematical Institute for Data Science, Johns Hopkins University. (zzhu29@jhu.edu, http://cis.jhu.edu/~zhihui/; rvidal@jhu.edu, http://cis.jhu.edu/~rvidal/).

[§]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong. (manchoso@se.cuhk.edu.hk, http://www.se.cuhk.edu.hk/~manchoso/).

an approach has already been explored in some early works on low-rank semidefinite programming (see, e.g., [5,6] and the references therein) but has gained renewed interest lately in the study of low-rank matrix recovery problems. For the purpose of illustration, let us first consider the case where the ground-truth matrix $\boldsymbol{X}^\star$ is symmetric positive semidefinite with rank r. Instead of optimizing, say, an ℓ_2-loss function involving an $n \times n$ symmetric positive semidefinite matrix variable $\boldsymbol{X}$ with either a constraint or a regularization term controlling the rank of $\boldsymbol{X}$, we consider the factorization $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}$ and optimize the loss function over the $n \times r$ matrix variable $\boldsymbol{U}$:

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r}}{\operatorname{minimize}} \left\{ \xi(\boldsymbol{U}) := \frac{1}{m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})\|_2^2 \right\}. \tag{1.2}$$

There are two obvious advantages with the formulation (1.2). First, the recovered matrix will automatically satisfy the rank and positive semidefinite constraints. Second, when the rank of the ground-truth matrix is small, the size of the variable $\boldsymbol{U}$ can be much smaller than that of $\boldsymbol{X}$. Although the quadratic nature of $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}$ renders the objective function ξ in (1.2) nonconvex, recent advances in the analysis of the landscapes of structured nonconvex functions allow one to show that when the linear measurement operator $\mathcal{A}$ satisfies certain restricted isometry property (RIP), local search algorithms (such as gradient descent) are guaranteed to find a global minimum of (1.2) and exactly recover the underlying low-rank matrix $\boldsymbol{X}^\star$ [4,19,35,41,52]. Moreover, it was shown in [42,50] that (1.2) satisfies an error bound condition, indicating that simple gradient descent with an appropriate initialization will converge to a global minimum at a linear rate; see [12] for a comprehensive review.

1.1. Our Goal and Main Results. In this paper, we consider the *robust low-rank matrix recovery problem*, in which the measurements are corrupted by *outliers*. Specifically, we assume that

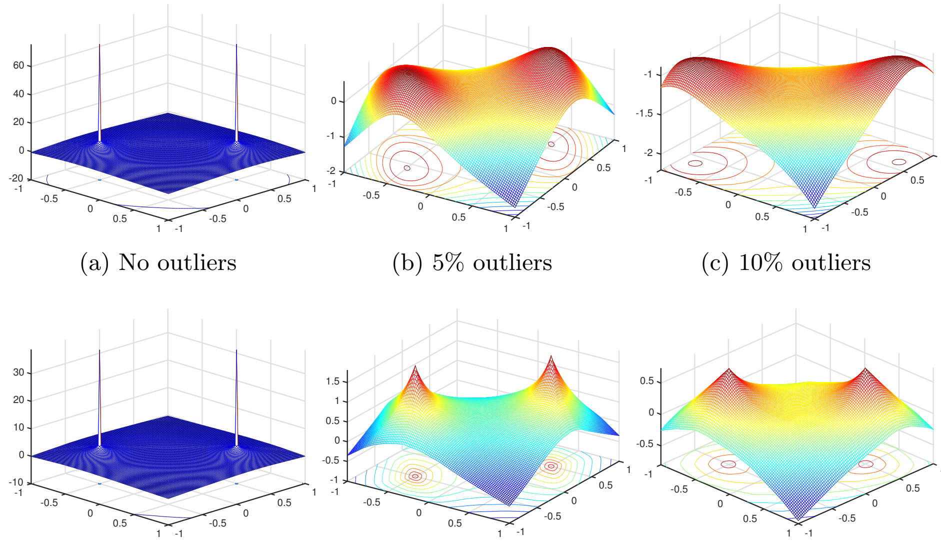
$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}^\star) + \boldsymbol{s}^\star, \tag{1.3}$$

where $\boldsymbol{s}^\star \in \mathbb{R}^m$ is an outlier vector such that a small fraction of its entries (the outliers) have an arbitrary magnitude and the remaining entries are zero. Moreover, the set of nonzero entries is assumed to be unknown. Outliers are prevalent in the context of sensor calibration [31] (because of sensor failure), face recognition [16] (due to self-shadowing, specularity, or saturations in brightness), video surveillance [26] (where the foreground objects are modeled as outliers), etc.

It is well known that the ℓ_2-loss function is sensitive to outliers, thus rendering (1.2) ineffective for recovering the underlying low-rank matrix. As illustrated in the top row of Figure 1, the global minima of ξ in (1.2) are perturbed away from the underlying low-rank matrix because of the outliers, and a larger fraction of outliers leads to a larger perturbation. By contrast, the ℓ_1-loss function is more robust against outliers and has been widely utilized for outlier detection [8,24,31]. This motivates us to adopt the ℓ_1-loss function together with the factored representation of the matrix variable to tackle the robust low-rank matrix recovery problem:

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r}}{\operatorname{minimize}} \left\{ f(\boldsymbol{U}) := \frac{1}{m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})\|_1 \right\}. \tag{1.4}$$

The robustness of the ℓ_1-loss function against outliers can be seen from the bottom row of Figure 1, where the global minima of (1.4) correspond precisely to the underlying



(a) No outliers (b) 5% outliers (c) 10% outliers

(d) No outliers (e) 5% outliers (f) 10% outliers

Fig. 1: Landscapes of the objective functions $\boldsymbol{U} \mapsto \xi(\boldsymbol{U}) = \frac{1}{m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})\|_2^2$ (top row) and $\boldsymbol{U} \mapsto f(\boldsymbol{U}) = \frac{1}{m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})\|_1$ (bottom row) for low-rank matrix recovery with different percentages of outliers in the measurement vector $\boldsymbol{y}$ (1.3). Here, the ground-truth matrix $\boldsymbol{X}^\star$ is given by $\boldsymbol{X}^\star = \boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}}$ with $\boldsymbol{U}^\star = [0.5 \quad 0.5]^{\mathrm{T}}$ and 40 measurements are taken to form $\boldsymbol{y}$. For display purpose, we plot $-\log(\xi(\boldsymbol{U}))$ and $-\log(f(\boldsymbol{U}))$ instead of $\xi(\boldsymbol{U})$ and $f(\boldsymbol{U})$.

low-rank matrix $\boldsymbol{X}^\star$ even in the presence of outliers. However, compared with (1.2), the exact recovery property of (1.4) (i.e., when the global minima of (1.4) yield the ground-truth matrix $\boldsymbol{X}^\star$) and the convergence behavior of local search algorithms for solving (1.4) are much less understood. This stems in part from the fact that (1.4) is a nonsmooth nonconvex optimization problem, but most of the algorithmic and analysis techniques developed in the recent literature on structured nonconvex optimization problems apply only to the smooth setting.

In view of the above discussion, we aim to (i) provide conditions in terms of the number of linear measurements m and the fraction of outliers that can guarantee the exact recovery property of (1.4) and (ii) design a first-order method to solve (1.4) and establish guarantees on its convergence performance. To achieve (i), we utilize the notion of ℓ_1/ℓ_2-restricted isometry property (ℓ_1/ℓ_2-RIP), which has been introduced previously in the context of low-rank matrix recovery [46, 48] and covariance estimation [11]. We show that if the fraction of outliers is slightly less than $\frac{1}{2}$, then as long as the measurement operator $\mathcal{A}$ and its restriction $\mathcal{A}_{\Omega^c}$ onto the complement of the support set Ω of the outlier vector $\boldsymbol{s}^\star$ possess the ℓ_1/ℓ_2-RIP, any global minimum $\boldsymbol{U}^\star$ of (1.4) must satisfy $\boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}} = \boldsymbol{X}^\star$. To tackle (ii), we propose to use a subgradient method (SubGM) to solve (1.4). As a key step in the convergence analysis of the SubGM, we show that under the aforementioned setting for the fraction of outliers and the ℓ_1/ℓ_2-RIP of the operators $\mathcal{A}$ and $\mathcal{A}_{\Omega^c}$, the objective function f in (1.4) is *sharp* (see Definition 1) and *weakly convex* (see Definition 2). Consequently, we can apply (a slight variant of) the analysis framework in [14] to show that when initialized close to the set of global minima of (1.4), the SubGM with geometrically diminishing step sizes will converge R-linearly to a global minimum. To the best of our knowledge, this is the first time an exact recovery condition (i.e., the ℓ_1/ℓ_2-RIP of $\mathcal{A}$ and $\mathcal{A}_{\Omega^c}$) for the optimization formulation (1.4) is shown to also imply its regularity (i.e., sharpness and weak convexity). We summarize the above results in the following theorem:

THEOREM 1 (informal; see Theorem 3 for the formal statement). *Consider the measurement model* (1.3)*, where the ground-truth matrix* $\boldsymbol{X}^\star$ *is symmetric positive semidefinite with rank* r*. Suppose that the fraction of outliers is less than half and both operators* $\mathcal{A}$ *and* $\mathcal{A}_{\Omega^c}$ *possess the* ℓ_1/ℓ_2*-RIP (see Subsection 3.1 and Subsection 3.2). Then, every global minimum of* (1.4) *corresponds to the ground-truth matrix* $\boldsymbol{X}^\star$ *and the objective function* f *is sharp (see Definition 1) and weakly convex (see Definition 2). Consequently, when applied to* (1.4)*, the SubGM with an appropriate initialization will converge to the ground-truth matrix* $\boldsymbol{X}^\star$ *at a linear rate.*

Before we proceed, several remarks are in order. First, for various random measurement operators $\mathcal{A}$, such as sub-Gaussian measurement operators and the quadratic measurement operators in [11], as long as the number of measurements is sufficiently large, the operators $\mathcal{A}$ and $\mathcal{A}_{\Omega^c}$ will possess the ℓ_1/ℓ_2-RIP with high probability. This is the case, for instance, when $\mathcal{A}$ is a Gaussian measurement operator with $m \gtrsim nr$ measurements.[1] In particular, when combined with Theorem 1, we see that the low-rank matrix $\boldsymbol{X}^\star$ in (1.3) can be recovered using an information-theoretically optimal number of measurements. Second, although at first glance (1.4) seems to be more difficult to solve than (1.2) because of nonsmoothness, Theorem 1 implies that (1.4) can be solved *as efficiently as* its smooth counterpart (1.2), in the sense that both can be solved by first-order methods that have a linear convergence guarantee.

Although Theorem 1 is concerned with the setting where $\boldsymbol{X}^\star$ is symmetric positive semidefinite, it can be extended to the general setting where $\boldsymbol{X}^\star$ is a rank-r $n_1 \times n_2$ matrix. Specifically, by using the factorization $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ with $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{V} \in \mathbb{R}^{n_2 \times r}$ and utilizing the nonsmooth regularizer $\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\|_F$ (or $\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\|_1$) to account for the ambiguities in the factorization caused by invertible transformations, we formulate the general robust low-rank matrix recovery problem as follows:

$$\text{(1.5)} \qquad \underset{\boldsymbol{U}\in\mathbb{R}^{n_1\times r},\boldsymbol{V}\in\mathbb{R}^{n_2\times r}}{\text{minimize}} \left\{ g(\boldsymbol{U},\boldsymbol{V}) := \frac{1}{m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})\|_1 + \lambda\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\|_F \right\}.$$

Here, $\lambda > 0$ is a regularization parameter. We remark that the regularizer used in the above formulation is motivated by but different from that used in [35,42,52]. The latter, which is given by $\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\|_F^2$, is smooth but is not as well suited for robustifying the solution against outliers. In Section 4 we show that all the results established for (1.4) in Theorem 1 carry over to (1.5) for any $\lambda > 0$ (but the choice of λ affects the sharpness and weak convexity parameters; see the discussion after Proposition 6).

1.2. Related Work. By analyzing the optimization geometry, recent works [4, 19, 28, 35, 42] have shown that many local search algorithms with either an appropriate initialization or a random initialization can provably solve the low-rank matrix recovery problem (1.2) when the measurement operator $\mathcal{A}$ satisfies the RIP. In particular, gradient descent with an appropriate initialization is shown to converge to a global optimum at a linear rate [42, 51], while quadratic convergence is established for the cubic regularization method [47]. Key to these results is certain error bound conditions, which elucidate the regularity properties of the underlying optimization problem. Recently, the above results have been extended to cover general

[1]See Subsection 1.3 for the meaning of the notation $\gtrsim$.

smooth low-rank matrix optimization problems whose objective functions satisfy the restricted strong convexity and smoothness properties [27, 51, 52].

For the robust low-rank matrix recovery problem, existing solution methods can be classified into two categories. The first is based on the convex approach [8, 25, 31]. Although such approach enjoys strong statistical guarantees, it is computational expensive and thus not scalable to practical problems. The second category is based on the nonconvex approach. This includes the alternating minimization methods [22, 33, 45, 49], which typically use projected gradient descent for low-rank matrix recovery and thresholding-based truncation for identification of outliers. However, these methods typically require performing an SVD in each iteration for projection onto the set of low-rank matrices. Recently, a median-truncated gradient descent method has been proposed in [30] to tackle (1.2), where the gradient is modified to alleviate the effect of outliers. The median-truncated gradient descent is shown to have a local linear convergence rate [30], but such guarantee requires $m \gtrsim nr \log n$ measurements. Moreover, the maximum number of outliers that can be tolerated is not explicitly given. By contrast, our result only requires $m \gtrsim nr$ measurements (which matches the optimal information-theoretic bound) and explicitly bounds the fraction of outliers that can be present. We also note that a SubGM has been proposed in [31] for solving (1.4) in the setting where $\mathcal{A}$ is a certain quadratic measurement operator. As reported in [31], the SubGM exhibits excellent empirical performance in terms of both computational efficiency and accuracy. In this paper, we provide a rigorous justification for the empirical success of the SubGM, thus answering a question that is left open in [31].

Finally, we remark that our work is closely related to the recent works [2, 14, 15, 54] on subgradient methods for nonsmooth nonconvex optimization. A projected subgradient method is proven to converge linearly for the robust subspace recovery problem [54] and sublinearly for orthonormal dictionary learning [2]. It is shown in [14, 15] that if the optimization problem at hand is sharp (see Definition 1) and weakly convex (see Definition 2), various subgradient methods for solving it will converge at a linear rate. Currently, only a few applications are known to give rise to sharp and weakly convex optimization problems, such as robust phase retrieval [15, 17] and robust covariance estimation with quadratic sampling [14]. Thus, our result expands the repertoire of optimization problems that are sharp and weakly convex and contributes to the growing literature on the geometry of structured nonsmooth nonconvex optimization problems.

1.3. Notation.

Let us introduce the notations used in this paper. Finite-dimensional vectors and matrices are indicated by bold characters. The symbols $\boldsymbol{I}$ and $\boldsymbol{0}$ represent the identity matrix and zero matrix/vector, respectively. The set of $r \times r$ orthogonal matrices is denoted by $\mathcal{O}_r := \{\boldsymbol{R} \in \mathbb{R}^{r\times r} : \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} = \boldsymbol{I}\}$. The subdifferential of the absolute value function $|\cdot|$ is denoted by Sign; i.e., $\mathrm{Sign}(a) := \begin{cases} a/|a|, & a \neq 0, \\ [-1,1], & a = 0. \end{cases}$ We use $\mathrm{Sign}(\boldsymbol{A})$ to denote the matrix obtained by applying the Sign function to each element of the matrix $\boldsymbol{A}$. Furthermore, we use $\|\boldsymbol{A}\|_F$ to denote the Frobenius norm of the matrix $\boldsymbol{A}$ and $\|\boldsymbol{a}\|$ to denote the ℓ_2-norm of the vector $\boldsymbol{a}$. Finally, we use $x \lesssim y$ (resp. $x \gtrsim y$) to indicate that $x \leq cy$ (resp. $x \geq cy$) for some universal constant $c > 0$.

2. Problem Setup and Preliminaries.

Consider the general optimization problem

$$\inf_{\boldsymbol{x}\in\mathbb{R}^n} h(\boldsymbol{x}), \tag{2.1}$$

where $h : \mathbb{R}^n \to \mathbb{R}$ is a lower semi-continuous, possibly nonsmooth and nonconvex, function. Let $h^\star$ denote the optimal value of (2.1) and

$$\mathcal{X} := \{\boldsymbol{z} \in \mathbb{R}^n : h(\boldsymbol{z}) \le h(\boldsymbol{x}),\ \forall \boldsymbol{x} \in \mathbb{R}^n\}$$

denote the set of global minima of h. We assume that $\mathcal{X} \neq \emptyset$. Given any $\boldsymbol{x} \in \mathbb{R}^n$, the distance between $\boldsymbol{x}$ and $\mathcal{X}$ is defined as

$$\operatorname{dist}(\boldsymbol{x}, \mathcal{X}) := \inf_{\boldsymbol{z} \in \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{z}\|.$$

Since h can be nonsmooth, we utilize tools from generalized differentiation to formulate the optimality condition of (2.1). The (Fréchet) subdifferential of h at $\boldsymbol{x}$ is defined as

$$\partial h(\boldsymbol{x}) := \left\{ \boldsymbol{d} \in \mathbb{R}^n : \liminf_{\boldsymbol{y} \to \boldsymbol{x}} \frac{h(\boldsymbol{y}) - h(\boldsymbol{x}) - \langle \boldsymbol{d}, \boldsymbol{y} - \boldsymbol{x} \rangle}{\|\boldsymbol{y} - \boldsymbol{x}\|} \ge 0 \right\}, \tag{2.2}$$

where each $\boldsymbol{d} \in \partial h(\boldsymbol{x})$ is called a subgradient of h at $\boldsymbol{x}$. We say that $\boldsymbol{x}$ is a critical point of h if $\boldsymbol{0} \in \partial h(\boldsymbol{x})$.

2.1. Sharpness and Weak Convexity.

Since our goal is to consider a set of problems that can be solved by the SubGM with a linear rate of convergence, let us introduce two regularity notions for h that are central to our study.

DEFINITION 1 (sharpness; cf. [7]). *We say that $h : \mathbb{R}^n \to \mathbb{R}$ is sharp with parameter $\alpha > 0$ if*

$$h(\boldsymbol{x}) - h^\star \ge \alpha \operatorname{dist}(\boldsymbol{x}, \mathcal{X})$$

for all $\boldsymbol{x} \in \mathbb{R}^n$.

DEFINITION 2 (weak convexity; see, e.g., [44]). *We say that $h : \mathbb{R}^n \to \mathbb{R}$ is weakly convex with parameter $\tau \ge 0$ if $\boldsymbol{x} \mapsto h(\boldsymbol{x}) + \frac{\tau}{2}\|\boldsymbol{x}\|^2$ is convex.*

Suppose that h is sharp and weakly convex with parameters $\alpha > 0$ and $\tau \ge 0$, respectively. It is known that for any $\boldsymbol{x} \notin \mathcal{X}$ with $\operatorname{dist}(\boldsymbol{x}, \mathcal{X}) < \frac{2\alpha}{\tau}$, we have $\boldsymbol{0} \notin \partial h(\boldsymbol{x})$; i.e., $\boldsymbol{x}$ is not a critical point of h [14, Lemma 3.1]. This suggests the possibility of finding a global minimum of h by initializing local search algorithms with a point that is close to $\mathcal{X}$. To explore such possibility, let us consider using the SubGM in Algorithm 2.1 to solve the nonsmooth nonconvex optimization problem (2.1).

Algorithm 2.1 Subgradient Method (SubGM) for Solving (2.1)

Initialization: set $\boldsymbol{x}_0$ and μ_0;
1: **for** $k = 0, 1, \ldots$ **do**
2: compute a subgradient $\boldsymbol{d}_k \in \partial h(\boldsymbol{x}_k)$;
3: update the step size μ_k according to a certain rule;
4: update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \mu_k \boldsymbol{d}_k$;
5: **end for**

2.2. Convergence of SubGM for Sharp Weakly Convex Functions.

Unlike gradient descent, the SubGM with a constant step size may not converge to a critical point of a nonsmooth function in general, even when the function is convex [3,32,38]. To ensure the convergence of the SubGM, a set of diminishing step sizes is generally needed [20, 38]. As it turns out, for a sharp weakly convex function h, the SubGM with step sizes that are diminishing at a geometric rate can still be shown to converge linearly to a global minimum when initialized close to $\mathcal{X}$. Specifically, let

$$\kappa := \sup \left\{ \|\boldsymbol{d}\| : \boldsymbol{d} \in \partial h(\boldsymbol{x}), \operatorname{dist}(\boldsymbol{x}, \mathcal{X}) < \frac{2\alpha}{\tau} \right\}, \tag{2.3}$$

which can be shown to satisfy $\kappa \geq \alpha$; cf. [14, Lemma 3.2]. Then, we have the following result:

THEOREM 2 (local linear convergence of SubGM). *Suppose that the function $h : \mathbb{R}^n \to \mathbb{R}$ is sharp and weakly convex with parameters $\alpha > 0$ and $\tau \geq 0$, respectively. Suppose further that the SubGM in Algorithm 2.1 is initialized with a point $\boldsymbol{x}_0$ satisfying $\operatorname{dist}(\boldsymbol{x}_0, \mathcal{X}) < \frac{2\alpha}{\tau}$ and uses the geometrically diminishing step sizes $\mu_k = \rho^k \mu_0$, where the initial step size μ_0 satisfies*

$$\mu_0 \leq \frac{\alpha^2}{2\tau\kappa^2}\left(1 - \left(\max\left\{\frac{\tau}{\alpha}\operatorname{dist}(\boldsymbol{x}_0, \mathcal{X}) - 1, 0\right\}\right)^2\right) \tag{2.4}$$

and the decay rate ρ satisfies

$$1 > \rho \geq \underline{\rho} := \sqrt{1 - \left(\frac{2\alpha}{\overline{\operatorname{dist}}_0} - \tau\right)\mu_0 + \frac{\kappa^2}{\overline{\operatorname{dist}}_0^2}\mu_0^2} \tag{2.5}$$

with

$$\overline{\operatorname{dist}}_0 = \max\left\{\operatorname{dist}(\boldsymbol{x}_0, \mathcal{X}), \mu_0 \frac{\max\{\kappa^2, 2\alpha^2\}}{\alpha}\right\}. \tag{2.6}$$

Then, the iterates $\{\boldsymbol{x}_k\}_{k\geq 0}$ generated by the SubGM will converge linearly to a point in $\mathcal{X}$:

$$\operatorname{dist}(\boldsymbol{x}_k, \mathcal{X}) \leq \rho^k \overline{\operatorname{dist}}_0, \ \forall k \geq 0.$$

We note that a similar result has been established in [14, Corollary 6.1]. Nevertheless, compared with [14, Corollary 6.1], which requires $\frac{\alpha}{\kappa} \leq \sqrt{\frac{1}{2-\gamma}}$ and $\operatorname{dist}(\boldsymbol{x}_0, \mathcal{X}) \leq \frac{\gamma\alpha}{\tau}$ for some $\gamma \in (0, 1)$, Theorem 2 is less restrictive and allows the larger initialization region $\operatorname{dist}(\boldsymbol{x}_0, \mathcal{X}) < \frac{2\alpha}{\tau}$. In particular, as $\frac{\alpha}{\kappa}$ tends to 1, so does γ, and the decay rate ρ in [14, Corollary 6.1] approaches 1. Thus, one can no longer use [14, Corollary 6.1] to conclude that the SubGM converges linearly when $\frac{\alpha}{\kappa} = 1$. By contrast, the linear convergence result in Theorem 2 is still valid in this case. Theorem 2 can be proven by refining the arguments in the proof of [14, Theorem 6.1]. We refer the reader to the companion technical report [29] of this paper for details.

Before we proceed, it is worth elaborating on the implication of Theorem 2 when h is convex. In this case, we can take $\tau = 0$, which, in view of (2.4), shows that μ_0 can be arbitrarily chosen. If we choose $\mu_0 \geq \frac{\alpha \operatorname{dist}(\boldsymbol{x}_0, \mathcal{X})}{\max\{\kappa^2, 2\alpha^2\}}$, then by (2.6) we have $\overline{\operatorname{dist}}_0 = \mu_0 \frac{\max\{\kappa^2, 2\alpha^2\}}{\alpha}$, which implies that the decay rate $\underline{\rho}$ satisfies

$$\underline{\rho} = \sqrt{1 - \frac{2\alpha^2}{\max\{\kappa^2, 2\alpha^2\}} + \frac{\kappa^2\alpha^2}{(\max\{\kappa^2, 2\alpha^2\})^2}} = \begin{cases} \sqrt{1 - \frac{\alpha^2}{\kappa^2}}, & \kappa^2 \geq 2\alpha^2, \\ \frac{\kappa}{2\alpha}, & \kappa^2 < 2\alpha^2. \end{cases}$$

In particular, this is in line with the results in [20, Theorem 4.4].

3. Nonconvex Robust Low-Rank Matrix Recovery: Symmetric Positive Semidefinite (PSD) Case.

In the last section we saw that the SubGM with suitable initialization and step sizes converges linearly to a global minimum of a sharp weakly convex function. Naturally, it is of interest to identify concrete problems that possess these two regularity properties. In this section we focus on the robust low-rank matrix recovery problem (1.4) and establish, for the first time, a connection between the exact recovery condition of ℓ_1/ℓ_2-RIP and the regularity properties of sharpness and weak convexity of the objective function f in (1.4). Specifically, we first show that if the fraction of outliers is slightly less than $\frac{1}{2}$ and certain measurement operators arising from the measurement model (1.3) possess the ℓ_1/ℓ_2-RIP, then the sharpness condition in Definition 1 holds for (1.4). Consequently, all global minima of (1.4) lead to the exact recovery of the ground-truth matrix $\boldsymbol{X}^\star$. We then show that (1.4) also satisfies the weak convexity condition in Definition 2. Hence, by the convergence result (Theorem 2) in the last section, we conclude that the SubGM can be utilized to find a global minimum of (1.4) efficiently.

To begin, let us collect some preparatory results. Let $\boldsymbol{X}^\star = \boldsymbol{U}^\star \boldsymbol{U}^{\star\mathrm{T}}$ be a factorization of $\boldsymbol{X}^\star$, where $\boldsymbol{U}^\star \in \mathbb{R}^{n\times r}$. Note that for any $\boldsymbol{R} \in \mathcal{O}_r$, we have $\boldsymbol{X}^\star = \boldsymbol{U}^\star \boldsymbol{R}(\boldsymbol{U}^\star \boldsymbol{R})^\mathrm{T}$. Thus, all elements in the set

$$\mathcal{U} := \{\boldsymbol{U}^\star \boldsymbol{R} : \boldsymbol{R} \in \mathcal{O}_r\}$$

are valid factors of $\boldsymbol{X}^\star$. Furthermore, it is clear that the function f in (1.4) is constant on the set $\mathcal{U}$. The following result connects $\operatorname{dist}(\boldsymbol{U},\mathcal{U})$ and the distance between $\boldsymbol{U}\boldsymbol{U}^\mathrm{T}$ and $\boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}}$ for any given $\boldsymbol{U} \in \mathbb{R}^{n\times r}$:

Lemma 1 ([42, Lemma 5.4]). *Given any $\boldsymbol{U}^\star \in \mathbb{R}^{n\times r}$, define $\boldsymbol{X}^\star = \boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}}$. Then, for any $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, we have*

$$2\left(\sqrt{2}-1\right)\sigma_r^2(\boldsymbol{X}^\star)\operatorname{dist}^2(\boldsymbol{U},\mathcal{U}) \le \|\boldsymbol{U}\boldsymbol{U}^\mathrm{T} - \boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}}\|_F^2,$$

where σ_r denotes the r-th largest singular value.

3.1. ℓ_1/ℓ_2-Restricted Isometry Property. Since the ℓ_1/ℓ_2-RIP [11, 46, 48] of the linear measurement operator $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ in (1.4) plays an important role in our subsequent analysis, let us first provide a condition under which $\mathcal{A}$ will possess such property. Recall that $\mathcal{A}$ can be specified by a collection of m $n\times n$ matrices $\boldsymbol{A}_1,\ldots,\boldsymbol{A}_m$. In other words, given any $\boldsymbol{X} \in \mathbb{R}^{n\times n}$, we have $\mathcal{A}(\boldsymbol{X}) = (\langle \boldsymbol{A}_1, \boldsymbol{X}\rangle, \ldots, \langle \boldsymbol{A}_m, \boldsymbol{X}\rangle)$. We now show that if $\boldsymbol{A}_1,\ldots,\boldsymbol{A}_m$ have independent and identically distributed (*i.i.d.*) standard Gaussian entries, then $\mathcal{A}$ will possess the ℓ_1/ℓ_2-RIP with high probability.

Proposition 1 (ℓ_1/ℓ_2-RIP of Gaussian measurement operators). *Let $r \ge 1$ be given. Suppose that $m \gtrsim nr$ and the matrices $\boldsymbol{A}_1,\ldots,\boldsymbol{A}_m \in \mathbb{R}^{n\times n}$ defining the linear measurement operator $\mathcal{A}$ have* i.i.d. *standard Gaussian entries. Then, for any $0 < \delta < \sqrt{\frac{2}{\pi}}$, there exists a universal constant $c > 0$ such that with probability exceeding $1-\exp(-c\delta^2 m)$, $\mathcal{A}$ will possess the ℓ_1/ℓ_2-RIP; i.e., the inequalities*

$$\left(\sqrt{\frac{2}{\pi}}-\delta\right)\|\boldsymbol{X}\|_F \le \frac{1}{m}\|\mathcal{A}(\boldsymbol{X})\|_1 \le \left(\sqrt{\frac{2}{\pi}}+\delta\right)\|\boldsymbol{X}\|_F \tag{3.1}$$

hold for any rank-$2r$ matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$.

The proof of Proposition 1 is given in Appendix A. It is worth noting that similar ℓ_1/ℓ_2-RIPs hold for other types of measurement operators such as the quadratic measurement operators in [11] and those defined by sub-Gaussian matrices. Thus, although our results are stated for Gaussian measurement operators, they can be readily extended to cover other measurement operators that possess similar RIPs.

3.2. Sharpness and Exact Recovery. Assuming that the linear measurement operator $\mathcal{A}$ possesses the ℓ_1/ℓ_2-RIP (3.1), our first goal is to identify further conditions on the measurement model (1.3) so that any global minimum $\boldsymbol{U}^\star$ of (1.4) can be used to recover the ground-truth matrix $\boldsymbol{X}^\star$ via $\boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}} = \boldsymbol{X}^\star$. Towards that end, let $\Omega \subseteq \{1, \dots, m\}$ denote the support of the outlier vector $\boldsymbol{s}^\star$ and $\Omega^c = \{1, \dots, m\} \setminus \Omega$. Furthermore, let $p = \frac{|\Omega|}{m}$ be the fraction of outliers in $\boldsymbol{y}$. Throughout, we do not make any assumption on the location of the non-zero entries of $\boldsymbol{s}^\star$. Instead, we assume that $\mathcal{A}_{\Omega^c}$, the linear operator defined by the matrices in $\{\boldsymbol{A}_i : i \in \Omega^c\}$, also possesses the ℓ_1/ℓ_2-RIP; i.e., we have

$$\left(\sqrt{\frac{2}{\pi}} - \delta\right) \|\boldsymbol{X}\|_F \le \frac{1}{m(1-p)} \|[\mathcal{A}(\boldsymbol{X})]_{\Omega^c}\|_1 \le \left(\sqrt{\frac{2}{\pi}} + \delta\right) \|\boldsymbol{X}\|_F \tag{3.2}$$

for any rank-$2r$ matrix $\boldsymbol{X}$. When each $\boldsymbol{A}_i$ is generated with *i.i.d.* standard Gaussian entries, Proposition 1 implies that $\mathcal{A}_{\Omega^c}$ will satisfy (3.2) with high probability as long as p is a constant. This follows from the fact that $|\Omega^c| = (1-p)m \gtrsim nr$ if $m \gtrsim nr$.

PROPOSITION 2 (sharpness and exact recovery with outliers: PSD case). *Let* $0 < \delta < \frac{1}{3}\sqrt{\frac{2}{\pi}}$ *be given. Suppose that the fraction of outliers* p *satisfies*

$$p < \frac{1}{2} - \frac{\delta}{\sqrt{2/\pi - \delta}}, \tag{3.3}$$

and that the linear operators $\mathcal{A}$ *and* $\mathcal{A}_{\Omega^c}$ *possess the* ℓ_1/ℓ_2*-RIP* (3.1) *and* (3.2)*, respectively. Then, the objective function* f *in* (1.4) *satisfies*

$$f(\boldsymbol{U}) - f(\boldsymbol{U}^\star) \ge \alpha \operatorname{dist}(\boldsymbol{U}, \mathcal{U})$$

for any $\boldsymbol{U} \in \mathbb{R}^{n \times r}$*, where*

$$\alpha = \sqrt{2\left(\sqrt{2} - 1\right)} \left(2(1-p)\left(\sqrt{\frac{2}{\pi}} - \delta\right) - \left(\sqrt{\frac{2}{\pi}} + \delta\right)\right) \sigma_r(\boldsymbol{X}^\star) > 0. \tag{3.4}$$

In particular, the set $\mathcal{U}$ *is precisely the set of global minima of* (1.4) *and the objective function* f *is sharp with parameter* $\alpha > 0$.

Proof of Proposition 2. Using (1.3) and (1.4), we compute

$$\begin{aligned}
f(\boldsymbol{U}) - f(\boldsymbol{U}^\star) &= \frac{1}{m} \left\|\mathcal{A}\left(\boldsymbol{U}\boldsymbol{U}^\mathrm{T} - \boldsymbol{X}^\star\right) - \boldsymbol{s}^\star\right\|_1 - \frac{1}{m}\|\boldsymbol{s}^\star\|_1 \\
&= \frac{1}{m}\left\|\left[\mathcal{A}\left(\boldsymbol{U}\boldsymbol{U}^\mathrm{T} - \boldsymbol{X}^\star\right)\right]_{\Omega^c}\right\|_1 + \frac{1}{m}\left\|\left[\mathcal{A}\left(\boldsymbol{U}\boldsymbol{U}^\mathrm{T} - \boldsymbol{X}^\star\right)\right]_{\Omega} - \boldsymbol{s}^\star\right\|_1 - \frac{1}{m}\|\boldsymbol{s}^\star\|_1 \\
&\ge \frac{1}{m}\left\|\left[\mathcal{A}\left(\boldsymbol{U}\boldsymbol{U}^\mathrm{T} - \boldsymbol{X}^\star\right)\right]_{\Omega^c}\right\|_1 - \frac{1}{m}\left\|\left[\mathcal{A}\left(\boldsymbol{U}\boldsymbol{U}^\mathrm{T} - \boldsymbol{X}^\star\right)\right]_{\Omega}\right\|_1 \\
&= \frac{2}{m}\left\|\left[\mathcal{A}\left(\boldsymbol{U}\boldsymbol{U}^\mathrm{T} - \boldsymbol{X}^\star\right)\right]_{\Omega^c}\right\|_1 - \frac{1}{m}\left\|\mathcal{A}\left(\boldsymbol{U}\boldsymbol{U}^\mathrm{T} - \boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}}\right)\right\|_1 \\
&\ge \left(2(1-p)\left(\sqrt{\frac{2}{\pi}} - \delta\right) - \left(\sqrt{\frac{2}{\pi}} + \delta\right)\right)\left\|\boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}} - \boldsymbol{U}\boldsymbol{U}^\mathrm{T}\right\|_F \\
&\ge \alpha \operatorname{dist}(\boldsymbol{U}, \mathcal{U}),
\end{aligned}$$

where the second inequality follows from the ℓ_1/ℓ_2-RIP of $\mathcal{A}$ and $\mathcal{A}_{\Omega^c}$ and the last inequality follows from Lemma 1. The characterization of the set of global minima of (1.4) follows immediately from the above inequality and the choice of p in (3.3). ∎

One interesting consequence of Proposition 2 is that for the robust low-rank matrix recovery problem (1.4), the sharpness condition (which characterizes the geometry of the optimization problem around the set of global minima) coincides with the exact recovery property (which is of statistical nature). Moreover, condition (3.3) suggests that the smaller δ is, the higher the outlier ratio p can be. On the other hand, given an outlier ratio p, condition (3.3) requires that $\delta < \sqrt{\frac{2}{\pi}} - \frac{\sqrt{2/\pi}}{3/2-p}$, which indirectly imposes a condition on the number of measurements m. Indeed, Proposition 1 implies that in order for a Gaussian measurement operator $\mathcal{A}$ to possess the ℓ_1/ℓ_2-RIP with positive probability, we need $m \gtrsim nr/\big(\sqrt{\frac{2}{\pi}} - \frac{\sqrt{2/\pi}}{3/2-p}\big)^2$ measurements. Putting it another way, the larger the number of measurements m is, the higher the outlier ratio p can be. We shall elaborate on this point with experiments in Section 5.

3.3. Weak Convexity. In the last subsection we established the sharpness of (1.4) and showed that any of its global minimum will lead to the exact recovery of the ground-truth matrix $\boldsymbol{X}^\star$, even when the fraction of outliers is up to almost $\frac{1}{2}$. In this subsection we further establish the weak convexity of (1.4), thus opening up the possibility of using the machinery developed in Section 2 to obtain provable convergence guarantees for the SubGM when it is applied to solve (1.4). Towards that end, we note that the ℓ_1-norm, being a convex function, is subdifferentially regular [37, Example 7.27] (see [37, Definition 7.25] for the definition of subdifferential regularity). Hence, by the chain rule for subdifferentials of subdifferentially regular functions [37, Corollary 8.11 and Theorem 10.6], we have

(3.5)
$$\partial f(\boldsymbol{U}) = \frac{1}{m}\left[\left(\mathcal{A}^*\left(\operatorname{Sign}\left(\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}) - \boldsymbol{y}\right)\right)\right)^{\mathrm{T}}\boldsymbol{U} + \mathcal{A}^*\left(\operatorname{Sign}\left(\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}) - \boldsymbol{y}\right)\right)\boldsymbol{U}\right].$$

We are now ready to prove the following result. Note that the weak convexity parameter τ in (3.6) is independent of the fraction of outliers.

PROPOSITION 3 (weak convexity: PSD case). *Suppose that the measurement operator $\mathcal{A}$ satisfies the ℓ_1/ℓ_2-RIP (3.1). Then, the objective function f in (1.4) is weakly convex with parameter*

(3.6)
$$\tau = 2\Big(\sqrt{\frac{2}{\pi}} + \delta\Big).$$

Proof of Proposition 3. For any $\boldsymbol{U}', \boldsymbol{U} \in \mathbb{R}^{n\times r}$, let $\boldsymbol{\Delta} = \boldsymbol{U}' - \boldsymbol{U}$. Then, we have

$$\begin{aligned}
f(\boldsymbol{U}') &= \frac{1}{m}\left\|\mathcal{A}(\boldsymbol{U}'\boldsymbol{U}'^{\mathrm{T}} - \boldsymbol{X}^\star) - \boldsymbol{s}^\star\right\|_1 \\
&= \frac{1}{m}\left\|\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} - \boldsymbol{X}^\star + \boldsymbol{U}\boldsymbol{\Delta}^{\mathrm{T}} + \boldsymbol{\Delta}\boldsymbol{U}^{\mathrm{T}} + \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}) - \boldsymbol{s}^\star\right\|_1 \\
&\geq \frac{1}{m}\left\|\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} - \boldsymbol{X}^\star + \boldsymbol{U}\boldsymbol{\Delta}^{\mathrm{T}} + \boldsymbol{\Delta}\boldsymbol{U}^{\mathrm{T}}) - \boldsymbol{s}^\star\right\|_1 - \frac{1}{m}\left\|\mathcal{A}(\boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}})\right\|_1 \\
&\geq \frac{1}{m}\left\|\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} - \boldsymbol{X}^\star + \boldsymbol{U}\boldsymbol{\Delta}^{\mathrm{T}} + \boldsymbol{\Delta}\boldsymbol{U}^{\mathrm{T}}) - \boldsymbol{s}^\star\right\|_1 - \left(\sqrt{\frac{2}{\pi}} + \delta\right)\left\|\boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\right\|_F \\
&\geq f(\boldsymbol{U}) + \frac{1}{m}\left\langle \boldsymbol{d}, \mathcal{A}(\boldsymbol{U}\boldsymbol{\Delta}^{\mathrm{T}} + \boldsymbol{\Delta}\boldsymbol{U}^{\mathrm{T}})\right\rangle - \frac{\tau}{2}\|\boldsymbol{\Delta}\|_F^2
\end{aligned}$$

for any $\boldsymbol{d} \in \operatorname{Sign}(\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}) - \boldsymbol{y})$, where the second inequality follows from the ℓ_1/ℓ_2-RIP of $\mathcal{A}$ and the last inequality is due to the convexity of the ℓ_1-norm and $\|\boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\|_F \leq$

$\|\boldsymbol{\Delta}\|_F^2$. Substituting (3.5) into the above equation gives

$$f(\boldsymbol{U}') \geq f(\boldsymbol{U}) + \langle \boldsymbol{D}, \boldsymbol{U}' - \boldsymbol{U} \rangle - \frac{\tau}{2}\|\boldsymbol{U}' - \boldsymbol{U}\|_F^2, \ \forall \ \boldsymbol{D} \in \partial f(\boldsymbol{U}).$$

This completes the proof. □

3.4. Putting Everything Together. With the results in Subsection 3.2 and Subsection 3.3 in place, in order to show that the SubGM enjoys the convergence guarantees in Theorem 2 when applied to the robust low-rank matrix recovery problem (1.4), it remains to determine κ, the bound on the norm of any subgradient of f in a neighborhood of $\mathcal{U}$; see (2.3). This is established by the following result:

PROPOSITION 4 (bound on subgradient norm: PSD case). *Suppose that the measurement operator $\mathcal{A}$ satisfies the ℓ_1/ℓ_2-RIP (3.1). Then, for any $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ satisfying $\operatorname{dist}(\boldsymbol{U}, \mathcal{U}) \leq \frac{2\alpha}{\tau}$, we have*

$$\|\boldsymbol{D}\|_F \leq \kappa = 2\left(\sqrt{\frac{2}{\pi}} + \delta\right)\left(\|\boldsymbol{U}^\star\|_F + \frac{2\alpha}{\tau}\right), \ \forall \boldsymbol{D} \in \partial f(\boldsymbol{U}). \tag{3.7}$$

Proof of Proposition 4. Recall from (2.2) that

$$\liminf_{\boldsymbol{U}' \to \boldsymbol{U}} \frac{f(\boldsymbol{U}') - f(\boldsymbol{U}) - \langle \boldsymbol{D}, \boldsymbol{U}' - \boldsymbol{U} \rangle}{\|\boldsymbol{U}' - \boldsymbol{U}\|_F} \geq 0 \tag{3.8}$$

for any $\boldsymbol{D} \in \partial f(\boldsymbol{U})$. Now, for any $\boldsymbol{U}' \in \mathbb{R}^{n\times r}$,

$$\begin{aligned}
|f(\boldsymbol{U}') - f(\boldsymbol{U})| &= \frac{1}{m}\left|\left\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}'\boldsymbol{U}'^{\mathrm{T}})\right\|_1 - \left\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})\right\|_1\right| \\
&\leq \frac{1}{m}\left\|\mathcal{A}(\boldsymbol{U}'\boldsymbol{U}'^{\mathrm{T}} - \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})\right\|_1 \\
&\leq \left(\sqrt{\frac{2}{\pi}} + \delta\right)\left\|\boldsymbol{U}'\boldsymbol{U}'^{\mathrm{T}} - \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\right\|_F \\
&= \left(\sqrt{\frac{2}{\pi}} + \delta\right)\left\|(\boldsymbol{U}' - \boldsymbol{U})\boldsymbol{U}^{\mathrm{T}} + \boldsymbol{U}'(\boldsymbol{U}' - \boldsymbol{U})^{\mathrm{T}}\right\|_F \\
&\leq \left(\sqrt{\frac{2}{\pi}} + \delta\right)(\|\boldsymbol{U}\| + \|\boldsymbol{U}'\|)\|\boldsymbol{U}' - \boldsymbol{U}\|_F,
\end{aligned}$$

where the second inequality follows from the ℓ_1/ℓ_2-RIP of $\mathcal{A}$. It follows that

$$\begin{aligned}
\liminf_{\boldsymbol{U}' \to \boldsymbol{U}} \frac{|f(\boldsymbol{U}') - f(\boldsymbol{U})|}{\|\boldsymbol{U} - \boldsymbol{U}'\|_F} &\leq \lim_{\boldsymbol{U}' \to \boldsymbol{U}} \frac{(\sqrt{2/\pi} + \delta)(\|\boldsymbol{U}\| + \|\boldsymbol{U}'\|)\|\boldsymbol{U}' - \boldsymbol{U}\|_F}{\|\boldsymbol{U}' - \boldsymbol{U}\|_F} \\
&= 2\left(\sqrt{\frac{2}{\pi}} + \delta\right)\|\boldsymbol{U}\|.
\end{aligned}$$

Upon taking $\boldsymbol{U}' = \boldsymbol{U} + t\boldsymbol{D}$, $t \to 0$ and invoking (3.8), we get

$$\|\boldsymbol{D}\|_F \leq 2\left(\sqrt{\frac{2}{\pi}} + \delta\right)\|\boldsymbol{U}\|, \ \forall \ \boldsymbol{D} \in \partial f(\boldsymbol{U}).$$

To complete the proof, it remains to note that for any $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ satisfying $\operatorname{dist}(\boldsymbol{U}, \mathcal{U}) \leq \frac{2\alpha}{\tau}$, where α, τ are given in (3.4), (3.6), respectively, the triangle inequality yields $\|\boldsymbol{U}\| \leq \|\boldsymbol{U}^\star\|_F + \frac{2\alpha}{\tau}$. □

By collecting Proposition 2, Proposition 3, and Proposition 4 together and invoking Theorem 2, we obtain the following guarantees for the SubGM[2] when it is applied to the robust low-rank matrix recovery problem (1.4):

THEOREM 3 (nonconvex robust low-rank matrix recovery: PSD case). *Consider the measurement model (1.3), where $\boldsymbol{X}^\star$ is an $n \times n$ rank-r symmetric positive semidefinite matrix. Let $0 < \delta < \frac{1}{3}\sqrt{\frac{2}{\pi}}$ be given. Suppose that the fraction of outliers p in the measurement vector $\boldsymbol{y}$ satisfies (3.3), and that the linear operators $\mathcal{A}$, $\mathcal{A}_{\Omega^c}$ possess the ℓ_1/ℓ_2-RIP (3.1), (3.2), respectively. Let α, τ, and κ be given by (3.4), (3.6), and (3.7), respectively. Under such setting, suppose that we apply the SubGM in Algorithm 2.1 to solve (1.4), where the initial point $\boldsymbol{U}_0$ satisfies $\operatorname{dist}(\boldsymbol{U}_0, \mathcal{U}) < \frac{2\alpha}{\tau}$ and the geometrically diminishing step sizes $\mu_k = \rho^k \mu_0$ are used with μ_0, ρ satisfying (2.4), (2.5), respectively. Then, the sequence of iterates $\{\boldsymbol{U}_k\}_{k\geq 0}$ generated by the SubGM will converge to a point in $\mathcal{U}$ at a linear rate:*

$$\operatorname{dist}(\boldsymbol{U}_k, \mathcal{U}) \leq \rho^k \max\left\{\operatorname{dist}(\boldsymbol{U}_0, \mathcal{U}), \mu_0 \frac{\max\{\kappa^2, 2\alpha^2\}}{\alpha}\right\}.$$

Moreover, the ground-truth matrix $\boldsymbol{X}^\star$ can be exactly recovered by any point $\boldsymbol{U}^\star \in \mathcal{U}$ via $\boldsymbol{X}^\star = \boldsymbol{U}^\star \boldsymbol{U}^{\star\mathrm{T}}$.

We remark that a similar result for the smooth counterpart (1.2) without any outliers is established in [42, Theorem 3.3]. Our Theorem 3 implies that the nonsmooth problem (1.4) can be solved *as efficiently as* its smooth counterpart (1.2), even in the presence of a substantial fraction of outliers in the measurement vector.

3.5. Initializing the SubGM. We now discuss some potential initialization strategies for the SubGM. A common approach to generating an appropriate initialization for matrix recovery-type problems is the spectral method. In our context, this entails simply computing the rank-r approximation of $\frac{1}{m}\mathcal{A}^*(\boldsymbol{y}) = \frac{1}{m}\sum_{i=1}^m y_i \boldsymbol{A}_i$, where $\mathcal{A}^*$ is the adjoint operator of $\mathcal{A}$. Specifically, let $\boldsymbol{P}\boldsymbol{\Pi}\boldsymbol{Q}^\mathrm{T}$ be a rank-r SVD of $\frac{1}{m}\mathcal{A}^*(\boldsymbol{y})$, where $\boldsymbol{P}, \boldsymbol{Q}$ have orthonormal columns and $\boldsymbol{\Pi}$ is an $r \times r$ diagonal matrix with the top r singular values of $\frac{1}{m}\mathcal{A}^*(\boldsymbol{y})$ along its diagonal. In the symmetric positive semidefinite case, we may assume without loss of generality that $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m$ are symmetric. Then, we can take $\boldsymbol{U}_0 = \boldsymbol{P}\boldsymbol{\Pi}^{1/2}$ as the initialization. The main idea behind this approach is that when there is no outlier (i.e., $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}^\star)$ as in (1.1)), we have $\frac{1}{m}\mathcal{A}^*(\boldsymbol{y}) = \frac{1}{m}\mathcal{A}^*(\mathcal{A}(\boldsymbol{X}^\star)) \approx \boldsymbol{X}^\star$ when $\frac{1}{m}\mathcal{A}^*\mathcal{A}$ is close to an unitary operator for low-rank matrices. Thus, $\boldsymbol{U}_0$ is also expected to be close to $\mathcal{U}$. However, when the measurements are corrupted by outliers, it is possible that $\frac{1}{m}\mathcal{A}^*(\boldsymbol{y})$ is perturbed away from $\frac{1}{m}\mathcal{A}^*(\mathcal{A}(\boldsymbol{X}^\star))$ and thus $\boldsymbol{U}_0$ may not be close enough to $\mathcal{U}$. To mitigate the influence of outliers, Li et al. [30] have recently proposed a truncated spectral method for initialization, in which the spectral method is applied to an operator that is formed by using those measurements whose absolute values do not deviate too much from the median of the absolute values of certain sampled measurements; see Algorithm 3.1. They showed that under appropriate conditions, the truncated spectral method can output an initialization that satisfies the requirement of Theorem 3.

THEOREM 4 (proximity of initialization to optimal set: PSD case; cf. [30, Theorem 3.3]). *Let $r \geq 1$ be given and set $\overline{c} = \frac{\|\boldsymbol{X}^\star\|_F}{\sqrt{r}\sigma_r(\boldsymbol{X}^\star)}$. Suppose that the matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m \in \mathbb{R}^{n\times n}$ defining the linear measurement operator $\mathcal{A}$ are symmetric and*

[2] In practice, we can just take $\operatorname{Sign}(0) = 0$ when applying the SubGM to solve (1.4).

Algorithm 3.1 Truncated Spectral Method for Initialization [30]

Input: measurement vector $\boldsymbol{y}$; sensing matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m$; threshold $\beta > 0$;

1: set $\boldsymbol{y}_1 = \{y_i\}_{i=1}^{\lfloor m/2 \rfloor}$, $\boldsymbol{y}_2 = \{y_i\}_{\lfloor m/2 \rfloor + 1}^{m}$;

2: Compute the rank-r SVD of

$$\boldsymbol{E} = \frac{1}{\lfloor m/2 \rfloor} \sum_{i=1}^{\lfloor m/2 \rfloor} y_i \boldsymbol{A}_i \mathbb{I}_{\{|y_i| \le \beta \cdot \mathrm{median}(|\boldsymbol{y}_2|)\}}$$

and denote it by $\boldsymbol{P}\boldsymbol{\Pi}\boldsymbol{Q}^T$, where

$$\mathbb{I}_{\{|y_i| \le \beta \cdot \mathrm{median}(|\boldsymbol{y}_2|)\}} = \begin{cases} 1 & \text{if } |y_i| \le \beta \cdot \mathrm{median}(|\boldsymbol{y}_2|), \\ 0 & \text{otherwise;} \end{cases}$$

Output: $\boldsymbol{U}_0 = \boldsymbol{P}\boldsymbol{\Pi}^{1/2}$, $\boldsymbol{V}_0 = \boldsymbol{Q}\boldsymbol{\Pi}^{1/2}$;

have i.i.d. standard Gaussian entries on and above the diagonal, and that the number of measurements m satisfies $m \gtrsim \beta^2 \overline{c}^2 n r^2 \log n$, where $\beta = 2\log\left(r^{1/4}\overline{c}^{1/2} + 20\right)$. Furthermore, suppose that the fraction of outliers p in the measurement vector $\boldsymbol{y}$ satisfies $p \lesssim \frac{1}{\sqrt{r\overline{c}}}$. Then, with overwhelming probability, Algorithm 3.1 outputs an initialization $\boldsymbol{U}_0 \in \mathbb{R}^{n\times r}$ satisfying $\mathrm{dist}(\boldsymbol{U}_0, \mathcal{U}) \lesssim \sigma_r(\boldsymbol{X}^\star)$ *and hence also the requirement of Theorem 3 (as $\sigma_r(\boldsymbol{X}^\star)$ is of the same order as $\frac{2\alpha}{\tau}$).*

Note that the requirements on the number of measurements and the fraction of outliers that can be tolerated are slightly more stringent than those in Proposition 1 and Theorem 3. However, as will be illustrated in Section 5, our numerical experiments show that even a randomly initialized SubGM can very efficiently find the global minimum and hence recover the ground-truth matrix $\boldsymbol{X}^\star$. A theoretical justification of such a phenomenon will be the subject of a future study. We suspect that it may be possible to relax the requirement on the initialization in Theorem 3 or to show that the SubGM enters the region $\left\{\boldsymbol{U} : \mathrm{dist}(\boldsymbol{U}, \mathcal{U}) < \frac{2\alpha}{\tau}\right\}$ very quickly even though the random initialization lies outside of this region.

4. Nonconvex Robust Low-Rank Matrix Recovery: General Case.

In this section we consider the general setting where $\boldsymbol{X}^\star$ is a rank-r $n_1 \times n_2$ matrix. To extend the nonsmooth nonconvex formulation (1.4) to this setting, a natural approach is to use the factorization $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ with $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n_2 \times r}$. However, such a factorization is ambiguous in the sense that if $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$, then $\boldsymbol{X} = (\boldsymbol{U}\boldsymbol{T})(\boldsymbol{V}\boldsymbol{T}^{-\mathrm{T}})^{\mathrm{T}}$ for any invertible matrix $\boldsymbol{T} \in \mathbb{R}^{r\times r}$. To address this issue, we introduce the nonsmooth nonconvex regularizer

$$\phi(\boldsymbol{U}, \boldsymbol{V}) := \|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\|_F, \tag{4.1}$$

which aims to balance the factors $\boldsymbol{U}$ and $\boldsymbol{V}$, and solve the following regularized problem:

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n_1\times r}, \boldsymbol{V}\in\mathbb{R}^{n_2\times r}}{\mathrm{minimize}} \left\{ g(\boldsymbol{U}, \boldsymbol{V}) := \frac{1}{m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})\|_1 + \lambda \|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\|_F \right\}. \tag{4.2}$$

Here, $\lambda > 0$ is a regularization parameter. We remark that a similar regularizer, namely,

$$\widetilde{\phi}(\boldsymbol{U}, \boldsymbol{V}) := \|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\|_F^2,$$

has been introduced in [35, 42, 52] to account for the ambiguities caused by invertible transformations when minimizing the squared ℓ_2-loss function $(\boldsymbol{U},\boldsymbol{V}) \mapsto \frac{1}{m}\|\boldsymbol{y}-\mathcal{A}(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})\|_2^2$. However, such a regularizer is not entirely suitable for the ℓ_1-loss function, as it is no longer clear that the resulting problem will satisfy the sharpness condition in Definition 1.

To simplify notation, we stack $\boldsymbol{U}$ and $\boldsymbol{V}$ together as $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}^{\mathrm{T}} & \boldsymbol{V}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}$ and write $g(\boldsymbol{W})$ for $g(\boldsymbol{U},\boldsymbol{V})$. Observe that the regularizer ϕ achieves its minimum value of 0 when $\boldsymbol{U}$ and $\boldsymbol{V}$ have the same Gram matrices; i.e., $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}$. Now, let $\boldsymbol{X}^\star = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{\mathrm{T}}$ be a rank-r SVD of $\boldsymbol{X}^\star$, where $\boldsymbol{\Phi} \in \mathbb{R}^{n_1\times r}, \boldsymbol{\Psi} \in \mathbb{R}^{n_2\times r}$ have orthonormal columns and $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$ is a diagonal matrix. Define

$$\boldsymbol{U}^\star = \boldsymbol{\Phi}\boldsymbol{\Sigma}^{1/2}, \quad \boldsymbol{V}^\star = \boldsymbol{\Psi}\boldsymbol{\Sigma}^{1/2}, \quad \boldsymbol{W}^\star = \begin{bmatrix}\boldsymbol{U}^{\star\mathrm{T}} & \boldsymbol{V}^{\star\mathrm{T}}\end{bmatrix}^{\mathrm{T}}.$$

The orthogonal invariance of g (i.e., $g(\boldsymbol{W}) = g(\boldsymbol{W}\boldsymbol{R})$ for any $\boldsymbol{R} \in \mathcal{O}_r$) implies that g is constant on the set

$$\mathcal{W} := \{\boldsymbol{W}^\star\boldsymbol{R} : \boldsymbol{R} \in \mathcal{O}_r\}.$$

4.1. Sharpness and Exact Recovery. Our immediate goal is to show that $\mathcal{W}$ is the set of global minima of (4.2). Towards that end, let $0 < \delta < \frac{1}{3}\sqrt{\frac{2}{\pi}}$ be given. Suppose that the fraction of outliers p in the measurement vector $\boldsymbol{y}$ satisfies (3.3), and that the linear operators $\mathcal{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ and $\mathcal{A}_{\Omega^c} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{|\Omega^c|}$ possess the ℓ_1/ℓ_2-RIP (3.1) and (3.2), respectively.[3] Using the argument in the proof of Proposition 2, we get

$$\overline{g}(\boldsymbol{W}) - \overline{g}(\boldsymbol{W}^\star) \geq \left(2(1-p)\left(\sqrt{\frac{2}{\pi}}-\delta\right) - \left(\sqrt{\frac{2}{\pi}}+\delta\right)\right)\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star\|_F, \tag{4.3}$$

where

$$\overline{g}(\boldsymbol{W}) = \frac{1}{m}\|\boldsymbol{y}-\mathcal{A}(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})\|_1.$$

In particular, we see that $\overline{g}(\boldsymbol{W}) > \overline{g}(\boldsymbol{W}^\star)$ whenever $\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} \neq \boldsymbol{X}^\star$. Since $\boldsymbol{U}^{\star\mathrm{T}}\boldsymbol{U}^\star = \boldsymbol{V}^{\star\mathrm{T}}\boldsymbol{V}^\star$ by construction, we conclude that $\boldsymbol{W}^\star$ is a global minimum of (4.2), as $\boldsymbol{W}^\star$ is a global minimum of both the first term $\overline{g}$ and the second term ϕ of g. It then follows from the orthogonal invariance of g that every element in $\mathcal{W}$ is a global minimum of (4.2). The following result further establishes that $\mathcal{W}$ is exactly the set of global minima of (4.2) and g is sharp.

PROPOSITION 5 (sharpness and exact recovery with outliers: general case). *Let $0 < \delta < \frac{1}{3}\sqrt{\frac{2}{\pi}}$ be given. Suppose that the fraction of outliers p satsifies* (3.3)*, and that the linear operators $\mathcal{A}$ and $\mathcal{A}_{\Omega^c}$ possess the ℓ_1/ℓ_2-RIP* (3.1) *and* (3.2)*, respectively. Then, the objective function g in* (4.2) *satisfies*

$$g(\boldsymbol{W}) - g(\boldsymbol{W}^\star) \geq \alpha\,\mathrm{dist}(\boldsymbol{W},\mathcal{W})$$

for any $\boldsymbol{W} \in \mathbb{R}^{(n_1+n_2)\times r}$, where

$$\alpha = \sqrt{\sqrt{2}-1}\cdot\min\left\{2(1-p)\left(\sqrt{\frac{2}{\pi}}-\delta\right) - \left(\sqrt{\frac{2}{\pi}}+\delta\right), 2\lambda\right\}\cdot\sigma_r(\boldsymbol{X}^\star) > 0. \tag{4.4}$$

[3] It can be shown that modulo the constants, the Gaussian measurement operator $\mathcal{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ will possess the ℓ_1/ℓ_2-RIPs (3.1) and (3.2) with high probability as long as $m \gtrsim \max\{n_1, n_2\}r$. To avoid any distraction caused by the new constants, we shall simply use the ℓ_1/ℓ_2-RIPs (3.1) and (3.2) in our derivation.

In particular, the set $\mathcal{W}$ is precisely the set of global minima of (4.2) and the objective function g is sharp with parameter $\alpha > 0$.

Proof of Proposition 5. Let $\zeta(p,\delta) = 2(1-p)\left(\sqrt{\frac{2}{\pi}} - \delta\right) - \left(\sqrt{\frac{2}{\pi}} + \delta\right)$. Since $\boldsymbol{U}^{\star \mathrm{T}}\boldsymbol{U}^\star = \boldsymbol{V}^{\star \mathrm{T}}\boldsymbol{V}^\star$, we have $\phi(\boldsymbol{W}^\star) = 0$ by (4.1) and

$$\begin{aligned}
g(\boldsymbol{W}) - g(\boldsymbol{W}^\star) &= \frac{1}{m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^\mathrm{T})\|_1 - \frac{1}{m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X}^\star)\|_1 + \lambda\|\boldsymbol{U}^\mathrm{T}\boldsymbol{U} - \boldsymbol{V}^\mathrm{T}\boldsymbol{V}\|_F \\
&\geq \zeta(p,\delta)\|\boldsymbol{X}^\star - \boldsymbol{U}\boldsymbol{V}^\mathrm{T}\|_F + \lambda\|\boldsymbol{U}^\mathrm{T}\boldsymbol{U} - \boldsymbol{V}^\mathrm{T}\boldsymbol{V}\|_F \\
&\geq \min\{\zeta(p,\delta), 2\lambda\}\left(\|\boldsymbol{X}^\star - \boldsymbol{U}\boldsymbol{V}^\mathrm{T}\|_F + \frac{1}{2}\|\boldsymbol{U}^\mathrm{T}\boldsymbol{U} - \boldsymbol{V}^\mathrm{T}\boldsymbol{V}\|_F\right) \\
&\geq \min\{\zeta(p,\delta), 2\lambda\}\sqrt{\|\boldsymbol{X}^\star - \boldsymbol{U}\boldsymbol{V}^\mathrm{T}\|_F^2 + \frac{1}{4}\|\boldsymbol{U}^\mathrm{T}\boldsymbol{U} - \boldsymbol{V}^\mathrm{T}\boldsymbol{V}\|_F^2} \\
&\geq \min\left\{\frac{\zeta(p,\delta)}{2}, \lambda\right\}\|\boldsymbol{W}\boldsymbol{W}^\mathrm{T} - \boldsymbol{W}^\star\boldsymbol{W}^{\star \mathrm{T}}\|_F \\
&\geq \min\left\{\frac{\zeta(p,\delta)}{2}, \lambda\right\}\sqrt{2\left(\sqrt{2}-1\right)\sigma_r(\boldsymbol{W}^\star)}\,\mathrm{dist}(\boldsymbol{W}, \mathcal{W}) \\
&= \min\{\zeta(p,\delta), 2\lambda\}\sqrt{\sqrt{2}-1}\sigma_r^{1/2}(\boldsymbol{X}^\star)\,\mathrm{dist}(\boldsymbol{W}, \mathcal{W}),
\end{aligned}$$

where the first inequality follws from (4.3), the fourth inequality follows from

$$\begin{aligned}
\|\boldsymbol{X}^\star - \boldsymbol{U}\boldsymbol{V}^\mathrm{T}\|_F^2 + \frac{1}{4}\|\boldsymbol{U}^\mathrm{T}\boldsymbol{U} - \boldsymbol{V}^\mathrm{T}\boldsymbol{V}\|_F^2 &= \|\boldsymbol{U}^\star\boldsymbol{V}^{\star \mathrm{T}} - \boldsymbol{U}\boldsymbol{V}^\mathrm{T}\|_F^2 + \frac{1}{4}\|\boldsymbol{U}^\mathrm{T}\boldsymbol{U} - \boldsymbol{V}^\mathrm{T}\boldsymbol{V}\|_F^2 \\
&= \frac{1}{4}\|\boldsymbol{W}\boldsymbol{W}^\mathrm{T} - \boldsymbol{W}^\star\boldsymbol{W}^{\star \mathrm{T}}\|_F^2 + \nu(\boldsymbol{W})
\end{aligned}$$

with

$$\begin{aligned}
\nu(\boldsymbol{W}) &= \frac{1}{2}\|\boldsymbol{U}\boldsymbol{V}^\mathrm{T} - \boldsymbol{U}^\star\boldsymbol{V}^{\star \mathrm{T}}\|_F^2 + \frac{1}{4}\|\boldsymbol{U}^\mathrm{T}\boldsymbol{U} - \boldsymbol{V}^\mathrm{T}\boldsymbol{V}\|_F^2 \\
&\quad - \frac{1}{4}\|\boldsymbol{U}\boldsymbol{U}^\mathrm{T} - \boldsymbol{U}^\star\boldsymbol{U}^{\star \mathrm{T}}\|_F^2 - \frac{1}{4}\|\boldsymbol{V}\boldsymbol{V}^\mathrm{T} - \boldsymbol{V}^\star\boldsymbol{V}^{\star \mathrm{T}}\|_F^2 \\
&= \frac{1}{2}\|\boldsymbol{U}^\mathrm{T}\boldsymbol{U}^\star\|_F^2 + \frac{1}{2}\|\boldsymbol{V}^\mathrm{T}\boldsymbol{V}^\star\|_F^2 - \left\langle \boldsymbol{U}\boldsymbol{V}^\mathrm{T}, \boldsymbol{U}^\star\boldsymbol{V}^{\star \mathrm{T}}\right\rangle \\
&\quad + \frac{1}{2}\|\boldsymbol{U}^\star\boldsymbol{V}^{\star \mathrm{T}}\|_F^2 - \frac{1}{4}\|\boldsymbol{U}^\star\boldsymbol{U}^{\star \mathrm{T}}\|_F^2 - \frac{1}{4}\|\boldsymbol{V}^\star\boldsymbol{V}^{\star \mathrm{T}}\|_F^2 \\
&= \frac{1}{2}\|\boldsymbol{U}^\mathrm{T}\boldsymbol{U}^\star - \boldsymbol{V}^\mathrm{T}\boldsymbol{V}^\star\|_F^2 + \frac{1}{2}\|\boldsymbol{U}^\star\boldsymbol{V}^{\star \mathrm{T}}\|_F^2 - \frac{1}{4}\|\boldsymbol{U}^\star\boldsymbol{U}^{\star \mathrm{T}}\|_F^2 - \frac{1}{4}\|\boldsymbol{V}^\star\boldsymbol{V}^{\star \mathrm{T}}\|_F^2 \\
&= \frac{1}{2}\|\boldsymbol{U}^\mathrm{T}\boldsymbol{U}^\star - \boldsymbol{V}^\mathrm{T}\boldsymbol{V}^\star\|_F^2 \geq 0
\end{aligned}$$

(recall that $\boldsymbol{U}^{\star \mathrm{T}}\boldsymbol{U}^\star = \boldsymbol{V}^{\star \mathrm{T}}\boldsymbol{V}^\star$), the fifth inequality is from Lemma 1, and the last equality follows from the fact that $\sigma_r(\boldsymbol{W}^\star) = \sqrt{2}\sigma_r^{1/2}(\boldsymbol{X}^\star)$. This completes the proof. ∎

By comparing Proposition 2 and Proposition 5, we see that the fraction of outliers that can be tolerated for exact recovery is the same in both the symmetric positive semidefinite and general cases. Moreover, the sharpness parameter α in (4.4) demonstrates the role that the regularizer ϕ plays: When the regularizer ϕ is absent (which corresponds to $\lambda = 0$), although every element in $\mathcal{W}$ is still a global minimum of (4.2), we cannot guarantee that there is no other global minimum. Indeed, when $\lambda = 0$, the pair $(\boldsymbol{U}^\star\boldsymbol{T}, \boldsymbol{V}^\star\boldsymbol{T}^{-\mathrm{T}})$ is a global minimum of (4.2) for any invertible matrix $\boldsymbol{T} \in \mathbb{R}^{r\times r}$. However, when $\lambda > 0$, the regularizer ϕ ensures that the pair $(\boldsymbol{U}^\star\boldsymbol{T}, \boldsymbol{V}^\star\boldsymbol{T}^{-\mathrm{T}})$ is a global minimum of (4.2) only when $\boldsymbol{T} \in \mathcal{O}_r$.

4.2. Weak Convexity. Let us now establish the weak convexity of the objective function g in (4.2).

PROPOSITION 6 (weak convexity: general case). *Suppose that the measurement operator $\mathcal{A}$ satisfies the ℓ_1/ℓ_2-RIP (3.1). Then, the objective function g in (4.2) is weakly convex with parameter*

$$\tau = \sqrt{\frac{2}{\pi}} + \delta + 2\lambda. \tag{4.5}$$

Proof of Proposition 6. Since $g = \overline{g} + \lambda\phi$, it suffices to show that $\overline{g}$ and ϕ are both weakly convex. Similar to (3.5), we apply the chain rule for subdifferentials [37, Corollary 8.11 and Theorem 10.6] to get

$$\partial\overline{g}(\boldsymbol{W}) = \frac{1}{m}\begin{bmatrix} \mathcal{A}^*\left(\mathrm{Sign}\left(\mathcal{A}(\boldsymbol{U}\boldsymbol{V}^\mathrm{T}) - \boldsymbol{y}\right)\right)\boldsymbol{V} \\ \left(\mathcal{A}^*\left(\mathrm{Sign}\left(\mathcal{A}(\boldsymbol{U}\boldsymbol{V}^\mathrm{T}) - \boldsymbol{y}\right)\right)\right)^\mathrm{T}\boldsymbol{U}\end{bmatrix}.$$

Using this and the argument in the proof of Proposition 3, we can show that for any $\boldsymbol{W}, \boldsymbol{W}' \in \mathbb{R}^{(n_1+n_2)\times r}$,

$$\begin{aligned}\overline{g}(\boldsymbol{W}') &\geq \overline{g}(\boldsymbol{W}) + \langle \boldsymbol{D}, \boldsymbol{W}' - \boldsymbol{W}\rangle - \left(\sqrt{\frac{2}{\pi}} + \delta\right)\|(\boldsymbol{U}' - \boldsymbol{U})(\boldsymbol{V}' - \boldsymbol{V})^\mathrm{T}\|_F \\ &\geq \overline{g}(\boldsymbol{W}) + \langle \boldsymbol{D}, \boldsymbol{W}' - \boldsymbol{W}\rangle - \left(\frac{\sqrt{2/\pi} + \delta}{2}\right)\|\boldsymbol{W}' - \boldsymbol{W}\|_F^2,\ \forall\ \boldsymbol{D} \in \partial\overline{g}(\boldsymbol{W});\end{aligned}$$

i.e., the function $\overline{g}$ is weakly convex with parameter $\tau_{\overline{g}} = \sqrt{\frac{2}{\pi}} + \delta$.

Next, define the matrices

$$\underline{\boldsymbol{W}} = \begin{bmatrix}\boldsymbol{U}^\mathrm{T} & -\boldsymbol{V}^\mathrm{T}\end{bmatrix}^\mathrm{T}, \quad \underline{\boldsymbol{W}'} = \begin{bmatrix}\boldsymbol{U}'^\mathrm{T} & -\boldsymbol{V}'^\mathrm{T}\end{bmatrix}^\mathrm{T}$$

and note that $\underline{\boldsymbol{W}}^\mathrm{T}\boldsymbol{W} = \boldsymbol{U}^\mathrm{T}\boldsymbol{U} - \boldsymbol{V}^\mathrm{T}\boldsymbol{V}$. Furthermore, define the function $\psi : \mathbb{R}^{r\times r} \to \mathbb{R}$ by $\psi(\boldsymbol{C}) = \|\boldsymbol{C}\|_F$, whose subdifferential is

$$\partial\psi(\boldsymbol{C}) = \begin{cases}\left\{\frac{\boldsymbol{C}}{\|\boldsymbol{C}\|_F}\right\}, & \boldsymbol{C} \neq \boldsymbol{0}, \\ \left\{\boldsymbol{B} \in \mathbb{R}^{r\times r} : \|\boldsymbol{B}\|_F \leq 1\right\}, & \boldsymbol{C} = \boldsymbol{0}.\end{cases}$$

Upon setting $\boldsymbol{\Delta} = \boldsymbol{W}' - \boldsymbol{W}$ and $\underline{\boldsymbol{\Delta}} = \underline{\boldsymbol{W}'} - \underline{\boldsymbol{W}}$, we compute

$$\begin{aligned}\phi(\boldsymbol{W}') &= \|\underline{\boldsymbol{W}'}^\mathrm{T}\boldsymbol{W}'\|_F \\ &= \|\underline{\boldsymbol{W}}^\mathrm{T}\boldsymbol{W} + \underline{\boldsymbol{W}}^\mathrm{T}\boldsymbol{\Delta} + \underline{\boldsymbol{\Delta}}^\mathrm{T}\boldsymbol{W} + \underline{\boldsymbol{\Delta}}^\mathrm{T}\boldsymbol{\Delta}\|_F \\ &\geq \|\underline{\boldsymbol{W}}^\mathrm{T}\boldsymbol{W} + \underline{\boldsymbol{W}}^\mathrm{T}\boldsymbol{\Delta} + \underline{\boldsymbol{\Delta}}^\mathrm{T}\boldsymbol{W}\|_F - \|\underline{\boldsymbol{\Delta}}^\mathrm{T}\boldsymbol{\Delta}\|_F \\ &\geq \|\underline{\boldsymbol{W}}^\mathrm{T}\boldsymbol{W}\|_F + \left\langle\boldsymbol{\Psi}, \underline{\boldsymbol{W}}^\mathrm{T}\boldsymbol{\Delta} + \underline{\boldsymbol{\Delta}}^\mathrm{T}\boldsymbol{W}\right\rangle - \|\underline{\boldsymbol{\Delta}}^\mathrm{T}\boldsymbol{\Delta}\|_F,\end{aligned} \tag{4.6}$$

where the last inequality holds for any $\boldsymbol{\Psi} \in \partial\psi(\underline{\boldsymbol{W}}^\mathrm{T}\boldsymbol{W})$ due to the convexity of the Frobenius norm. Since the Frobenius norm is subdifferentially regular [37, Example 7.27], the chain rule for subdifferentials [37, Corollary 8.11 and Theorem 10.6] yields

$$\partial\phi(\boldsymbol{W}) = \left\{\underline{\boldsymbol{W}}(\boldsymbol{\Psi} + \boldsymbol{\Psi}^\mathrm{T}) : \boldsymbol{\Psi} \in \partial\psi(\underline{\boldsymbol{W}}^\mathrm{T}\boldsymbol{W})\right\}. \tag{4.7}$$

It follows from (4.6) and (4.7) that

$$\begin{aligned}\phi(\boldsymbol{W}') &\geq \phi(\boldsymbol{W}) + \langle \boldsymbol{\Phi}, \boldsymbol{W}' - \boldsymbol{W}\rangle - \|\underline{\boldsymbol{\Delta}}^{\mathrm{T}}\boldsymbol{\Delta}\|_F \\ &\geq \phi(\boldsymbol{W}) + \langle \boldsymbol{\Phi}, \boldsymbol{W}' - \boldsymbol{W}\rangle - \|\boldsymbol{W}' - \boldsymbol{W}\|_F^2, \ \forall \boldsymbol{\Phi} \in \partial\phi(\boldsymbol{W});\end{aligned}$$

i.e., the function ϕ is weakly convex with parameter $\tau_\phi = 2$.

Putting the above results together, we conclude that $g = \overline{g} + \lambda\phi$ is weakly convex with parameter $\tau = \tau_{\overline{g}} + \lambda\tau_\phi$, as desired. ∎

Unlike the sharpness condition in Proposition 5 that requires $\lambda > 0$, the weak convexity condition in Proposition 6 holds even when $\lambda = 0$. Although the parameters α and τ in (4.4) and (4.5) increase as λ increases from 0, the former becomes constant when $\lambda \geq \frac{2(1-p)\left(\sqrt{2/\pi}-\delta\right)-\left(\sqrt{2/\pi}+\delta\right)}{2}$. In view of Theorem 2, it is desirable to choose λ so that the local linear convergence region $\left\{\boldsymbol{x} : \mathrm{dist}(\boldsymbol{x}, \mathcal{X}) < \frac{2\alpha}{\tau}\right\}$ of the SubGM is as large as possible. Such consideration suggests that we should set

$$\lambda = \frac{2(1-p)\left(\sqrt{2/\pi}-\delta\right)-\left(\sqrt{2/\pi}+\delta\right)}{2}.$$

4.3. Putting Everything Together. As in Subsection 3.4, before we can invoke Theorem 2 to establish convergence guarantees for the SubGM when applied to the general robust low-rank matrix recovery problem (4.2), we need to bound the norm of any subgradient of g in a neighborhood of $\mathcal{W}$. This is achieved by the following result:

PROPOSITION 7 (bound on subgradient norm: general case). *Suppose that the measurement operator $\mathcal{A}$ satisfies the ℓ_1/ℓ_2-RIP (3.1). Then, for any $\boldsymbol{W} \in \mathbb{R}^{(n_1+n_2)\times r}$ satisfying* $\mathrm{dist}(\boldsymbol{W}, \mathcal{W}) \leq \frac{2\alpha}{\tau}$, *we have*

$$\|\boldsymbol{D}\|_F \leq \kappa = \max\left\{\sqrt{\frac{2}{\pi}} + \delta, \lambda\right\}\left(\|\boldsymbol{W}^\star\|_F + \frac{2\alpha}{\tau}\right), \ \forall\ \boldsymbol{D} \in \partial g(\boldsymbol{W}). \tag{4.8}$$

Proof of Proposition 7. Observe that for any $\boldsymbol{W}, \boldsymbol{W}' \in \mathbb{R}^{(n_1+n_2)\times r}$,

$$\begin{aligned}&|g(\boldsymbol{W}') - g(\boldsymbol{W})| \leq |\overline{g}(\boldsymbol{W}') - \overline{g}(\boldsymbol{W})| + \lambda\,|\phi(\boldsymbol{W}') - \phi(\boldsymbol{W})| \\ &\leq \frac{1}{m}\left\|\mathcal{A}(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{U}'\boldsymbol{V}'^{\mathrm{T}})\right\|_1 + \lambda\left(\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{U}'^{\mathrm{T}}\boldsymbol{U}'\right\|_F + \left\|\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} - \boldsymbol{V}'^{\mathrm{T}}\boldsymbol{V}'\right\|_F\right) \\ &\leq \left(\sqrt{\frac{2}{\pi}} + \delta\right)\left\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{U}'\boldsymbol{V}'^{\mathrm{T}}\right\|_F + \lambda\left(\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{U}'^{\mathrm{T}}\boldsymbol{U}'\right\|_F + \left\|\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} - \boldsymbol{V}'^{\mathrm{T}}\boldsymbol{V}'\right\|_F\right) \\ &\leq \left(\sqrt{\frac{2}{\pi}} + \delta\right)\left(\|\boldsymbol{V}\|_F\|\boldsymbol{U} - \boldsymbol{U}'\|_F + \|\boldsymbol{U}'\|_F\|\boldsymbol{V} - \boldsymbol{V}'\|_F\right) \\ &\quad + \lambda\left(\|\boldsymbol{U}\|_F + \|\boldsymbol{U}'\|_F\right)\|\boldsymbol{U} - \boldsymbol{U}'\|_F + \lambda\left(\|\boldsymbol{V}\|_F + \|\boldsymbol{V}'\|_F\right)\|\boldsymbol{V} - \boldsymbol{V}'\|_F \\ &\leq \max\left\{\sqrt{\frac{2}{\pi}} + \delta, \lambda\right\}\left(\|\boldsymbol{W}\|_F + \|\boldsymbol{W}'\|_F\right)\|\boldsymbol{W} - \boldsymbol{W}'\|_F,\end{aligned}$$

where the third inequality follows from the ℓ_1/ℓ_2-RIP (3.1). Thus, similar to the derivation of (3.7), for any $\boldsymbol{W} \in \mathbb{R}^{(n_1+n_2)\times r}$ satisfying $\mathrm{dist}(\boldsymbol{W}, \mathcal{W}) \leq \frac{2\alpha}{\tau}$, where α

and τ are given in (4.4) and (4.5), respectively, we have

$$\|\boldsymbol{D}\|_F \leq \max\left\{\sqrt{\frac{2}{\pi}}+\delta,\lambda\right\}\|\boldsymbol{W}\|_F$$
$$\leq \max\left\{\sqrt{\frac{2}{\pi}}+\delta,\lambda\right\}\left(\|\boldsymbol{W}^\star\|_F+\frac{2\alpha}{\tau}\right),\ \forall\ \boldsymbol{D}\in\partial g(\boldsymbol{W}).\qquad\square$$

By collecting Proposition 5, Proposition 6, and Proposition 7 together and invoking Theorem 2, we obtain the following guarantees when the SubGM is used to solve the general robust low-rank matrix recovery problem (4.2):

THEOREM 5 (nonconvex robust low-rank matrix recovery: general case). *Consider the measurement model (1.3), where $\boldsymbol{X}^\star$ is an $n_1\times n_2$ rank-r matrix. Let $0<\delta<\frac{1}{3}\sqrt{\frac{2}{\pi}}$ be given. Suppose that the fraction of outliers p in the measurement vector $\boldsymbol{y}$ satisfies (3.3), and that the linear operators $\mathcal{A}$, $\mathcal{A}_{\Omega^c}$ possess the ℓ_1/ℓ_2-RIP (3.1), (3.2), respectively. Let α, τ, and κ be given by (4.4), (4.5), and (4.8), respectively. Under such setting, suppose that we apply the SubGM in Algorithm 2.1 to solve (4.2), where the initial point $\boldsymbol{W}_0$ satisfies $\operatorname{dist}(\boldsymbol{W}_0,\mathcal{W})<\frac{2\alpha}{\tau}$ and the geometrically diminishing step sizes $\mu_k=\rho^k\mu_0$ are used with μ_0, ρ satisfying (2.4), (2.5), respectively. Then, the sequence of iterates $\{\boldsymbol{W}_k\}_{k\geq 0}$ generated by the SubGM will converge to a point in $\mathcal{W}$ at a linear rate:*

$$\operatorname{dist}(\boldsymbol{W}_k,\mathcal{W})\leq\rho^k\max\left\{\operatorname{dist}(\boldsymbol{W}_0,\mathcal{W}),\mu_0\frac{\max\{\kappa^2,2\alpha^2\}}{\alpha}\right\}.$$

Moreover, the ground-truth matrix $\boldsymbol{X}^\star$ can be exactly recovered by any point $\boldsymbol{W}^\star\in\mathcal{W}$ via $\boldsymbol{X}^\star=\boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}}$.

4.4. Initializing the SubGM. In the general case, we can still use the truncated spectral method in Algorithm 3.1 to obtain a good initialization for the SubGM. Specifically, we take $\boldsymbol{W}_0=\begin{bmatrix}\boldsymbol{U}_0^{\mathrm{T}} & \boldsymbol{V}_0^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}$ as the initialization, where $\boldsymbol{U}_0,\boldsymbol{V}_0$ are the outputs of Algorithm 3.1. Then, we have the following result, which is essentially a restatement of [30, Theorem 3.3]:

THEOREM 6 (proximity of initialization to optimal set: general case). *Let $r\geq 1$ be given and set $n=n_1+n_2$, $\overline{c}=\frac{\|\boldsymbol{X}^\star\|_F}{\sqrt{r}\sigma_r(\boldsymbol{X}^\star)}$. Suppose that the matrices $\boldsymbol{A}_1,\ldots,\boldsymbol{A}_m\in\mathbb{R}^{n_1\times n_2}$ defining the linear measurement operator $\mathcal{A}$ have i.i.d. standard Gaussian entries, and that the number of measurements m satisfies $m\gtrsim\beta^2\overline{c}^2nr^2\log n$, where $\beta=2\log\left(r^{1/4}\overline{c}^{1/2}+20\right)$. Furthermore, suppose that the fraction of outliers p in the measurement vector $\boldsymbol{y}$ satisfies $p\lesssim\frac{1}{\sqrt{r}\overline{c}}$. Then, with overwhelming probability, Algorithm 3.1 outputs an initialization $\boldsymbol{W}_0\in\mathbb{R}^{(n_1+n_2)\times r}$ satisfying $\operatorname{dist}(\boldsymbol{W}_0,\mathcal{U})\lesssim\sigma_r(\boldsymbol{X}^\star)$ and hence also the requirement of Theorem 5.*

5. Experiments. In this section we conduct experiments to illustrate the performance of the SubGM when applied to robust low-rank matrix recovery problems. The experiments on synthetic data show that the SubGM can exactly and efficiently recover the underlying low-rank matrix from its linear measurements even in the presence of outliers, thus corroborating the result in Theorem 3.

We generate the underlying low-rank matrix $\boldsymbol{X}^\star=\boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}}$ by generating $\boldsymbol{U}^\star\in\mathbb{R}^{n\times r}$ with *i.i.d.* standard Gaussian entries. Similarly, we generate the entries of the m sensing matrices $\boldsymbol{A}_1,\ldots,\boldsymbol{A}_m\in\mathbb{R}^{n\times n}$ (which define the linear measurement operator

$\mathcal{A}$) in an *i.i.d.* fashion according to the standard Gaussian distribution. To generate the outlier vector $\boldsymbol{s}^\star \in \mathbb{R}^m$, we first randomly select pm locations. Then, we fill each of the selected location with an *i.i.d.* mean 0 and variance 100 Gaussian entry, while the remaining locations are set to 0. Here, p is the ratio of the nonzero elements in $\boldsymbol{s}^\star$. According to (1.3), the measurement vector $\boldsymbol{y}$ is then generated by $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}^\star) + \boldsymbol{s}^\star$; i.e., $y_i = \langle \boldsymbol{A}_i, \boldsymbol{X}^\star \rangle + s_i^\star$ for $i = 1, \ldots, m$.

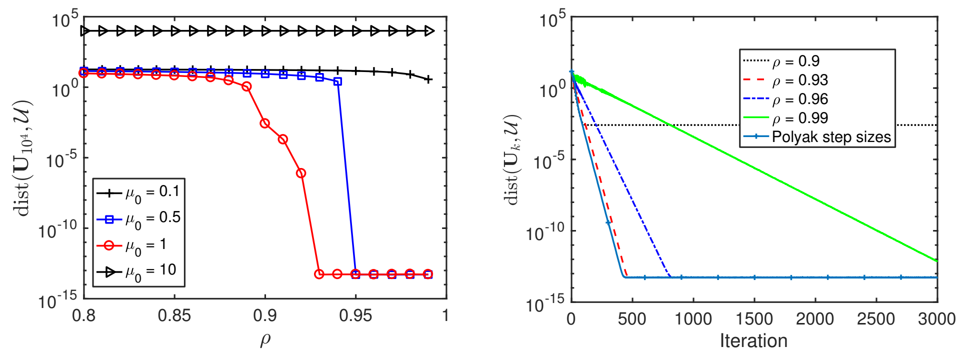
To illustrate the performance of the SubGM for recovering the underlying low-rank matrix $\boldsymbol{X}^\star$ from $\boldsymbol{y}$, we first set $n = 50$, $r = 5$, and $p = 0.3$. Throughout the experiments, we initialize the SubGM with a randomly generated standard Gaussian vector, as it gives similar practical performance as the one obtained by the truncated spectral method in Algorithm 3.1. We first run the SubGM for 10^4 iterations using the geometrically diminishing step sizes $\mu_k = \rho^k \mu_0$, where the initial step size μ_0 and decay rate ρ are selected from $\{0.1, 0.5, 1, 10\}$ and $\{0.80, 0.81, 0.82, \ldots, 0.99\}$, respectively. For each pair of parameters (μ_0, ρ), we plot the distance of the last iterate to $\mathcal{U}$ (i.e., $\operatorname{dist}(\boldsymbol{U}_{10^4}, \mathcal{U})$) in Figure 2a. When the SubGM diverges, we simply set $\operatorname{dist}(\boldsymbol{U}_{10^4}, \mathcal{U}) = 10^4$ for the purpose of presenting all results in the same figure. As observed from Figure 2a, the SubGM diverges when μ_0 is large, say, $\mu_0 = 10$. On the other hand, it converges to a global minimum when $\mu_0 = 1$, $\rho \in [0.93, 0.99]$ and $\mu_0 = 0.5$, $\rho \in [0.95, 0.99]$. It is worth noting that the SubGM converges to a global minimum when $\mu_0 = 1, \rho = 0.93$, but not when $\mu_0 = 0.5, \rho = 0.93$. This is consistent with Theorem 2, which shows that a larger initial step size μ_0 allows for a smaller decay rate ρ. Such a phenomenon can also be observed in the case where $\mu_0 = 0.1$, for which the SubGM fails to find a global minimum even when $\rho \in [0.95, 0.99]$.

In Figure 2b, we fix $\mu_0 = 1$ and plot the convergence behavior of the SubGM with $\rho \in \{0.9, 0.93, 0.96, 0.99\}$. As observed from the figure, when ρ is not too small (say, larger than 0.93), the distances $\{\operatorname{dist}(\boldsymbol{U}_k, \mathcal{U})\}_{k \geq 0}$ converge to 0 at a linear rate, thus implying that the SubGM with geometrically diminishing step sizes can exactly recover the underlying low-rank matrix $\boldsymbol{X}^\star$. We observe that a smaller ρ gives faster convergence. This corroborates the results in Theorem 2, which guarantee that $\{\operatorname{dist}(\boldsymbol{U}_k, \mathcal{U})\}_{k \geq 0}$ decays at the rate $O(\rho^k)$ as long as ρ is not too small (i.e., satisfying (2.5)). We also consider the SubGM with the Polyak step size rule [36], which, in the context of (1.4), is given by $\mu_k = \frac{f(\boldsymbol{U}_k) - f^\star}{\|\boldsymbol{d}_k\|^2}$, where $f^\star$ is the optimal value of (1.4) and $\boldsymbol{d}_k \in \partial f(\boldsymbol{U}_k)$ (the method terminates when $\boldsymbol{d}_k = \boldsymbol{0}$). The convergence rate of such method for sharp weakly convex minimization has been analyzed in [14]. We plot the convergence behavior of the SubGM with the Polyak step size rule in Figure 2b, which also shows its linear convergence. However, we note that the Polyak step size rule is generally not easy to implement, as it requires the knowledge of $f^\star$.

Then, we consider the SubGM with piecewise geometrically diminishing step sizes, which dates as far back as to the work [39] and has recently been used in [54]. Specifically, we set $\mu_k = \frac{1}{2^{\lfloor k/N \rfloor}}$ with $N \in \{50, 100, 200\}$. Compared to the vanilla strategy $\mu_k = \rho^k \mu_0$, the piecewise strategy allows for a smaller decay rate ρ (here, we use $\rho = \frac{1}{2}$) and keeps the same step size for N iterations. As can be seen from Figure 2c, the method converges at a piecewise linear rate. Nevertheless, we observe that the piecewise strategy is slightly less efficient than the vanilla one in general.

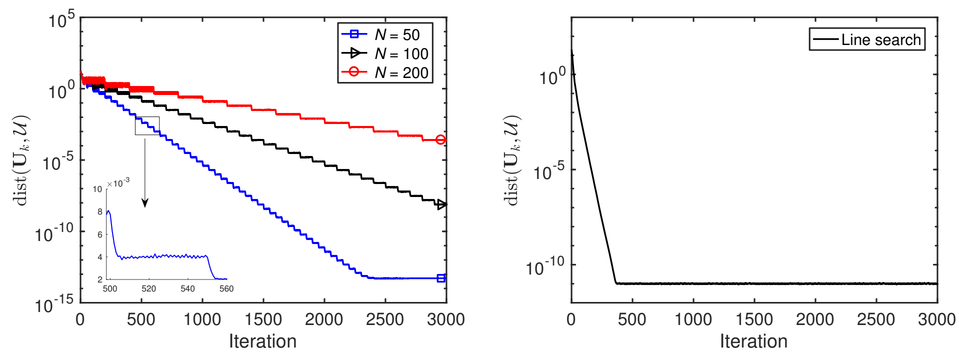
We also consider a modified backtracking line search strategy in [34] to choose the step size. Although such a strategy is generally designed for smooth problems, it is empirically used in [54] for a nonsmooth nonconvex optimization problem to achieve fast convergence. Inspired by the strategy of choosing geometrically diminishing step sizes, we modify the backtracking line search strategy in [34] by (i) setting $\mu_k = \mu_{k-1}$

and (ii) reducing it according to $\mu_k \leftarrow \mu_k\rho$ until the condition $f(\boldsymbol{U}_k - \mu_k \boldsymbol{d}_k) > f(\boldsymbol{U}_k) - \eta\mu_k\|\boldsymbol{d}_k\|$ is satisfied. We set $\eta = 10^{-3}$, $\rho = 0.85$, $\mu_0 = 1$ and plot the convergence behavior of the resulting method in Figure 2d. As can be seen from the figure, the method converges at a linear rate. Moreover, we observe empirically that the choice of parameters above works for other settings (i.e., different n, r, m, p). We leave the convergence analysis of the SubGM with backtracking line search as a future work.



(a) Distance of last iterate to optimal set with $\mu_0 \in \{0.1, 0.5, 1, 10\}$ and $\rho \in \{0.80, 0.81, \ldots, 0.99\}$

(b) Convergence of SubGM with geometrically diminishing ($\mu_k = \rho^k$, $\rho \in \{0.90, 0.93, 0.96, 0.99\}$) and Polyak step sizes

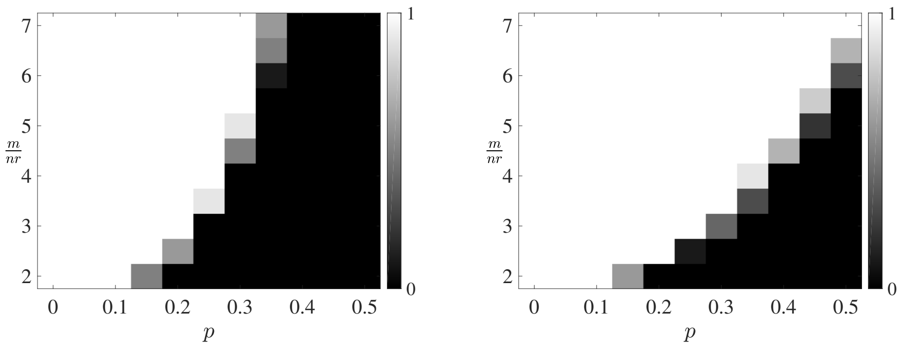


(c) Convergence of SubGM with piecewise geometrically diminishing ($\mu_k = \frac{1}{2^{\lfloor k/N \rfloor}}$, $N \in \{50, 100, 200\}$) step sizes

(d) Convergence of SubGM with modified backtracking line search ($\eta = 10^{-3}$, $\rho = 0.85$, $\mu_0 = 1$)

Fig. 2: Behavior of SubGM when applied to robust low-rank matrix recovery with $n = 50$, $r = 5$, $m = 5nr$, and $p = 0.3$.

Next, we study the performance of the SubGM with geometrically diminishing step sizes by varying the outlier ratio p and the number of measurements m. In these experiments we run the SubGM for 2×10^3 iterations with initial step size $\mu_0 = 1$ and decay rate $\rho = 0.99$. We also conduct experiments on the median-truncated gradient descent (MTGD) with the setting used in [30]. In particular, we initialize the MTGD with the truncated spectral method in Algorithm 3.1 and run it for 10^4 iterations. For each pair of p and m, 10 Monte Carlo trials are carried out, and for each trial we declare the recovery to be successful if the relative reconstruction error satisfies $\frac{\|\widehat{\boldsymbol{X}}-\boldsymbol{X}^\star\|_F}{\|\boldsymbol{X}^\star\|_F} \leq 10^{-6}$, where $\widehat{\boldsymbol{X}}$ is the reconstructed matrix. Figure 3 displays the phase transition of MTGD and SubGM using the average result of 10 independent trials. In this figure, white indicates successful recovery while black indicates failure. It is of interest to observe that when the outlier ratio p is small, both the SubGM and MTGD can exactly recover the underlying low-rank matrix $\boldsymbol{X}^\star$ even with only $m = 2nr$ measurements. On the other hand, given sufficiently large number of measurements (say $m = 7nr$), the SubGM is able to exactly recover the ground-truth matrix even when half of the measurements are corrupted by outliers, while the MTGD fails in this case. In particular, by comparing Figure 3a with Figure 3b, we observe that the SubGM is more robust to outliers than MTGD, especially in the case of high outlier ratio. We also observe from Figure 3 that with more measurements, the robust low-rank matrix recovery formulation (1.4) can tolerate not only more outliers but also a higher fraction of outliers. This provides further explanation to the observations made after the proof of Proposition 2.



(a) (b)

Fig. 3: Phase transition of robust low-rank matrix recovery using (a) median-truncated gradient descent (MTGD) [30] and (b) SubGM. Here, we fix $n = 50$, $r = 5$ and vary the outlier ratio p from 0 to 0.5. In addition, we vary m so that the ratio $\frac{m}{nr}$ varies from 2 to 7. Successful recovery is indicated by white and failure by black. Results are averaged over 10 independent trials.

6. Conclusion. In this paper we gave a nonsmooth nonconvex formulation of the problem of recovering a rank-r matrix $\boldsymbol{X}^\star \in \mathbb{R}^{n_1 \times n_2}$ from corrupted linear measurements. The formulation enforces the low-rank property of the solution by using a factored representation of the matrix variable and employs an ℓ_1-loss function to robustify the solution against outliers. We showed that even when close to half of the measurements are arbitrarily corrupted, as long as certain measurement operators arising from the measurement model satisfy the ℓ_1/ℓ_2-RIP, the formulation will be sharp and weakly convex. Consequently, the ground-truth matrix can be exactly recovered from any of its global minimum. Moreover, when suitably initialized, the SubGM with geometrically diminishing step sizes will converge to the ground-truth matrix at a linear rate.

7. Acknowledgment. We thank the Associate Editor and two anonymous reviewers for their detailed and helpful comments.

REFERENCES

[1] S. Aaronson, *The Learnability of Quantum States*, in Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 463, 2007, pp. 3089–3114.

[2] Y. Bai, Q. Jiang, and J. Sun, *Subgradient Descent Learns Orthogonal Dictionaries*, International Conference on Learning Representations (ICLR), (2019).

[3] D. P. Bertsekas, *Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization*, in Optimization for Machine Learning, S. Sra, S. Nowozin, and S. J. Wright, eds., Neural Information Processing Series, MIT Press, Cambridge, Massachusetts, 2012, pp. 85–119.

[4] S. Bhojanapalli, B. Neyshabur, and N. Srebro, *Global Optimality of Local Search for Low Rank Matrix Recovery*, in Advances in Neural Information Processing Systems 29 (NIPS), D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds., 2016, pp. 3873–3881.

[5] S. Burer and R. D. Monteiro, *A Nonlinear Programming Algorithm for Solving Semidefinite Programs via Low-Rank Factorization*, Mathematical Programming, 95 (2003), pp. 329–357.

[6] S. Burer and R. D. Monteiro, *Local Minima and Convergence in Low-Rank Semidefinite Programming*, Mathematical Programming, 103 (2005), pp. 427–444.

[7] J. V. Burke and M. C. Ferris, *Weak Sharp Minima in Mathematical Programming*, SIAM Journal on Control and Optimization, 31 (1993), pp. 1340–1359.

[8] E. J. Candès, X. Li, Y. Ma, and J. Wright, *Robust Principal Component Analysis?*, Journal of the ACM, 58 (2011), p. Article 11.

[9] E. J. Candès and Y. Plan, *Tight Oracle Inequalities for Low-Rank Matrix Recovery from a Minimal Number of Noisy Random Measurements*, IEEE Transactions on Information Theory, 57 (2011), pp. 2342–2359.

[10] E. J. Candès and B. Recht, *Exact Matrix Completion via Convex Optimization*, Foundations of Computational Mathematics, 9 (2009), pp. 717–772.

[11] Y. Chen, Y. Chi, and A. J. Goldsmith, *Exact and Stable Covariance Estimation from Quadratic Sampling via Convex Programming*, IEEE Transactions on Information Theory, 61 (2015), pp. 4034–4059.

[12] Y. Chi, Y. M. Lu, and Y. Chen, *Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview*, arXiv preprint arXiv:1809.09573, (2018).

[13] M. A. Davenport and J. Romberg, *An Overview of Low-Rank Matrix Recovery from Incomplete Observations*, IEEE Journal of Selected Topics in Signal Processing, 10 (2016), pp. 608–622.

[14] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette, *Subgradient Methods for Sharp Weakly Convex Functions*, Journal of Optimization Theory and Applications, 179 (2018), pp. 962–982.

[15] D. Davis, D. Drusvyatskiy, and C. Paquette, *The Nonsmooth Landscape of Phase Retrieval*, arXiv preprint arXiv:1711.03247, (2017).

[16] F. De La Torre and M. J. Black, *A Framework for Robust Subspace Learning*, International Journal of Computer Vision, 54 (2003), pp. 117–142.

[17] J. C. Duchi and F. Ruan, *Solving (Most) of a Set of Quadratic Equalities: Composite Optimization for Robust Phase Retrieval*, Information and Inference: A Journal of the IMA, (2018), p. iay015, https://doi.org/10.1093/imaiai/iay015.

[18] M. Fazel, H. Hindi, and S. Boyd, *Rank Minimization and Applications in System Theory*, in Proceedings of the 2004 American Control Conference, vol. 4, IEEE, 2004, pp. 3273–3278.

[19] R. Ge, J. D. Lee, and T. Ma, *Matrix Completion has No Spurious Local Minima*, in Advances in Neural Information Processing Systems, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds., 2016, pp. 2973–2981.

[20] J.-L. Goffin, *On Convergence Rates of Subgradient Optimization Methods*, Mathematical programming, 13 (1977), pp. 329–347.

[21] D. Gross, *Recovering Low-Rank Matrices from Few Coefficients in Any Basis*, IEEE Transactions on Information Theory, 57 (2011), pp. 1548–1566.

[22] Q. Gu, Z. W. Wang, and H. Liu, *Low-Rank and Sparse Structure Pursuit via Alternating Minimization*, in Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS 2016), 2016, pp. 600–609.

[23] B. Haeffele, E. Young, and R. Vidal, *Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing*, in Proceedings of the 31st International Conference on Machine Learning (ICML 2014), 2014, pp. 2007–2015.

[24] C. Josz, Y. Ouyang, R. Zhang, J. Lavaei, and S. Sojoudi, *A Theory on the Absence of Spurious Solutions for Nonconvex and Nonsmooth Optimization*, in Advances in Neural Information Processing Systems 31 (NeurIPS), S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., 2018, pp. 2441–2449.
[25] Q. Ke and T. Kanade, *Robust L_1 Norm Factorization in the Presence of Outliers and Missing Data by Alternative Convex Programming*, in Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, IEEE, 2005, pp. 739–746.
[26] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, *Statistical Modeling of Complex Backgrounds for Foreground Object Detection*, IEEE Transactions on Image Processing, 13 (2004), pp. 1459–1472.
[27] Q. Li, Z. Zhu, and G. Tang, *The Non-Convex Geometry of Low-Rank Matrix Optimization*, Information and Inference: A Journal of the IMA, (2018), p. iay003, https://doi.org/10.1093/imaiai/iay003.
[28] X. Li, J. Lu, R. Arora, J. Haupt, H. Liu, Z. Wang, and T. Zhao, *Symmetry, Saddle Points, and Global Optimization Landscape of Nonconvex Matrix Factorization*, IEEE Transactions on Information Theory, 65 (2019), pp. 3489–3514.
[29] X. Li, Z. Zhu, A. M.-C. So, and R. Vidal, *Nonconvex Robust Low-Rank Matrix Recovery*. Companion technical report, available at https://arxiv.org/abs/1809.09237, 2018.
[30] Y. Li, Y. Chi, H. Zhang, and Y. Liang, *Nonconvex Low-Rank Matrix Recovery with Arbitrary Outliers via Median-Truncated Gradient Descent*, Information and Inference: A Journal of the IMA, (2019), p. iaz009, https://doi.org/10.1093/imaiai/iaz009.
[31] Y. Li, Y. Sun, and Y. Chi, *Low-Rank Positive Semidefinite Matrix Recovery from Corrupted Rank-One Measurements*, IEEE Transactions on Signal Processing, 65 (2017), pp. 397–408.
[32] A. Nedić and D. Bertsekas, *Convergence Rate of Incremental Subgradient Algorithms*, in Stochastic Optimization: Algorithms and Applications, S. Uryasev and P. M. Pardalos, eds., vol. 54 of Applied Optimization, Springer Science+Business Media, Dordrecht, 2001.
[33] P. Netrapalli, U. N. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, *Non-Convex Robust PCA*, in Advances in Neural Information Processing Systems 27 (NIPS), Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds., 2014, pp. 1107–1115.
[34] J. Nocedal and S. Wright, *Numerical optimization*, Springer Science & Business Media, 2006.
[35] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, *Non-Square Matrix Sensing without Spurious Local Minima via the Burer-Monteiro Approach*, in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017), 2017, pp. 65–74.
[36] B. T. Polyak, *Minimization of Unsmooth Functions*, USSR Computational Mathematics and Mathematical Physics, 9 (1969), pp. 14–29.
[37] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317 of Grundlehren der mathematischen Wissenschaften, Springer-Verlag, Berlin Heidelberg, second ed., 2004.
[38] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, vol. 3 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin Heidelberg, 1985.
[39] N. Z. Shor and M. B. Shchepakin, *Algorithms for the Solution of the Two-Stage Problem in Stochastic Programming*, Kibernetika, 4 (1968), pp. 56–58.
[40] N. Srebro, J. Rennie, and T. S. Jaakkola, *Maximum-Margin Matrix Factorization*, in Advances in Neural Information Processing Systems 17 (NIPS), L. K. Saul, Y. Weiss, and L. Bottou, eds., 2004, pp. 1329–1336.
[41] R. Sun and Z.-Q. Luo, *Guaranteed Matrix Completion via Non–Convex Factorization*, IEEE Transactions on Information Theory, 62 (2016), pp. 6535–6579.
[42] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, *Low-Rank Solutions of Linear Matrix Equations via Procrustes Flow*, in Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), 2016, pp. 964–973.
[43] R. Vershynin, *Introduction to the Non-Asymptotic Analysis of Random Matrices*, in Compressed Sensing: Theory and Applications, Y. C. Eldar and G. Kutyniok, eds., Cambridge University Press, New York, 2012, pp. 210–268.
[44] J.-P. Vial, *Strong and Weak Convexity of Sets and Functions*, Mathematics of Operations Research, 8 (1983), pp. 231–259.
[45] X. Yi, D. Park, Y. Chen, and C. Caramanis, *Fast Algorithms for Robust PCA via Gradient Descent*, in Advances in Neural Information Processing Systems 29 (NIPS), D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds., 2016, pp. 4152–4160.
[46] M.-C. Yue and A. M.-C. So, *A Perturbation Inequality for Concave Functions of Singular*

Values and Its Applications in Low-Rank Matrix Recovery, Applied and Computational Harmonic Analysis, 40 (2016), pp. 396–416.

[47] M.-C. Yue, Z. Zhou, and A. M.-C. So, *On the Quadratic Convergence of the Cubic Regularization Method under a Local Error Bound Condition*, SIAM Journal on Optimization, 29 (2019), pp. 904–932.

[48] M. Zhang, Z.-H. Huang, and Y. Zhang, *Restricted p-Isometry Properties of Nonconvex Matrix Recovery*, IEEE Transactions on Information Theory, 59 (2013), pp. 4316–4323.

[49] X. Zhang, L. Wang, and Q. Gu, *A Unified Framework for Nonconvex Low-Rank plus Sparse Matrix Recovery*, in Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS 2018), 2018, pp. 1097–1107.

[50] Q. Zheng and J. Lafferty, *A Convergent Gradient Descent Algorithm for Rank Minimization and Semidefinite Programming from Random Linear Measurements*, in Advances in Neural Information Processing Systems 28 (NIPS), C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds., 2015, pp. 109–117.

[51] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, *The Global Optimization Geometry of Low-Rank Matrix Optimization*, arXiv preprint arXiv:1703.01256, (2017).

[52] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, *Global Optimality in Low-Rank Matrix Optimization*, IEEE Transactions on Signal Processing, 66 (2018), pp. 3614–3628.

[53] Z. Zhu, A. M.-C. So, and Y. Ye, *Fast and Near–Optimal Matrix Completion via Randomized Basis Pursuit*, in Fifth International Congress of Chinese Mathematicians, L. Ji, Y. S. Poon, L. Yang, and S.-T. Yau, eds., vol. 51, Part 2 of AMS/IP Studies in Advanced Mathematics, American Mathematical Society and International Press, 2012, pp. 859–882.

[54] Z. Zhu, Y. Wang, D. Robinson, D. Naiman, R. Vidal, and M. Tsakiris, *Dual Principal Component Pursuit: Improved Analysis and Efficient Algorithms*, in Advances in Neural Information Processing Systems 31 (NeurIPS), S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., 2018, pp. 2171–2181.

Appendix A. Proof of Proposition 1.

A.1. Preliminaries. We say that a random variable X is sub-Gaussian if $\Pr\left[|X| > t\right] \leq \exp\left(1 - \frac{t^2}{K_1^2}\right)$, $\forall\ t \geq 0$ for some constant $K_1 > 0$. This is equivalent to

$$(\mathrm{E}[|X|^p])^{1/p} \leq K_2\sqrt{p},\ \forall\ p \geq 1 \tag{A.1}$$

for some constant $K_2 > 0$. The constants K_1 and K_2 differ from each other by at most an absolute constant factor; see [43, Lemma 5.5]. The sub-Gaussian norm of a sub-Gaussian random variable X is defined as $\|X\|_{\psi_2} = \sup_{p\geq 1}\left\{p^{-1/2}\,\mathrm{E}[|X|^p]^{1/p}\right\}$. We then have the following Hoeffding-type inequalty:

Lemma 2 ([43, Proposition 5.10]). *Let $X_1, \ldots, X_m$ be independent sub-Gaussian random variables with $\mathrm{E}[X_i] = 0$ for $i = 1, \ldots, m$ and $K = \max_{i\in\{1,\ldots,m\}} \|X_i\|_{\psi_2}$. Then, for any $t > 0$, we have*

$$\Pr\left[\frac{1}{m}\left|\sum_{i=1}^m X_i\right| > t\right] \leq 2\exp\left(-\frac{cmt^2}{K^2}\right) \tag{A.2}$$

for some constant $c > 0$.

We also need the following result on the covering number of the set of low-rank matrices:

Lemma 3 ([9, Lemma 3.1]). *Let $\mathbb{S}_r = \{\boldsymbol{X} \in \mathbb{R}^{n\times n} : \|\boldsymbol{X}\|_F = 1, \operatorname{rank}(\boldsymbol{X}) \leq r\}$. Then, there exists an ϵ-net $\overline{\mathbb{S}}_{r,\epsilon} \subset \mathbb{S}_r$ with respect to the Frobenius norm (i.e., for any $\boldsymbol{X} \in \mathbb{S}_r$, there exists an $\overline{\boldsymbol{X}} \in \overline{\mathbb{S}}_{r,\epsilon}$ such that $\|\boldsymbol{X} - \overline{\boldsymbol{X}}\|_F \leq \epsilon$) satisfying $|\overline{\mathbb{S}}_{r,\epsilon}| \leq \left(\frac{9}{\epsilon}\right)^{(2n+1)r}$.*

A.2. Isometry Property of a Given Matrix.

LEMMA 4. *Suppose that the matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_m \in \mathbb{R}^{n\times n}$ defining the linear measurement operator $\mathcal{A}$ have* i.i.d. *standard Gaussian entries. Then, for any $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ and $0 < \delta < 1$, there exists a constant $c_1 > 0$ such that with probability exceeding $1 - 2\exp(-c_1\delta^2 m)$, we have*

$$\left(\sqrt{\frac{2}{\pi}} - \delta\right) \|\boldsymbol{X}\|_F \le \frac{1}{m}\|\mathcal{A}(\boldsymbol{X})\|_1 \le \left(\sqrt{\frac{2}{\pi}} + \delta\right) \|\boldsymbol{X}\|_F. \tag{A.3}$$

Proof of Lemma 4. Since $\boldsymbol{A}_i$ has *i.i.d.* standard Gaussian entries, the random variable $\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle$ is Gaussian with mean zero and variance $\|\boldsymbol{X}\|_F^2$. It follows that

$$\mathrm{E}[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle|] = \sqrt{\frac{2}{\pi}}\|\boldsymbol{X}\|_F, \quad \mathrm{E}[\|\mathcal{A}(\boldsymbol{X})\|_1] = m\sqrt{\frac{2}{\pi}}\|\boldsymbol{X}\|_F. \tag{A.4}$$

Now, let $Z_i = |\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle| - \mathrm{E}[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle|]$, which satisfies $\mathrm{E}[Z_i] = 0$. We claim that Z_i is a sub-Gaussian random variable. To establish the claim, it suffices to bound the sub-Gaussian norm of Z_i. Towards that end, we first observe that $\Pr[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle| > t] \le 2\exp\left(-\frac{t^2}{2\|\boldsymbol{X}\|_F^2}\right)$. Together with (A.4), this implies that for any $t > \mathrm{E}[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle|]$,

$$\begin{aligned}
\Pr[|Z_i| > t] &= \Pr[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle| > t + \mathrm{E}[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle|]] + \Pr[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle| < -t + \mathrm{E}[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle|]] \\
&\le 2\exp\left(-\frac{(t + \mathrm{E}[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle|])^2}{2\|\boldsymbol{X}\|_F^2}\right) + \Pr[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle| < -t + \mathrm{E}[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle|]] \\
&\le 2\exp\left(-\frac{(t + \mathrm{E}[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle|])^2}{2\|\boldsymbol{X}\|_F^2}\right) \le \exp\left(1 - \frac{t^2}{\|\boldsymbol{X}\|_F^2}\right),
\end{aligned}$$

where the second inequality follows because $\Pr[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle| < -t + \mathrm{E}[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle|]] = 0$ for all $t > \mathrm{E}[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle|]$. Since $\exp\left(1 - \frac{t^2}{\|\boldsymbol{X}\|_F^2}\right) \ge 1$ for all $t \le E[|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle|] = \sqrt{\frac{2}{\pi}}\|\boldsymbol{X}\|_F$, we then have $\Pr[|Z_i| > t] \le \exp\left(1 - \frac{t^2}{\|\boldsymbol{X}\|_F^2}\right)$, $\forall t \ge 0$. This, together with (A.1), implies that $(\mathrm{E}[|Z_i|^p])^{1/p} \le cp^{1/2}\|\boldsymbol{X}\|_F$, $\forall\, p \ge 1$, where $c > 0$ is a constant. It follows that $\|Z_i\|_{\psi_2} \le c\|\boldsymbol{X}\|_F$; i.e., Z_i is a sub-Gaussian random variable, as desired.

Now, applying the Hoeffding-type inequality in Lemma 2 with $t = \delta\|\boldsymbol{X}\|_F$ and $K = c\|\boldsymbol{X}\|_F$ gives

$$\Pr\left[\frac{1}{m}\left|\|\mathcal{A}(\boldsymbol{X})\|_1 - \mathrm{E}[\|\mathcal{A}(\boldsymbol{X})\|_1]\right| > \delta\|\boldsymbol{X}\|_F\right] \le 2\exp(-c_1 m\delta^2)$$

for some constant $c_1 > 0$. Using (A.4), we conclude that (A.3) holds with probability at least $1 - 2\exp(-c_1 m\delta^2)$. This completes the proof. □

A.3. Proof of Proposition 1. We now utilize an ϵ-net argument to show that (A.3) holds for all rank-r matrices with high probability as long as $m \gtrsim nr$. Since the inequality (A.3) is scale invariant, without loss of generality, we may assume that $\|\boldsymbol{X}\|_F = 1$ and focus on the set $\mathbb{S}_r$ defined in Lemma 3.

Proof of Proposition 1. We begin by showing that (A.3) holds for all $\boldsymbol{X} \in \overline{\mathbb{S}}_{r,\epsilon}$ with high probability. Indeed, upon setting $\epsilon = \frac{\delta\sqrt{\pi}}{16}$ in Lemma 3 and utilizing a union bound together with Lemma 4, we have

$$\text{(A.5)}\quad \begin{aligned} &\Pr\left[\max_{\overline{\boldsymbol{X}}\in\overline{\mathbb{S}}_{r,\epsilon}} \frac{1}{m}\left|\|\mathcal{A}(\overline{\boldsymbol{X}})\|_1 - m\sqrt{\frac{2}{\pi}}\|\overline{\boldsymbol{X}}\|_F\right| \geq \frac{\delta}{2}\right] \leq 2|\overline{\mathbb{S}}_{r,\epsilon}|\exp(-c_1 m\delta^2) \\ &\leq 2\left(\frac{9}{\epsilon}\right)^{(2n+1)r} \exp(-c_1 m\delta^2) \leq \exp(-c_2 m\delta^2) \end{aligned}$$

whenever $m \gtrsim nr$.

Next, we show that (A.3) holds for all $\boldsymbol{X} \in \mathbb{S}_r$. Towards that end, set

$$\text{(A.6)}\quad \kappa_r = \frac{1}{m}\sup_{\boldsymbol{X}\in\mathbb{S}_r} \|\mathcal{A}(\boldsymbol{X})\|_1$$

and let $\boldsymbol{X} \in \mathbb{S}_r$ be *arbitrary*. Then, there exists an $\overline{\boldsymbol{X}} \in \overline{\mathbb{S}}_{r,\epsilon}$ such that $\|\boldsymbol{X} - \overline{\boldsymbol{X}}\|_F \leq \epsilon$. It follows from (A.5) that with high probability,

$$\text{(A.7)}\quad \begin{aligned} \frac{1}{m}\|\mathcal{A}(\boldsymbol{X})\|_1 &= \frac{1}{m}\|\mathcal{A}(\boldsymbol{X} - \overline{\boldsymbol{X}}) + \mathcal{A}(\overline{\boldsymbol{X}})\|_1 \leq \frac{1}{m}\|\mathcal{A}(\boldsymbol{X} - \overline{\boldsymbol{X}})\|_1 + \frac{1}{m}\|\mathcal{A}(\overline{\boldsymbol{X}})\|_1 \\ &\leq \frac{1}{m}\|\mathcal{A}(\boldsymbol{X} - \overline{\boldsymbol{X}})\|_1 + \sqrt{\frac{2}{\pi}} + \frac{\delta}{2}. \end{aligned}$$

Noting that $\boldsymbol{X} - \overline{\boldsymbol{X}}$ has rank at most $2r$, we can decompose it as $\boldsymbol{X} - \overline{\boldsymbol{X}} = \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2$, where $\langle \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2 \rangle = 0$ and $\operatorname{rank}(\boldsymbol{\Delta}_1), \operatorname{rank}(\boldsymbol{\Delta}_2) \leq r$ (this follows essentially from the SVD). Hence, we can compute

$$\begin{aligned} &\frac{1}{m}\|\mathcal{A}(\boldsymbol{X} - \overline{\boldsymbol{X}})\|_1 \leq \frac{1}{m}[\|\mathcal{A}(\boldsymbol{\Delta}_1)\|_1 + \|\mathcal{A}(\boldsymbol{\Delta}_2)\|_1] \\ &= \frac{1}{m}[\|\boldsymbol{\Delta}_1\|_F\|\mathcal{A}(\boldsymbol{\Delta}_1/\|\boldsymbol{\Delta}_1\|_F)\|_1 + \|\boldsymbol{\Delta}_2\|_F\|\mathcal{A}(\boldsymbol{\Delta}_2/\|\boldsymbol{\Delta}_2\|_F)\|_1] \\ &\leq \kappa_r(\|\boldsymbol{\Delta}_1\|_F + \|\boldsymbol{\Delta}_2\|_F) \leq \sqrt{2}\kappa_r\epsilon, \end{aligned}$$

where the last inequality is due to $\|\boldsymbol{\Delta}_1\|_F^2 + \|\boldsymbol{\Delta}_2\|_F^2 = \|\boldsymbol{X} - \overline{\boldsymbol{X}}\|_F^2 \leq \epsilon^2$. This, together with (A.7), gives

$$\text{(A.8)}\quad \frac{1}{m}\|\mathcal{A}(\boldsymbol{X})\|_1 \leq \sqrt{\frac{2}{\pi}} + \frac{\delta}{2} + \sqrt{2}\kappa_r\epsilon.$$

In particular, using the definition of κ_r in (A.6), we obtain $\kappa_r \leq \sqrt{\frac{2}{\pi}} + \frac{\delta}{2} + \sqrt{2}\kappa_r\epsilon$, or equivalently, $\kappa_r \leq \frac{\sqrt{2/\pi}+\delta/2}{1-\sqrt{2}\epsilon}$. Plugging in our choice of ϵ yields $\sqrt{2}\kappa_r\epsilon \leq \frac{\delta}{2}$. This, together with (A.8) and the fact that $\|\boldsymbol{X}\|_F = 1$, implies

$$\frac{1}{m}\|\mathcal{A}(\boldsymbol{X})\|_1 \leq \left(\sqrt{\frac{2}{\pi}} + \delta\right)\|\boldsymbol{X}\|_F.$$

Similarly, using (A.5), we have

$$\begin{aligned} \frac{1}{m}\|\mathcal{A}(\boldsymbol{X})\|_1 &\geq \frac{1}{m}\|\mathcal{A}(\overline{\boldsymbol{X}})\|_1 - \frac{1}{m}\|\mathcal{A}(\boldsymbol{X} - \overline{\boldsymbol{X}})\|_1 \\ &\geq \sqrt{\frac{2}{\pi}} - \frac{\delta}{2} - \frac{1}{m}\|\mathcal{A}(\boldsymbol{X} - \overline{\boldsymbol{X}})\|_1 \\ &\geq \sqrt{\frac{2}{\pi}} - \frac{\delta}{2} - \sqrt{2}\kappa_r\epsilon \geq \sqrt{\frac{2}{\pi}} - \delta = \left(\sqrt{\frac{2}{\pi}} - \delta\right)\|\boldsymbol{X}\|_F \end{aligned}$$

with high probability. This completes the proof. ∎