

Guarantees of a Preconditioned Subgradient Algorithm for Overparameterized Asymmetric Low-rank Matrix Recovery

Paris Giampouras¹ HanQin Cai² René Vidal³

Abstract

In this paper, we focus on a matrix factorization-based approach to recover low-rank *asymmetric* matrices from corrupted measurements. We propose an *Overparameterized Preconditioned Subgradient Algorithm (OPSA)* and provide, for the first time in the literature, linear convergence rates independent of the rank of the sought asymmetric matrix in the presence of gross corruptions. Our work goes beyond existing results in preconditioned-type approaches addressing their current limitation, i.e., the lack of convergence guarantees in the case of *asymmetric matrices of unknown rank*. By applying our approach to (robust) matrix sensing, we highlight its merits when the measurement operator satisfies a mixed-norm restricted isometry property. Lastly, we present extensive numerical experiments that validate our theoretical results and demonstrate the effectiveness of our approach for different levels of overparameterization and outlier corruptions.

1. Introduction

Low-rank matrix recovery has been a ubiquitous problem showing up in numerous applications in the fields of signal/image processing, machine learning, and data science (Recht et al., 2010; Chen et al., 2013; Davenport & Romberg, 2016; Cai et al., 2021a; Smith et al., 2024; Wang et al., 2024). For instance, problems such as matrix sensing (Jain et al., 2013; Wei et al., 2016; Park et al., 2017; Li et al., 2020), matrix completion (Candes & Recht, 2008; Nie et al., 2012;

Kümmerle & Verdun, 2021; Cai et al., 2023), and robust principal component analysis (Candès et al., 2011; Netrapalli et al., 2014; Giampouras et al., 2018; Cai et al., 2019; 2021b), can all be cast as low-rank matrix recovery problems and then solved using minimization algorithms that seek a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ that is as close as possible to the unknown low-rank matrix $\mathbf{X}_* \in \mathbb{R}^{m \times n}$.

A major challenge in low-rank matrix recovery concerns the computational complexity and memory requirements of the proposed algorithms when the size of the unknown matrix significantly increases. To address this, matrix factorization-based approaches have been proposed, (Chi et al., 2019; Park et al., 2016), which, given $r \geq \text{rank}(\mathbf{X}_*)$, use matrices $\mathbf{L} \in \mathbb{R}^{m \times r}$ and $\mathbf{R} \in \mathbb{R}^{n \times r}$ such that $\mathbf{X} = \mathbf{L}\mathbf{R}^\top$. These approaches enable the development of iterative algorithms with significantly reduced computational complexity and memory storage requirements. However, they lead to non-convex formulations of the optimization problems, which pose their own challenges in the derivation of theoretical guarantees and fast rates of convergence.

Alternating gradient-based algorithms have been a standard approach for solving matrix factorization-based problems. A well-known issue with these approaches lies in the dependence of their rate of convergence on the condition number of the unknown matrix $\mathbf{X}_*$ (Tong et al., 2021a). This makes convergence too slow in problems where the sought matrix is ill-conditioned. Several works, (Park et al., 2016; Zhang et al., 2023), address this issue by moving beyond vanilla gradient updates and resorting to *preconditioned* approaches. In (Tong et al., 2021a), preconditioned gradient algorithms have been proposed that consist of updates in the following form:

$$\begin{aligned} \mathbf{L}_{t+1} &= \mathbf{L}_t - \eta \nabla_{\mathbf{L}} \mathcal{L}(\mathbf{L}_t \mathbf{R}_t^\top) \underbrace{(\mathbf{R}_t^\top \mathbf{R}_t)^{-1}}_{\text{preconditioner}}, \\ \mathbf{R}_{t+1} &= \mathbf{R}_t - \eta \nabla_{\mathbf{R}} \mathcal{L}(\mathbf{L}_t \mathbf{R}_t^\top) \underbrace{(\mathbf{L}_t^\top \mathbf{L}_t)^{-1}}_{\text{preconditioner}}. \end{aligned} \quad (1)$$

Preconditioned-based approaches, e.g. (Chi et al., 2019; Cai et al., 2021c; Zhang et al., 2023; 2024; Cai et al., 2024), lead to linear rates of convergence that do not depend on the condition number of $\mathbf{X}_*$. Recently, these approaches have been extended to the *robust low-rank matrix recovery*

¹Department of Computer Science, University of Warwick, Coventry, England, UK ²Department of Statistics and Data Science and Department of Computer Science, University of Central Florida, Orlando, FL, USA ³Center for Innovation in Data Engineering and Science (IDEAS), Departments of ESE and Radiology, University of Pennsylvania, Philadelphia, PA, USA. Correspondence to: René Vidal <vidalr@seas.upenn.edu>.

framework, aiming to address more challenging scenarios that account for the presence of grossly corrupted data. In these settings, (Tong et al., 2021b) reported similar rates of convergence, which are however negatively influenced as the level of corruption increases.

A fundamental assumption made in the above-mentioned approaches is that the rank r of the sought matrix $\mathbf{X}_*$ is known. Clearly, in the case that r is underestimated, the algorithm will only be able to find a low-rank approximation of the ground truth matrix. To address this shortcoming, in (Ma & Fattahi, 2021), the authors overparameterized the rank as $d \geq r = \text{rank}(\mathbf{X}_*)$, and established a linear rate of convergence for robust low-rank matrix recovery for *symmetric* matrices using a vanilla subgradient algorithm. Further works, build on this observation aiming to establish convergence to the ground truth matrix $\mathbf{X}_*$ under weaker conditions, (Ding et al., 2021; Xiong et al., 2023), or to explore the intriguing implicit regularization phenomena in this setting, (Soltanolkotabi et al., 2023).

Note that due to non-invertibility issues as the iterates $\mathbf{L}_t, \mathbf{R}_t$ converge to $\mathbf{L}_*, \mathbf{R}_*$, the overparameterized scenario cannot be directly adopted as such in the previously proposed preconditioned-based methods that rely on updates given in (1). To address this issue, recent works, (Zhang et al., 2023; Xu et al., 2023), proposed an overparameterized preconditioned algorithm focusing on symmetric matrices and assuming smooth loss functions, with updates in the form:

$$\mathbf{L}_{t+1} = \mathbf{L}_t - \eta \nabla_{\mathbf{L}} \mathcal{L}(\mathbf{L}_t \mathbf{L}_t^\top) (\mathbf{L}_t^\top \mathbf{L}_t + \lambda \mathbf{I})^{-1}, \quad (2)$$

and showed linear convergence at a rate independent of the condition number of $\mathbf{X}_*$ and the overparameterization of the true rank. Focusing again on *smooth* problems, in (Cheng & Zhao, 2024), the authors proposed an extension of these works to the case of asymmetric matrices coming up with an alternating algorithm with regularized preconditioners leading to updates:

$$\begin{aligned} \mathbf{L}_{t+1} &= \mathbf{L}_t - \eta \nabla_{\mathbf{L}} \mathcal{L}(\mathbf{L}_t \mathbf{R}_t^\top) (\mathbf{R}_t^\top \mathbf{R}_t + \lambda_t \mathbf{I})^{-1}, \\ \mathbf{R}_{t+1} &= \mathbf{R}_t - \eta \nabla_{\mathbf{R}} \mathcal{L}(\mathbf{L}_t \mathbf{R}_t^\top) (\mathbf{L}_t^\top \mathbf{L}_t + \lambda_t \mathbf{I})^{-1}. \end{aligned} \quad (3)$$

In this work, we depart from previous works by focusing on *robust* low-rank matrix recovery in the presence of outliers and propose an overparameterized preconditioned-based algorithm in the unknown rank regime. Our work aims to address the following question (Q):

Q: Can we establish linear rates of convergence to the ground truth $\mathbf{X}_$, in the case of **non-smooth** minimization problems in the overparameterized regime with **unknown rank**, and for **asymmetric** matrices $\mathbf{X}_*$?*

1.1. Main Contributions

In this work, we advance beyond previous work by focusing on robust low-rank matrix recovery with a non-smooth objective function, addressing the unique challenges of recovering *asymmetric matrices with unknown rank* (see Table 1 and comparison with SOTA). Our preconditioners naturally arise by adopting quasi-Newton-type updates in an implicitly regularized objective function. Our main contributions are summarized as follows:

- We propose a novel algorithm, coined *Overparameterized Preconditioned Subgradient Algorithm (OPSA)*, that minimizes a robust ℓ_1 loss function. To account for overparameterization caused by rank overestimation, we propose a novel distance metric and assume that the matrix factors are initialized sufficiently close to the ground truth (which can be easily attained by spectral initialization). In Theorem 5.4, we show that *OPSA converges linearly* to the low-rank ground truth matrix $\mathbf{X}_*$ using an adaptive Polyak’s step size. Note that our main result holds for general non-smooth loss functions under certain conditions such as the restricted rank- d sharpness condition and restricted Lipschitz continuity. Moreover, our results extend the preconditioned subgradient method (Tong et al., 2021b) from the exact known rank setting to the overparameterized regime.
- For theoretical results, we focus on robust matrix sensing and show that linear convergence holds for OPSA both in the noiseless case and in the presence of gross corruptions/outliers when the measurement matrices satisfy a mixed-norm restricted isometry property (RIP). In this setting, we unveil how the iteration complexity is affected by overparameterization. Moreover, our results showcase that the tolerance of measurement matrices in outliers is another important factor for sharpness around $\mathbf{X}_*$, which is a necessary condition for exact convergence.
- In the experimental section, we empirically showcase the favorable performance of the proposed OPSA against the state-of-the-art under different levels of overparameterization d , for the problem of robust matrix sensing with Gaussian measurements. We also demonstrate that OPSA constantly enjoys linear convergence with varying condition numbers κ , parameters λ , and outlier densities even when the rank is heavily overestimated. The experiments provided further confirm our theoretical findings.

1.2. Notation

The transpose of a vector or matrix is denoted as $(\cdot)^\top$. The Euclidean vector norm is denoted as $\|\cdot\|_2$. The Frobenius and operator matrix norms are denoted as $\|\cdot\|_F$ and $\|\cdot\|_{op}$,

Table 1. Comparison of theoretical convergence properties of SOTA algorithms for low-rank matrix estimation. Columns indicate if algorithms handle asymmetric matrices, outliers, unknown rank, and if convergence is independent of $\kappa(\mathbf{X}_*)$, the condition number of $\mathbf{X}_*$. The proposed OPSA method addresses all these challenges.

Algorithm	Asymmetric	Outliers	Unknown Rank	No dependency on $\kappa(\mathbf{X}_*)$
VANILLA GD (Stöger & Soltanolkotabi, 2021)	✓	✓	✓	✗
SCALEDGD (Tong et al., 2021a)	✓	✗	✗	✓
PRECEDGD (Zhang et al., 2023)	✗	✗	✓	✓
SCALEDGD(λ) (Xiong et al., 2023)	✗	✗	✓	✓
SCALEDSDM (Tong et al., 2021b)	✓	✓	✗	✓
OPSA (Proposed)	✓	✓	✓	✓

respectively. We also denote as $\sigma_i(\mathbf{X})$ the singular values of matrix $\mathbf{X}$ -assuming a decreasing order as i increases- and as $\kappa(\mathbf{X})$ its condition number. We denote the trace of a matrix as $\text{tr}(\cdot)$, the trace of the inner product between two matrices $\mathbf{A}, \mathbf{B}$ as $\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^\top \mathbf{B})$, and the $m \times n$ -dimensional Euclidean space as $\mathbb{R}^{m \times n}$. $G(d)$ denotes the set of invertible matrices in $\mathbb{R}^{d \times d}$.

2. Related Work

In this section, we provide some interesting insights into the connection of our approach with prior research works.

Preconditioned gradient and subgradient methods for low-rank matrix recovery. Preconditioned-based methods have attracted significant interest over the last few years since they allow for establishing rates of convergence that do not depend on the condition number of $\mathbf{X}_*$, (Tong et al., 2021a; Zhang, 2021; Zhang et al., 2023). For a thorough review of these methods, we refer the reader to (Chi et al., 2019). Preconditioned methods have been extended to non-smooth problems such as robust low-rank matrix recovery with ℓ_1 loss. However, they focus on either the known-rank asymmetric regime or assume an unknown-rank with sought symmetric matrix $\mathbf{X}$. Relaxing the symmetric assumption on $\mathbf{X}$ to the more challenging asymmetric one in unknown-rank regimes is the main contribution of our work. Recently, these approaches have been shown to offer significant improvements in the low-rank adaptation (LoRA) for parameter-efficient fine-tuning foundation models, (Zhang & Pilanci, 2024). Even though this problem is out of the scope of the current paper, extending current approaches, which rely on a fixed rank, to the overparameterized preconditioning framework is a promising future research direction.

Overparameterized (robust) low-rank matrix recovery. Recently, several works have focused on robust low-rank matrix recovery in the unknown rank regime. In (Ma & Fattahi, 2021), the authors focus on robust matrix sensing and report the convergence of a vanilla subgradient algorithm in the overparameterized setting for symmetric

matrices, which suggests an implicit regularization behavior. In (Ding et al., 2021), improved results are obtained, again for the symmetric case, by relaxing the conditions imposed on measurement matrices. In (Zhang et al., 2023), with the aim to reduce the negative effect of overparameterization and ill-conditioning, the authors focused on symmetric matrices and generalized the preconditioned-based approach in the overparameterized setting using updates in the form of (2). Similar to our work, the authors in (Cheng & Zhao, 2024) recently, proposed an overparameterized preconditioned approach for asymmetric matrix factorization establishing linear convergence with update in the form of (3). However, unlike our work, they focused on smooth losses, which pose less challenges, and enabled them to use a Polyak-Lojasiewicz (PL)-type condition for deriving the convergence rate.

3. Problem Formulation

We focus on the low-rank matrix estimation problem, assuming that the true rank r is unknown. We denote the ground truth matrix as $\mathbf{X}_*$, and assume a singular value decomposition

$$\mathbf{X}_* = \mathbf{U}_* \mathbf{\Sigma}_* \mathbf{V}_*^\top, \quad (4)$$

where $\mathbf{U}_* \in \mathbb{R}^{m \times d}$ contains $d \geq r$ left singular vectors, $\mathbf{\Sigma}_* \in \mathbb{R}^{d \times d}$ is a diagonal matrix consisting of d singular values of $\mathbf{X}_*$ presented in an non-ascending order. Since $\text{rank}(\mathbf{X}_*) = r$ and $d \geq r$ we have $\sigma_i(\mathbf{X}_*) = 0$ for $i = r + 1, \dots, d$.

The low-rank matrix estimation problem w.r.t. the space of $\mathbf{X} \in \mathbb{R}^{m \times n}$ is defined as

$$\min_{\substack{\mathbf{X} \in \mathbb{R}^{m \times n} \\ \text{rank}(\mathbf{X}) \leq d}} \mathcal{L}(\mathbf{X}), \quad (5)$$

where $\mathcal{L}(\mathbf{X})$ is a general loss function that is convex w.r.t. $\mathbf{X}$ and possibly non-smooth in order to allow the use of robust loss functions such as the ℓ_1 norm.

Here, we solve a problem equivalent to problem (5), defined

over matrix factors $\mathbf{L} \in \mathbb{R}^{m \times d}$ and $\mathbf{R} \in \mathbb{R}^{n \times d}$,

$$\min_{\mathbf{L} \in \mathbb{R}^{m \times d}, \mathbf{R} \in \mathbb{R}^{n \times d}} \mathcal{L}(\mathbf{L}\mathbf{R}^\top). \quad (6)$$

3.1. Matrix Sensing

Next, we focus on the *matrix sensing* problem i.e., we assume that we have access to observations $\mathbf{y} = \{y_i\}_{i=1}^p$ of a low-rank matrix $\mathbf{X}_\star \in \mathbb{R}^{m \times n}$, given as

$$y_i = \mathcal{A}_i(\mathbf{X}_\star) + s_i, \quad 1 \leq i \leq p, \quad (7)$$

where $\mathcal{A}_i$ is the measurement operator,

$$\mathcal{A}_i(\mathbf{X}_\star) = \frac{1}{p} \langle \mathbf{A}_i \mathbf{X}_\star \rangle,$$

$\mathbf{A}_i \in \mathbb{R}^{m \times n}$ is the i -th measurement matrix, and s_i 's correspond to arbitrary and sparse corruptions.

The observation model above can be written in a more compact form as

$$\mathbf{y} = \mathcal{A}(\mathbf{X}_\star) + \mathbf{s}. \quad (8)$$

Our goal is to find $\mathbf{X}_\star$ given $\mathbf{y}$ and the measurement ensemble $\mathcal{A}(\cdot) = \{\mathcal{A}_i(\cdot)\}_{i=1}^p$. We formulate the problem as

$$\min_{\mathbf{X} \in \mathbb{R}^{m \times n}} \|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_1, \quad (9)$$

where we have used the ℓ_1 norm as the loss term $l(\cdot)$, also known as the residual sum of absolute errors, which is known to be robust to the presence of arbitrary sparse corruptions (Candès et al., 2011).

Recall that we assume we do not know the true rank r of the unknown matrix $\mathbf{X}_\star$, and we also solve the problem in the space of matrix factors $\mathbf{L} \in \mathbb{R}^{m \times d}$ and $\mathbf{R} \in \mathbb{R}^{n \times d}$ whose product equals $\mathbf{X}$ i.e., $\mathbf{X} = \mathbf{L}\mathbf{R}^\top$ with d being an overestimate of r . We thus formulate matrix sensing as

$$\min_{\mathbf{L} \in \mathbb{R}^{m \times d}, \mathbf{R} \in \mathbb{R}^{n \times d}} \|\mathbf{y} - \mathcal{A}(\mathbf{L}\mathbf{R}^\top)\|_1. \quad (10)$$

4. Overparameterized Preconditioned Subgradient Algorithm

To minimize the objective function given in (6), we use quasi-Newton type updates. Hence, we use local upper-bounds of the objective function, which lead to preconditioned-type updates for the matrix factors $\mathbf{L}$ and $\mathbf{R}$, i.e.,

$$\begin{aligned} (\mathbf{L}_{t+1}, \mathbf{R}_{t+1}) \equiv \arg \min_{\mathbf{L}, \mathbf{R}} & \mathcal{L}(\mathbf{L}_t \mathbf{R}_t^\top) + \langle \partial_{\mathbf{L}} \mathcal{L}(\mathbf{L}_t \mathbf{R}_t^\top), \mathbf{L} - \mathbf{L}_t \rangle \\ & + \langle \partial_{\mathbf{R}} \mathcal{L}(\mathbf{L}_t \mathbf{R}_t^\top), (\mathbf{R} - \mathbf{R}_t) \rangle \\ & + \frac{1}{2\eta_t} \left(\|\mathbf{L} - \mathbf{L}_t\|_F^2 + \|\mathbf{R}_t^\top \mathbf{R}_t + \lambda \mathbf{I}\|_F^2 \right. \\ & \left. + \|\mathbf{R} - \mathbf{R}_t\|_F^2 + \|\mathbf{L}_t^\top \mathbf{L}_t + \lambda \mathbf{I}\|_F^2 \right), \end{aligned}$$

where η_t is the step size and $\partial_{\mathbf{L}} \mathcal{L}(\mathbf{L}_t \mathbf{R}_t^\top)$ and $\partial_{\mathbf{R}} \mathcal{L}(\mathbf{L}_t \mathbf{R}_t^\top)$ denote subgradients of the objective function $\mathcal{L}(\mathbf{L}_t \mathbf{R}_t^\top)$ w.r.t. $\mathbf{L}$ and $\mathbf{R}$, respectively.

Note that the RHS of the above optimization problem corresponds to upper bounds of the original objective and leads to quasi-Newton-type updates, (Giampouras et al., 2020). We use a Polyak's type step size, and get a similar form to the one in (Tong et al., 2021b), i.e.,

$$\eta_t = \frac{\mathcal{L}(\mathbf{L}_t \mathbf{R}_t^\top) - \mathcal{L}(\mathbf{L}_\star \mathbf{R}_\star^\top)}{\gamma_t}, \quad (11)$$

where

$$\begin{aligned} \gamma_t = & \|\mathbf{S}_t \mathbf{R}_t (\mathbf{R}_t^\top \mathbf{R}_t + \lambda \mathbf{I})^{-\frac{1}{2}}\|_F^2 \\ & + \|\mathbf{S}_t^\top \mathbf{L}_t (\mathbf{L}_t^\top \mathbf{L}_t + \lambda \mathbf{I})^{-\frac{1}{2}}\|_F^2 \end{aligned}$$

with $\mathbf{S}_t$ denoting a subgradient of the objective $\mathcal{L}(\mathbf{X}_t)$, i.e., $\mathbf{S}_t \in \partial_{\mathbf{X}} \mathcal{L}(\mathbf{X}_t)$.

The proposed algorithm is given in Algorithm 1.

Algorithm 1 Overparameterized Preconditioned Subgradient Algorithm (OPSA)

- 1: **Input:** Data ($\mathbf{y} \in \mathbb{R}^p$ in the matrix sensing case), d : overestimated rank, η : stepsize.
 - 2: **Initialize** $\mathbf{L}_0$ and $\mathbf{R}_0$, set $t = 0$.
 - 3: **while** ! Stop Condition **do**
 - 4: $\mathbf{L}_{t+1} = \mathbf{L}_t - \eta_t \mathbf{S}_t \mathbf{R}_t (\mathbf{R}_t^\top \mathbf{R}_t + \lambda \mathbf{I})^{-1}$
 - 5: $\mathbf{R}_{t+1} = \mathbf{R}_t - \eta_t \mathbf{S}_t^\top \mathbf{L}_t (\mathbf{L}_t^\top \mathbf{L}_t + \lambda \mathbf{I})^{-1}$
 - 6: $t = t + 1$
 - 7: **end while**
 - 8: **Output:** $\hat{\mathbf{X}} = \mathbf{L}_t \mathbf{R}_t^\top$.
-

Remark 4.1. In practice, if $\mathcal{L}(\mathbf{L}_\star \mathbf{R}_\star^\top)$ is unknown, e.g., due to noise or additional regularization terms, Polyak's type step size may be hard to apply. To address that, one can use a geometrically decaying step size schedule to match the expected linear convergence. Such a step size schedule was introduced in (Goffin, 1977) and has been widely used in the literature on subgradient methods.

5. Convergence Analysis

In this section, we present the convergence analysis of the proposed Overparameterized Preconditioned Subgradient Algorithm (OPSA).

5.1. Landscape Assumptions

In the following, we provide the assumptions used in our theoretical results.

Assumption 5.1 (Restricted Lipschitz Continuity). A function $\mathcal{L}(\cdot) : \mathbb{R}^{m \times n} \mapsto \mathbb{R}$ is rank- d restricted L -Lipschitz

continuous for some quantity $L > 0$ if

$$|\mathcal{L}(\mathbf{X}_1) - \mathcal{L}(\mathbf{X}_2)| \leq L \|\mathbf{X}_1 - \mathbf{X}_2\|_F \quad (12)$$

holds for any $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{m \times n}$ such that $\mathbf{X}_1 - \mathbf{X}_2$ has rank at most $2d$.

Assumption 5.2 (Restricted Sharpness). A function $\mathcal{L}(\cdot) : \mathbb{R}^{m \times n} \mapsto \mathbb{R}$ is rank- d restricted μ -sharp w.r.t. $\mathbf{X}_\star$ for some $\mu > 0$ if

$$\mathcal{L}(\mathbf{X}) - \mathcal{L}(\mathbf{X}_\star) \geq \mu \|\mathbf{X} - \mathbf{X}_\star\|_F \quad (13)$$

holds for any $\mathbf{X} \in \mathbb{R}^{m \times n}$ with rank at most d .

5.2. Main Results

Let $\mathbf{F}_\star = \begin{bmatrix} \mathbf{L}_\star \\ \mathbf{R}_\star \end{bmatrix} \in \mathbb{R}^{(m+n) \times d}$, where $\mathbf{L}_\star = \mathbf{U}_\star \Sigma_\star^{\frac{1}{2}}$, $\mathbf{R}_\star = \mathbf{V}_\star \Sigma_\star^{\frac{1}{2}}$, and $\mathbf{F} = \begin{bmatrix} \mathbf{L} \\ \mathbf{R} \end{bmatrix} \in \mathbb{R}^{(m+n) \times d}$.

In our results, we propose a modified version of the distance metric introduced in (Tong et al., 2021b), defined as

$$\begin{aligned} \text{dist}^2(\mathbf{F}, \mathbf{F}_\star) &= \inf_{\mathbf{Q} \in G(d)} \|(\mathbf{L}\mathbf{Q} - \mathbf{L}_\star)(\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_F^2 \\ &\quad + \|(\mathbf{R}\mathbf{Q}^\top - \mathbf{R}_\star)(\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_F^2. \end{aligned}$$

Matrix $\mathbf{Q}$ is known as the alignment matrix and has long been used in similar distance metrics, (Tong et al., 2021a).

Lemma 5.3. *The proposed distance metric is an upper bound of the distance $\|\mathbf{L}\mathbf{R}^\top - \mathbf{X}_\star\|_F$. Namely, assuming that $\text{dist}(\mathbf{F}, \mathbf{F}_\star) \leq \lambda\epsilon$, it holds*

$$\|\mathbf{L}\mathbf{R}^\top - \mathbf{X}_\star\|_F \leq \left(1 + \frac{\epsilon}{2}\right) \sqrt{2} \text{dist}(\mathbf{F}, \mathbf{F}_\star). \quad (14)$$

Proof: The proof is deferred to supplementary material.

We next provide our main result, which relies on the assumption that the Lipschitz continuity and restricted sharpness condition are satisfied and ensure exact convergence.

Theorem 5.4 (Convergence of OPSA). *Let $\mathcal{L}(\mathbf{X}) : \mathbb{R}^{m \times n} \mapsto \mathbb{R}$ be convex w.r.t $\mathbf{X}$, and assume that it satisfies the rank- d restricted L -Lipschitz continuity assumption and the rank- d restricted μ -sharpness condition, defined above. Let also $\lambda = \frac{1}{20}$, and without loss of generality $\sigma_r(\mathbf{X}_\star) = 1$. Let $\epsilon = \frac{10^{-4}}{\chi \|\mathbf{X}_\star\|_{op}^{\frac{1}{2}}}$, and*

$$\text{dist}(\mathbf{F}_0, \mathbf{F}_\star) \leq \lambda\epsilon = \frac{10^{-4}}{\chi} \frac{\|\mathbf{X}_\star\|_{op}^{-\frac{1}{2}}}{20}, \quad (15)$$

where $\chi = \frac{L}{\mu}$. Then for the Overparameterized Preconditioned Subgradient Algorithm (OPSA) given in Algorithm 1,

with the Polyak's step-size defined in (11), we have

$$\begin{aligned} \text{dist}(\mathbf{F}_t, \mathbf{F}_\star) &\leq \left(1 - \frac{0.12}{\chi^2}\right)^{\frac{t}{2}} \frac{10^{-4} \|\mathbf{X}_\star\|_{op}^{-\frac{1}{2}}}{20 \cdot \chi}, \\ \|\mathbf{L}_t \mathbf{R}_t^\top - \mathbf{X}_\star\|_F &\leq \left(1 - \frac{0.12}{\chi^2}\right)^{\frac{t}{2}} \frac{1.5 \times 10^{-4} \|\mathbf{X}_\star\|_{op}^{-\frac{1}{2}}}{20 \cdot \chi}. \end{aligned}$$

Proof. The rate can be derived from its general form (see (16)) by setting $\lambda = \frac{\|\mathbf{X}_\star\|_{op}}{\bar{\lambda}}$, $\bar{\lambda} = c \|\mathbf{X}_\star\|_{op}$, $c = 20$. The detailed proof is deferred to supplementary material. $\square$

Remark 5.5. Our derived rate of convergence requires a good initialization that will satisfy condition (15). In practice, this condition can be satisfied by using a truncated spectral method as in (Zhang et al., 2016). It should be also noted that, even though the rate is independent of the condition number of $\mathbf{X}_\star$, the initialization condition is negatively affected as this condition number increases, requiring initializations closer to $\mathbf{X}_\star$.

Technical innovation. It should be noted that the derivation of the convergence rate in the *overparameterized and asymmetric* case that we focus on, requires a novel definition of the distance metric given in (23). Namely, previous convergence metrics dealing with the imbalance issue that shows up in the asymmetric low-rank matrix estimation problems, cannot be applied in the overparameterized setting due to overparameterization and non-invertibility of $\Sigma_\star$. In order to prove the contraction of this convergence metric we derive novel and non-trivial perturbation-type bounds of matrix norms, which pose their own challenges and move beyond existing approaches that address either the known-rank or unknown-rank and symmetric matrices regime (see Table 1).

Note also that the derived rate of convergence is valid in the overparameterized rank setting where $d \geq r$ and the value of λ should be positive i.e., $\lambda > 0$. The positive value of λ is required so that the matrix $\Sigma_\star + \lambda \mathbf{I}$ that is used in the convergence metric is always invertible. When $d = r$ the rate of OPSA can take the same form as the one of the scaled subgradient method of (Tong et al., 2021b), that is derived in the known rank regime. Namely, when $d = r$ then $\Sigma_\star$ is invertible and λ could be 0. To make the theorem applicable for $d = r$, we could replace the λ parameter appearing in (15), with $\lambda' = \lambda + \sigma_d(\mathbf{X}_\star)$. In that case, when $d > r$, we have $\lambda' = \lambda$ since $\sigma_d(\mathbf{X}_\star) = 0$. But when $d = r$, $\lambda' = \sigma_r(\mathbf{X}_\star)$ (since $\lambda = 0$), and the initialization condition and rate of convergence (see supplementary material) boils down to a similar form, up to some constants, to the one in (Tong et al., 2021b).

Remark 5.6. As is shown in Theorem 5.4, the rate is independent of the condition number of $\mathbf{X}_\star$. The linear rate provided in Theorem 5.4 is a simplified form of the general

expression:

$$\begin{aligned} \rho(\chi, \epsilon, \lambda) = & 1 \\ & - \frac{1}{2\chi^2} \sqrt{\frac{\sqrt{2}-1}{1+2\lambda}} \left(\sqrt{\frac{\sqrt{2}-1}{1+2\lambda}} \left(2 - \frac{1}{(1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\lambda}^{\frac{1}{4}}))^2} \right) \right. \\ & \left. - 2\chi\sqrt{2} \left(\frac{3}{2}\epsilon + 2 \frac{\sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\lambda}^{\frac{1}{4}})}{1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\lambda}^{\frac{1}{4}})} \right) \right). \end{aligned} \quad (16)$$

Note that this general form holds for any value of $\lambda \geq 0$, and its detailed derivation can be found in the supplementary material (Theorem A.3.1).

Remark 5.7. The level of overparameterization influences the rate of convergence through χ . Specifically, χ depends on the landscape properties of the objective as expressed by the Lipschitz constant L and the restricted sharpness constant μ . Generally speaking, $\chi = \frac{L}{\mu}$ increases with d ($L \uparrow, \mu \downarrow$) making the rate of convergence slower. This effect of the overestimated rank d on the rate of convergence is demonstrated in the numerical experiments section (see Figure 2).

5.3. Iteration Complexity for Matrix Sensing

In this section, we analyze the iteration complexity, i.e., the number of iterates T required to reach $\|L_T \mathbf{R}_T^\top - \mathbf{X}_*\|_F \leq \epsilon$, namely ϵ -accuracy. This study will focus on the matrix sensing problem.

5.3.1. NOISELESS MATRIX SENSING

Herein, we focus on the noiseless case without the presence of outliers, where it holds that $\mathbf{y} = \mathcal{A}(\mathbf{X}_*)$.

Definition 5.8 (Mixed-norm RIP). Let $2d > 0$ denote the rank of $\mathbf{X}$ and $\mathcal{A}(\cdot)$ a linear map. We define the $\delta_{2d}^-, \delta_{2d}^+$, as the lower and upper uniform bounds, respectively, of the quantity $\frac{\|\mathcal{A}(\mathbf{X})\|_1}{\|\mathbf{X}\|_F}$ for all matrices $\mathbf{X}$ of rank at most $2d$.

The mixed-norm RIP is empirically verified for the Gaussian linear map in Figure 1.

Proposition 5.9 (Lipschitz Continuity and Restricted Sharpness–No Outliers). *If $\mathcal{A}(\cdot)$ satisfies the mixed-norm property with constants $\delta_{2d}^-, \delta_{2d}^+$, then $\mathcal{L}(\mathbf{X}) = \|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_1$ satisfies the rank- d restricted L -Lipschitz continuity and rank- d restricted sharpness with constants,*

$$L = \delta_{2d}^+ \text{ and } \mu = \delta_{2d}^-. \quad (17)$$

Proof: Similar to Proposition 1 in (Tong et al., 2021b).

Corollary 5.10 (Iteration Complexity–No Outliers). *Using the same setting as in Theorem 5.4, we use the Lipschitz constant and restricted sharpness constant of Proposition 5.9, we get iteration complexity for noiseless matrix sensing is*

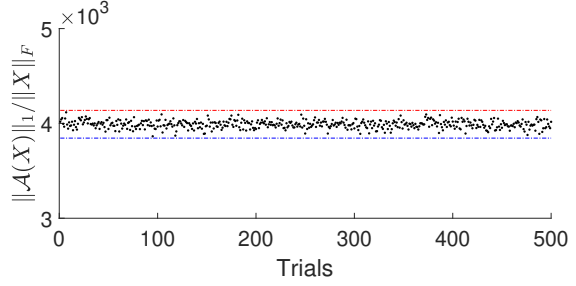
$$\mathcal{O}\left(\left(\frac{\delta_{2d}^+}{\delta_{2d}^-}\right)^2 \log\left(\frac{1}{\epsilon}\right)\right).$$


Figure 1. Values of $\frac{\|\mathcal{A}(\mathbf{X})\|_1}{\|\mathbf{X}\|_F}$ for Gaussian map $\mathcal{A} : \mathbb{R}^{500 \times 500} \rightarrow \mathbb{R}^{5000}$, where $\text{rank}(\mathbf{X}) = 10$. Each point represents one result of 500 random trials. Blue and red dash lines are the lower and upper uniform bounds, respectively.

5.3.2. ROBUST MATRIX SENSING

Here we consider the presence of outliers. We thus first define the $\mathcal{S}$ -outlier type bound condition. The $\mathcal{S}$ -outlier bound property has been used in the robust low-rank matrix recovery problem in prior works, e.g., (Tong et al., 2021b). It actually encodes a property of that allows restricted sharpness condition to be satisfied in matrix sensing problems in the presence of outliers. A detailed derivation of this condition, as a natural generalization of RIP can be found in (Charisopoulos et al., 2021).

Definition 5.11 ($\mathcal{S}$ -outlier Type Bound). The linear map $\mathcal{A}(\cdot)$ satisfies the rank- $2d$ $\mathcal{S}$ -outlier type bound w.r.t. a set $\mathcal{S}$ with a constant δ^0 if for all matrices $\mathbf{X} \in \mathbb{R}^{m \times n}$ of rank at most $2d$ we have

$$\delta^0 \|\mathbf{X}\|_F \leq \|\mathcal{A}_{\mathcal{S}^c}(\mathbf{X})\|_1 - \|\mathcal{A}_{\mathcal{S}}(\mathbf{X})\|_1, \quad (18)$$

where $\mathcal{A}_{\mathcal{S}}(\mathbf{X}) = \{\mathcal{A}_i(\mathbf{X})\}_{i \in \mathcal{S}}$ and $\mathcal{A}_{\mathcal{S}^c}(\mathbf{X}) = \{\mathcal{A}_i(\mathbf{X})\}_{i \in \mathcal{S}^c}$.

Proposition 5.12 (Matrix Sensing with Outliers). *Let $\mathcal{A}(\cdot)$ the rank- $2d$ mixed-norm RIP with $(\delta_{2d}^-, \delta_{2d}^+)$ and the $\mathcal{S}$ -outlier bound property defined above with δ^0 . Then $\mathcal{L}(\mathbf{X})$ satisfies the rank- d restricted L -Lipschitz continuity and μ -sharpness with*

$$L = \delta_{2d}^+ \text{ and } \mu = \delta^0. \quad (19)$$

Proof: The Lipschitz constant can be derived following similar steps as in the noiseless case. For the restricted sharpness constant μ we use Proposition 2 of (Tong et al., 2021b).

Corollary 5.13 (Iteration Complexity with Outliers). *Under the same setting as in Theorem 5.4, and by using the Lipschitz constant and restricted sharpness constant of Propositions 5.12, we get iteration complexity for matrix sensing, in the presence of outliers, is*

$$\mathcal{O}\left(\left(\frac{\delta_{2d}^+}{\delta^0}\right)^2 \log\left(\frac{1}{\epsilon}\right)\right).$$

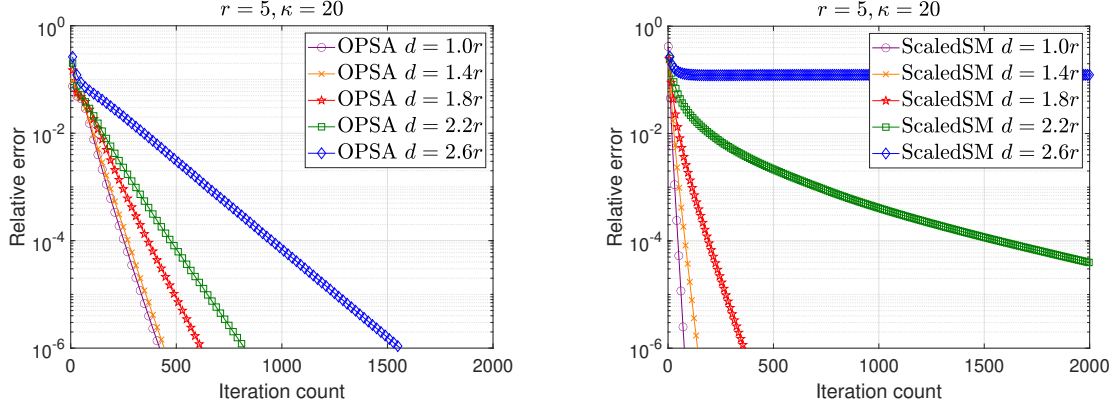


Figure 2. Performance comparison between OPSA (top) and ScaledSM (bottom) with different overparameterization d , where $n, r, \kappa, \lambda, \text{outlier} = 100, 5, 20, 2, 10\%$.

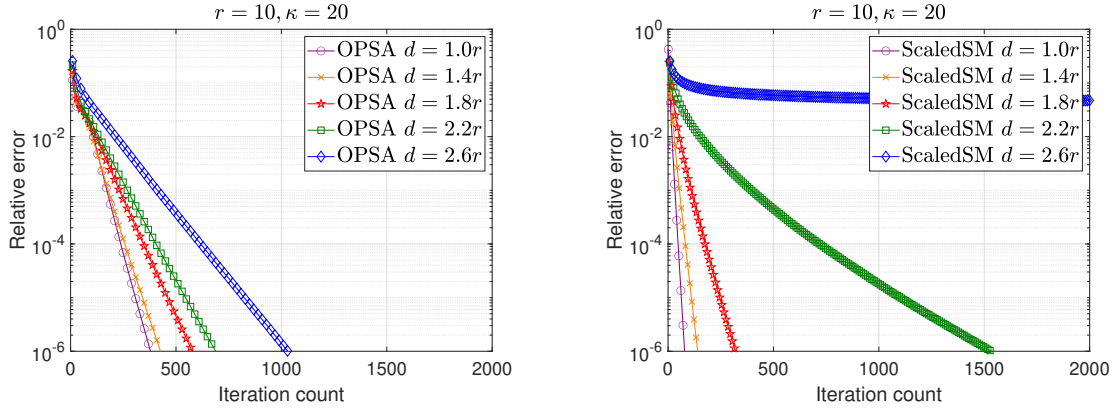


Figure 3. Performance comparison between OPSA (top) and ScaledSM (bottom) with different overparameterization d , where $n, r, \kappa, \lambda, \text{outlier} = 100, 10, 20, 2, 10\%$.

It should be noted that Corollary 5.13 implies that in the presence of outliers, the iteration complexity increases at an amount depending on the properties of the measurement matrices (through δ^0).

Remark 5.14. Assuming a measurement operator $\mathcal{A}_i(\mathbf{X}) = \frac{1}{p} \langle \mathbf{A}_i, \mathbf{X} \rangle$, where $\mathbf{A}_i$ has Gaussian i.i.d. entries $\mathcal{N}(0, 1)$, we can invoke the results of (Charisopoulos et al., 2021; Tong et al., 2021b), which show that $\mathcal{A}(\cdot)$ satisfies the mixed-RIP and $\mathcal{S}$ -outlier bound conditions with

$$\delta_{2d}^- \gtrsim 1, \quad \delta_{2d}^+ \lesssim 1, \quad \delta_0 \gtrsim 1 - 2p_s,$$

where $p_s \in [0, \frac{1}{2})$ is the fraction of outliers, as long as

$$p \gtrsim \frac{(m+n)d}{(1-2p_s)^2} \log \left(\frac{1}{1-2p_s} \right).$$

Hence, under the same setting as in Theorem 5.4 OPSA converges linearly to ϵ -accuracy in

$$O \left(\frac{1}{(1-2p_s)^2} \log \frac{1}{\epsilon} \right)$$

iterations assuming that it is initialized appropriately (see statement of Theorem 5.4).

6. Numerical Experiments

In this section, we verify the empirical performance of the proposed overparameterized Preconditioned Subgradient Algorithm (OPSA), i.e., Algorithm 1.

Experimental setup. The ground truth $\mathbf{X}_\star$ is generated as a product of two $n \times r$ random matrices, and then the condition number is adjusted to be κ by altering the singular values of $\mathbf{X}_\star$. The observations of matrix sensing are obtained as described in (8), that is

$$\mathbf{y}_i = \mathcal{A}_i(\mathbf{X}_\star) + \mathbf{s}_i, \quad 1 \leq i \leq p,$$

where the sensing operator $\mathcal{A}_i(\mathbf{X}_\star) = \frac{1}{p} \langle \mathbf{A}_i, \mathbf{X}_\star \rangle$ and $\mathbf{A}_i$ is a Gaussian random matrix. The corruption vector $\mathbf{s} \in \mathbb{R}^p$ contains randomly positioned outliers whose values are drawn uniformly at random from the interval $[-10\|\mathcal{A}(\mathbf{X}_\star)\|_\infty, 10\|\mathcal{A}_i(\mathbf{X}_\star)\|_\infty]$. Through this section,

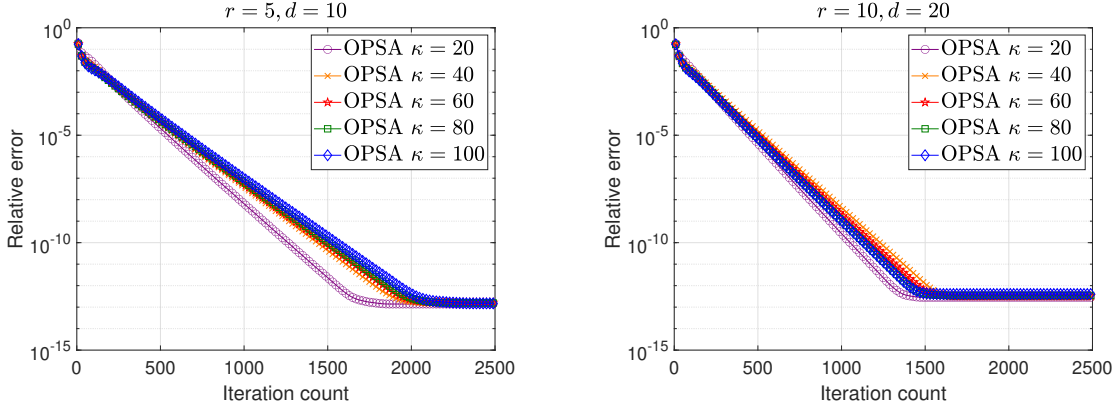


Figure 4. OPSA performance with different condition numbers κ , where $n, \lambda, \text{outlier} = 100, 2, 10\%$. **Top:** $r, d = 5, 10$. **Bottom:** $r, d = 10, 20$.

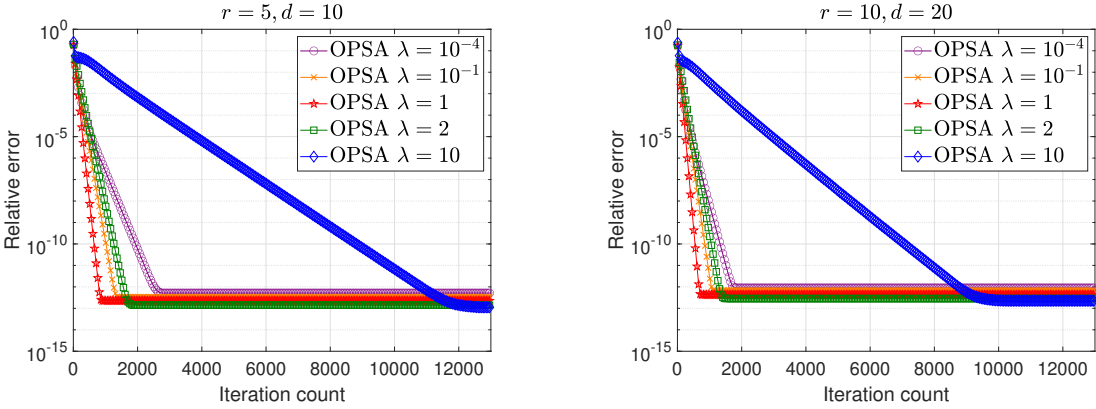


Figure 5. OPSA performance with different λ , where $n, \kappa, \text{outlier} = 100, 20, 10\%$. **Top:** $r, d = 5, 10$. **Bottom:** $r, d = 10, 20$.

$m = 8nr$ measurements are used in all tests. The code of ScaledSM is obtained from the authors’ website. For the sake of fairness, Polyak’s stepsizes are used for all tested algorithms, thus ScaledSM requires no extra parameter tuning. For OPSA, λ is the only parameter to be tuned, and an experiment in this section shows that λ can be easily tuned in a wide favorable range. All the tests are conducted with Matlab 2024a on a mobile workstation equipped with an Intel i9-12950HX CPU and 64GB of RAM. The Matlab implementation for the proposed OPSA is available online at <https://github.com/caesarcai/OPSA>.

Performance with different overparameterizations d . We compare the convergence performance of OPSA against ScaledSM (Tong et al., 2021b), the state-of-the-art subgradient method for robust matrix sensing, with different levels of overparameterization d . The convergence is evaluated with respect to the relative error $\|\mathbf{L}_t \mathbf{R}_t^\top - \mathbf{X}_*\|_F / \|\mathbf{X}_*\|_F$. The comparison results with true rank $r = 5$ and $r = 10$ are reported in Figures 2 and 3, respectively. For both algorithms, one can see that more overparameterization leads to

slower convergence. However, OPSA constantly achieves linear convergence while ScaledSM fails to converge when major overparameterization happens.

Performance with different condition numbers κ . In this experiment, we test the convergence performance of OPSA with different condition numbers of $\mathbf{X}_*$. The results with conditional numbers up to 100 are reported in Figure 4. One can observe that the convergence behavior and final accuracy of OPSA are not distinctly affected by larger condition numbers, even when the rank is heavily overestimated. This observation matches our main theoretical result. In Figure 7, we conduct additional experiments for OPSA with extremely large condition numbers κ up to 10,000. The results are consistent with Figure 4. Even with $\kappa = 1,000$ or $\kappa = 10,000$, OPSA still delivers stable convergence and achieves the same final accuracy.

Performance with different parameters λ . Notice that λ is the only parameter to be tuned in OPSA. In Figure 5, we test the convergence performance of OPSA with different λ . One can observe that a mild λ value (e.g., 0.1, 1, 2) helps

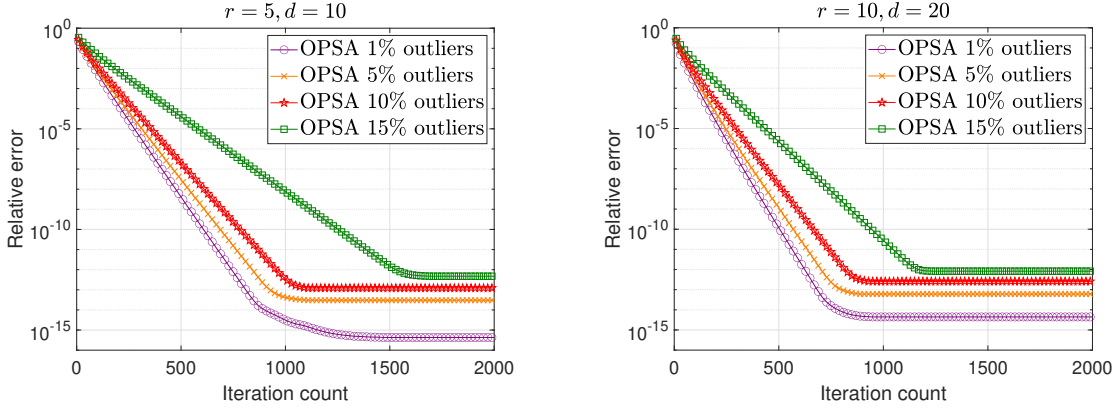


Figure 6. OPSPA performance with different outlier densities, where $n, \kappa, \lambda = 100, 20, 2$. **Top:** $r, d = 5, 10$. **Bottom:** $r, d = 10, 20$.

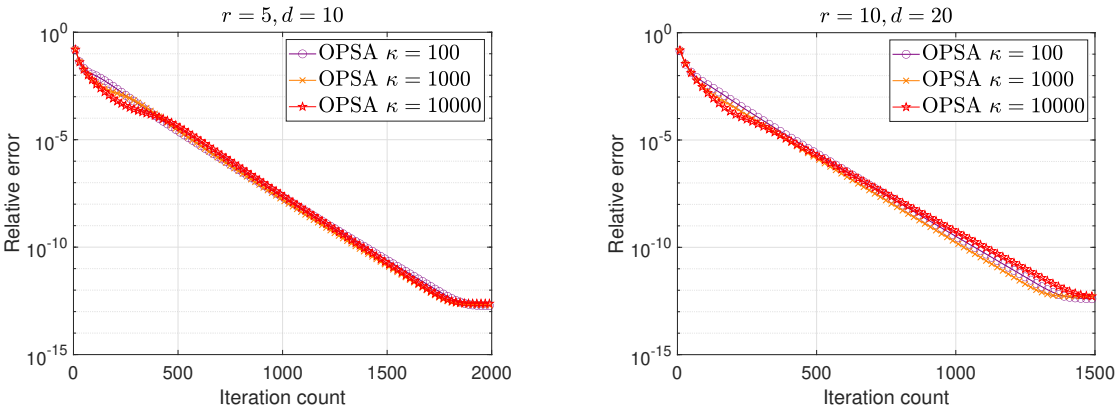


Figure 7. OPSPA performance with different large condition numbers κ , where $n, \lambda, \text{outlier} = 100, 2, 10\%$. **Left:** $r, d = 5, 10$. **Right:** $r, d = 10, 20$.

OPSPA maintain steep linear convergence when the rank is heavily overestimated. However, if the λ parameter is set too large (e.g., 10) or too tiny (e.g., 10^{-4}), OPSPA converges much slower, although still at a linear rate. Overall, OPSPA’s performance is favorable with a wide range of λ values, and thus the parameter tuning is easy for the proposed OPSPA.

Performance with different outlier densities. In Figure 6, OPSPA is tested against different outlier densities. OPSPA successfully recovers $\mathbf{X}_*$ and maintains linear convergence in all tests. Note that more iterations are needed, and slightly worse final accuracy can be achieved when more outliers are present. This is as expected since more outliers lead to harder recovery problems.

More numerical results. Additional numerical experiments on partially observed video background subtraction tasks, have been conducted and reported in the supplementary material.

7. Conclusions

In this paper, we proposed an overparameterized Preconditioned Subgradient Algorithm (OPSPA) for robust low-rank matrix recovery. Our work goes beyond existing SOTA works by addressing the challenging scenario of robust low-rank matrix recovery in the case of *asymmetric* matrices of *unknown rank* using a preconditioned-type approach. Under certain landscape assumptions i.e., Lipschitz continuity and restricted sharpness conditions, we a) established a linear rate of convergence that is independent of the condition number of the unknown matrix, and b) derived the iteration complexity matrix sensing, in the noiseless setting and in the presence of outliers. Numerical results corroborate our theoretical findings for different levels of overparameterization of the rank, and outliers, and the independence of the rate of convergence on the condition number of the sought matrix.

Acknowledgements

H.Q. Cai acknowledges partial support from the National Science Foundation (NSF) through grant DMS-2304489. R. Vidal acknowledges partial support of NSF grants 2212457 and 2031985, and Simons Foundation grant 814201.

Impact Statement

Low-rank matrix recovery has been broadly applied as one of the fundamental techniques in data mining and machine learning research. Hence, through its applications, the proposed preconditioned subgradient algorithm has potential broader impacts that data mining and machine learning have, as well as their limitations.

References

- Bhatia, R. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.
- Birman, M. S., Koplienko, L. S., and Solomyak, M. Z. Estimates of the spectrum of a difference of fractional powers of selfadjoint operators. *Izvestiya Vysshikh Uchebnykh Zavedenii. Matematika*, (3):3–10, 1975.
- Cai, H., Cai, J.-F., and Wei, K. Accelerated alternating projections for robust principal component analysis. *Journal of Machine Learning Research*, 20(1):685–717, 2019.
- Cai, H., Cai, J.-F., Wang, T., and Yin, G. Accelerated structured alternating projections for robust spectrally sparse signal recovery. *IEEE Transactions on Signal Processing*, 69:809–821, 2021a.
- Cai, H., Hamm, K., Huang, L., Li, J., and Wang, T. Rapid robust principal component analysis: CUR accelerated inexact low rank estimation. *IEEE Signal Processing Letters*, 28:116–120, 2021b.
- Cai, H., Liu, J., and Yin, W. Learned robust PCA: A scalable deep unfolding approach for high-dimensional outlier detection. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16977–16989, 2021c.
- Cai, H., Huang, L., Li, P., and Needell, D. Matrix completion with cross-concentrated sampling: Bridging uniform sampling and CUR sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10100–10113, 2023.
- Cai, H., Huang, L., Lu, X., and You, J. Accelerating ill-conditioned Hankel matrix recovery via structured Newton-like descent. *arXiv preprint arXiv:2406.07409*, 2024.
- Cai, H., Kundu, C., Liu, J., and Yin, W. Deeply learned robust matrix completion for large-scale low-rank data recovery. *arXiv preprint arXiv:2501.00677*, 2025.
- Candes, E. J. and Recht, B. Exact low-rank matrix completion via convex optimization. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 806–812. IEEE, 2008.
- Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3): 1–37, 2011.
- Charisopoulos, V., Chen, Y., Davis, D., Díaz, M., Ding, L., and Drusvyatskiy, D. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, 21(6):1505–1593, 2021.
- Chen, Y., Jalali, A., Sanghavi, S., and Caramanis, C. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7):4324–4337, 2013.
- Cheng, C. and Zhao, Z. Accelerating gradient descent for over-parameterized asymmetric low-rank matrix sensing via preconditioning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7705–7709, 2024.
- Chi, Y., Lu, Y. M., and Chen, Y. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019. doi: 10.1109/TSP.2019.2937282.
- Davenport, M. A. and Romberg, J. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4): 608–622, 2016.
- Ding, L., Jiang, L., Chen, Y., Qu, Q., and Zhu, Z. Rank over-specified robust matrix recovery: Subgradient method and exact recovery. *Advances in Neural Information Processing Systems*, 34:26767–26778, 2021.
- Giampouras, P., Vidal, R., Rontogiannis, A., and Haeffele, B. A novel variational form of the Schatten- p quasi-norm. *Advances in Neural Information Processing Systems*, 33: 21453–21463, 2020.
- Giampouras, P. V., Rontogiannis, A. A., and Koutroumbas, K. D. Robust pca via alternating iteratively reweighted low-rank matrix factorization. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 3383–3387, 2018. doi: 10.1109/ICIP.2018.8451212.
- Goffin, J.-L. On convergence rates of subgradient optimization methods. *Mathematical programming*, 13:329–347, 1977.

- Jain, P., Netrapalli, P., and Sanghavi, S. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674, 2013.
- Kümmerle, C. and Verdun, C. M. A scalable second order method for ill-conditioned matrix completion from few samples. In *International Conference on Machine Learning*, pp. 5872–5883. PMLR, 2021.
- Li, X., Zhu, Z., Man-Cho So, A., and Vidal, R. Nonconvex robust low-rank matrix recovery. *SIAM Journal on Optimization*, 30(1):660–686, 2020.
- Ma, J. and Fattahi, S. Implicit regularization of sub-gradient method in robust matrix recovery: Don’t be afraid of outliers. *arXiv preprint arXiv:2102.02969*, 95, 2021.
- Netrapalli, P., UN, N., Sanghavi, S., Anandkumar, A., and Jain, P. Non-convex robust pca. *Advances in neural information processing systems*, 27, 2014.
- Nie, F., Huang, H., and Ding, C. Low-rank matrix recovery via efficient Schatten p -norm minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pp. 655–661, 2012.
- Park, D., Kyrillidis, A., Bhojanapalli, S., Caramanis, C., and Sanghavi, S. Provable burer-monteiro factorization for a class of norm-constrained matrix problems. *arXiv preprint arXiv:1606.01316*, 2016.
- Park, D., Kyrillidis, A., Caramanis, C., and Sanghavi, S. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In *Artificial Intelligence and Statistics*, pp. 65–74. PMLR, 2017.
- Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.
- Smith, C., Cai, H., and Tasissa, A. Riemannian optimization for non-convex euclidean distance geometry with global recovery guarantees. *arXiv preprint arXiv:2410.06376*, 2024.
- Soltanolkotabi, M., Stöger, D., and Xie, C. Implicit balancing and regularization: Generalization and convergence guarantees for overparameterized asymmetric matrix sensing. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 5140–5142. PMLR, 2023.
- Stöger, D. and Soltanolkotabi, M. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34:23831–23843, 2021.
- Tong, T., Ma, C., and Chi, Y. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021a.
- Tong, T., Ma, C., and Chi, Y. Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number. *IEEE Transactions on Signal Processing*, 69:2396–2409, 2021b.
- Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pp. 964–973. PMLR, 2016.
- Wang, X., Zhou, T., Zhu, J., Liu, J., Yuan, K., Yao, T., Yin, W., Jin, R., and Cai, H. S³Attention: Improving long sequence attention with smoothed skeleton sketching. *IEEE Journal of Selected Topics in Signal Processing*, 18(6): 985–996, 2024.
- Wei, K., Cai, J.-F., Chan, T. F., and Leung, S. Guarantees of riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.
- Xiong, N., Ding, L., and Du, S. S. How overparameterization slows down gradient descent in matrix sensing: The curses of symmetry and initialization. *arXiv preprint arXiv:2310.01769*, 2023.
- Xu, X., Shen, Y., Chi, Y., and Ma, C. The power of preconditioning in overparameterized low-rank matrix sensing. In *International Conference on Machine Learning*, pp. 38611–38654. PMLR, 2023.
- Zhang, F. and Pilanci, M. Riemannian preconditioned loRA for fine-tuning foundation models. In *Forty-first International Conference on Machine Learning*, 2024.
- Zhang, G., Fattahi, S., and Zhang, R. Y. Preconditioned gradient descent for overparameterized nonconvex burer-monteiro factorization with global optimality certification. *Journal of Machine Learning Research*, 24(163):1–55, 2023.
- Zhang, H., Chi, Y., and Liang, Y. Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow. In *International conference on machine learning*, pp. 1022–1031. PMLR, 2016.
- Zhang, J., Zhang, R. Y., and Chiu, H.-M. Fast and accurate estimation of low-rank matrices from noisy measurements via preconditioned non-convex gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 3772–3780. PMLR, 2024.

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

Supplementary Materials for Guarantees of a Preconditioned Subgradient Algorithm for Overparameterized Asymmetric Low-rank Matrix Recovery

A.1. Problem Formulation

We focus on the following low-rank matrix estimation problem

$$\min_{\mathbf{X}} \mathcal{L}(\mathbf{Y}, \mathbf{X}) \equiv \min_{\mathbf{L}, \mathbf{R}, \mathbf{L}\mathbf{R}^\top = \mathbf{X}} \mathcal{L}(\mathbf{Y}, \mathbf{L}\mathbf{R}^\top). \quad (20)$$

We assume a preconditioned alternating gradient descent algorithm consisting of the following steps:

$$\mathbf{L}_{t+1} = \mathbf{L}_t - \eta_t (\mathbf{S}_t \mathbf{R}_t) (\mathbf{R}_t^\top \mathbf{R}_t + \lambda \mathbf{I})^{-1}, \quad (21)$$

$$\mathbf{R}_{t+1} = \mathbf{R}_t - \eta_t (\mathbf{S}_t^\top \mathbf{L}_t) (\mathbf{L}_t^\top \mathbf{L}_t + \lambda \mathbf{I})^{-1}, \quad (22)$$

with $\mathbf{S}_t \in \partial_{\mathbf{X}} l(\mathbf{Y}, \mathbf{X}_t)$.

Let $\mathbf{X}_\star$ be the ground truth matrix, $\mathbf{F}_\star = \begin{bmatrix} \mathbf{L}_\star \\ \mathbf{R}_\star \end{bmatrix} \in \mathbb{R}^{(m+n) \times d}$, where $d \geq r = \text{rank}(\mathbf{X}_\star)$, $\mathbf{L}_\star = \mathbf{U}_\star \Sigma_\star^{\frac{1}{2}}$, $\mathbf{R}_\star = \mathbf{V}_\star \Sigma_\star^{\frac{1}{2}}$, and $\mathbf{F} = \begin{bmatrix} \mathbf{L} \\ \mathbf{R} \end{bmatrix} \in \mathbb{R}^{(m+n) \times d}$. We define the following convergence metric,

$$\text{dist}(\mathbf{F}, \mathbf{F}_\star) = \sqrt{\inf_{\mathbf{Q} \in \mathbf{G}(d)} \|(\mathbf{L}\mathbf{Q} - \mathbf{L}_\star)(\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_F^2 + \|(\mathbf{R}\mathbf{Q}^{-\top} - \mathbf{R}_\star)(\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_F^2}. \quad (23)$$

In our analysis we focus on the Polyak's type step size, defined as follows,

$$\eta_t = \frac{\mathcal{L}(\mathbf{L}_t \mathbf{R}_t^\top) - \mathcal{L}(\mathbf{L}_\star \mathbf{R}_\star^\top)}{\|\mathbf{S}_t \mathbf{R}_t (\mathbf{R}_t^\top \mathbf{R}_t + \lambda \mathbf{I})^{-\frac{1}{2}}\|_F^2 + \|\mathbf{S}_t^\top \mathbf{L}_t (\mathbf{L}_t^\top \mathbf{L}_t + \lambda \mathbf{I})^{-\frac{1}{2}}\|_F^2}. \quad (24)$$

Algorithm 1 Overparameterized Preconditioned Subgradient Algorithm (OPSA)

- 1: **Input:** $\mathbf{y} \in \mathbb{R}^p$, d : overestimated rank, $\lambda > 0$: regularization parameter, η : stepsize.
 - 2: **Initialize** $\mathbf{L}_0$ and $\mathbf{R}_0$, set $t = 0$.
 - 3: **while** ! Stop Condition **do**
 - 4: $\mathbf{L}_{t+1} = \mathbf{L}_t - \eta_t \mathbf{S}_t \mathbf{R}_t (\mathbf{R}_t^\top \mathbf{R}_t + \lambda \mathbf{I})^{-1}$
 - 5: $\mathbf{R}_{t+1} = \mathbf{R}_t - \eta_t \mathbf{S}_t^\top \mathbf{L}_t (\mathbf{L}_t^\top \mathbf{L}_t + \lambda \mathbf{I})^{-1}$
 - 6: $t = t + 1$
 - 7: **end while**
 - 8: **Output:** $\mathbf{X} = \mathbf{L}_t \mathbf{R}_t^\top$.
-

A.2. Auxiliary Lemmata

Lemma A.2.1 (Theorem 1 (Birman et al., 1975); Theorem X.1.1, (Bhatia, 2013)). *Let A, B positive definite matrices. It holds,*

$$\|A^{\frac{1}{2}} - B^{\frac{1}{2}}\|_{op} \leq \|A - B\|_{op}^{\frac{1}{2}}. \quad (25)$$

Lemma A.2.2. *Let $\|X - X_\star\|_{op} \leq \frac{1}{2}\sigma_r(X_\star)$, then it holds $\|X - X_\star\|_F \geq \sqrt{\frac{(\sqrt{2}-1)\sigma_r(X_\star)}{\sigma_r(X_\star)+2\lambda}} \text{dist}(F, F_\star)$.*

Proof: Let us denote $Q_\star$ as the optimal alignment matrix for $\text{dist}(F, F_\star)$. We will have,

$$\begin{aligned} \text{dist}^2(F, F_\star) &= \|(LQ_\star - L_\star)(\Sigma_\star + \lambda I)^{\frac{1}{2}}\|_F^2 + \|(RQ_\star^{-\top} - R_\star)(\Sigma_\star + \lambda I)^{\frac{1}{2}}\|_F^2 \\ &= \text{tr}\{(\Sigma_\star + \lambda I)^{\frac{1}{2}}(LQ_\star - L_\star)^\top(LQ_\star - L_\star)(\Sigma_\star + \lambda I)^{\frac{1}{2}}\} \\ &\quad + \text{tr}\{(\Sigma_\star + \lambda I)^{\frac{1}{2}}(RQ_\star^{-\top} - R_\star)^\top(RQ_\star^{-\top} - R_\star)(\Sigma_\star + \lambda I)^{\frac{1}{2}}\} \\ &= \text{tr}\{(LQ_\star - L_\star)^\top(LQ_\star - L_\star)(\Sigma_\star + \lambda I)\} + \text{tr}\{(RQ_\star^{-\top} - R_\star)^\top(RQ_\star^{-\top} - R_\star)(\Sigma_\star + \lambda I)\} \\ &= \|(LQ_\star - L_\star)\Sigma_\star^{\frac{1}{2}}\|_F^2 + \|(RQ_\star^{-\top} - R_\star)\Sigma_\star^{\frac{1}{2}}\|_F^2 + \lambda(\|(LQ_\star - L_\star)\|_F^2 + \|(RQ_\star^{-\top} - R_\star)\|_F^2). \end{aligned}$$

By using the definition of the distance and Lemma 11 in (Tong et al., 2021a), and Lemma 5.14 in (Tu et al., 2016), we have

$$\text{dist}^2(F, F_\star) \leq \frac{1}{\sqrt{2}-1}\|X - X_\star\|_F^2 + \frac{2\lambda}{(\sqrt{2}-1)\sigma_r(X_\star)}\|X - X_\star\|_F^2. \quad (26)$$

Hence, we have

$$\text{dist}^2(F, F_\star) \leq \left(\frac{1}{\sqrt{2}-1} + \frac{2\lambda}{(\sqrt{2}-1)\sigma_r(X_\star)}\right)\|X - X_\star\|_F^2 = \frac{\sigma_r(X_\star) + 2\lambda}{(\sqrt{2}-1)\sigma_r(X_\star)}\|X - X_\star\|_F^2, \quad (27)$$

which leads us to inequality

$$\|X - X_\star\|_F \geq \sqrt{\frac{(\sqrt{2}-1)\sigma_r(X_\star)}{\sigma_r(X_\star) + 2\lambda}} \text{dist}(F, F_\star). \quad (28)$$

Lemma A.2.3. *The proposed distance metric given in (23), is an upper bound of the distance $\|LR^\top - X_\star\|_F$. Namely, assuming that $\text{dist}(F, F_\star) \leq \lambda\epsilon$, it holds*

$$\|LR^\top - X_\star\|_F \leq \left(1 + \frac{\epsilon}{2}\right) \sqrt{2} \text{dist}(F, F_\star). \quad (29)$$

Proof: We use the following distance metric,

$$\text{dist}(F, F_\star) = \sqrt{\inf_{Q \in G(d)} \|(LQ - L_\star)(\Sigma_\star + \lambda I)^{\frac{1}{2}}\|_F^2 + \|(RQ^{-\top} - R_\star)(\Sigma_\star + \lambda I)^{\frac{1}{2}}\|_F^2} \quad (30)$$

and we assume that $\text{dist}(F, F_\star) \leq \lambda\epsilon$.

By using the following known norm inequalities

$$\|AB\|_F \geq \|A\|_F \sigma_r(B) \geq \|A\|_{op} \sigma_r(B), \quad (31)$$

we can get

$$\|(L - L_\star)(\Sigma_\star + \lambda I)^{-\frac{1}{2}}\|_{op}^2 \lambda^2 + \|(R - R_\star)(\Sigma_\star + \lambda I)^{-\frac{1}{2}}\|_{op}^2 \lambda^2 \leq \lambda^2 \epsilon^2, \quad (32)$$

which implies

$$\max\{\|(L - L_\star)(\Sigma_\star + \lambda I)^{-\frac{1}{2}}\|_{op}, \|(R - R_\star)(\Sigma_\star + \lambda I)^{-\frac{1}{2}}\|_{op}\} \leq \epsilon. \quad (33)$$

Next, we show the relation between $\|L_t R_t^\top - X_\star\|_F$ and $\text{dist}(F_t, F_\star)$,

$$\begin{aligned}
 \|L_t R_t^\top - X_\star\|_F &= \|L R^\top - X_\star\|_F = \|\Delta_L R_\star^\top + L_\star \Delta_R^\top + \Delta_L \Delta_R^\top\|_F \\
 &\leq \|\Delta_L R_\star^\top\|_F + \|L_\star \Delta_R^\top\|_F + \|\Delta_L \Delta_R^\top\|_F = \|\Delta_L R_\star^\top\|_F + \|L_\star \Delta_R^\top\|_F \\
 &\quad + \|\frac{1}{2} \Delta_L (\Sigma_\star + \lambda I)^{\frac{1}{2}} (\Sigma_\star + \lambda I)^{-\frac{1}{2}} \Delta_R^\top + \frac{1}{2} \Delta_L (\Sigma_\star + \lambda I)^{\frac{1}{2}} (\Sigma_\star + \lambda I)^{-\frac{1}{2}} \Delta_R^\top\|_F \\
 &\leq \|\Delta_L (\Sigma_\star + \lambda I)^{\frac{1}{2}}\|_F + \|\Delta_R (\Sigma_\star + \lambda I)^{\frac{1}{2}} I\|_F + \frac{1}{2} \|\Delta_L (\Sigma_\star + \lambda I)^{\frac{1}{2}}\|_F \|\Delta_R (\Sigma_\star + \lambda I)^{-\frac{1}{2}}\|_{op} \\
 &\quad + \frac{1}{2} \|\Delta_L (\Sigma_\star + \lambda I)^{-\frac{1}{2}}\|_{op} \|\Delta_R (\Sigma_\star + \lambda I)^{\frac{1}{2}}\|_F \\
 &\leq \left(1 + \frac{1}{2} \max\{\|\Delta_R (\Sigma_\star + \lambda I)^{-\frac{1}{2}}\|_{op}, \|\Delta_L (\Sigma_\star + \lambda I)^{-\frac{1}{2}}\|_{op}\}\right) \left(\|\Delta_L (\Sigma_\star + \lambda I)^{\frac{1}{2}}\|_F + \|\Delta_R (\Sigma_\star + \lambda I)^{\frac{1}{2}}\|_F\right) \\
 &\leq \left(1 + \frac{\epsilon}{2}\right) \sqrt{2} \text{dist}(F, F_\star) \leq \left(1 + \frac{\epsilon}{2}\right) \sqrt{2} \lambda \epsilon,
 \end{aligned}$$

where he have used the initialization condition, the fact that

$$\max\{\|\Delta_R (\Sigma_\star + \lambda I)^{-\frac{1}{2}}\|_{op}, \|\Delta_L (\Sigma_\star + \lambda I)^{-\frac{1}{2}}\|_{op}\} \leq \epsilon \quad (34)$$

and the inequality

$$\begin{aligned}
 &\|\Delta_L (\Sigma_\star + \lambda I)^{\frac{1}{2}}\|_F + \|\Delta_R (\Sigma_\star + \lambda I)^{\frac{1}{2}}\|_F \\
 &\leq \sqrt{2} \sqrt{\|\Delta_L (\Sigma_\star + \lambda I)^{\frac{1}{2}}\|_F^2 + \|\Delta_R (\Sigma_\star + \lambda I)^{\frac{1}{2}}\|_F^2}.
 \end{aligned} \quad (35)$$

This finishes the proof. $\square$

Lemma A.2.4. Let us assume $\text{dist}(F, F_\star) \leq \epsilon \lambda$, where $\lambda = \frac{\|X_\star\|_{op}}{\lambda}$. Let $\sigma_r(X_\star) = 1$ and denote $\Delta_\lambda R = (R^\top R + \lambda I)^{1/2} - (R_\star^\top R_\star + \lambda I)^{1/2}$, $\Delta_\lambda L = (L^\top L + \lambda I)^{1/2} - (L_\star^\top L_\star + \lambda I)^{1/2}$, then it holds

$$\begin{aligned}
 \|(\Sigma_\star + \lambda I)^{-\frac{1}{2}} \Delta_\lambda L\|_{op} &\leq \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\lambda}^{\frac{1}{4}}), \\
 \|(\Sigma_\star + \lambda I)^{-\frac{1}{2}} \Delta_\lambda R\|_{op} &\leq \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\lambda}^{\frac{1}{4}}).
 \end{aligned} \quad (36)$$

Proof: We have

$$\|(\Sigma_\star + \lambda I)^{-\frac{1}{2}} \Delta_\lambda L\|_{op} = \|((L^\top L + \lambda I)(\Sigma_\star + \lambda I)^{-1})^{\frac{1}{2}} - ((L_\star^\top L_\star + \lambda I)(\Sigma_\star + \lambda I)^{-1})^{\frac{1}{2}}\|_{op}. \quad (37)$$

Let us denote $A = (L^\top L + \lambda I)(\Sigma_\star + \lambda I)^{-1}$ and $B = (L_\star^\top L_\star + \lambda I)(\Sigma_\star + \lambda I)^{-1}$. From Lemma A.2.1, we have that $\|A^{\frac{1}{2}} - B^{\frac{1}{2}}\|_{op} \leq \|A - B\|_{op}^{\frac{1}{2}}$.

Hence,

$$\begin{aligned}
 &\|((L^\top L + \lambda I)(\Sigma_\star + \lambda I)^{-1})^{\frac{1}{2}} - ((L_\star^\top L_\star + \lambda I)(\Sigma_\star + \lambda I)^{-1})^{\frac{1}{2}}\|_{op} \\
 &\leq \|(L^\top L + \lambda I)(\Sigma_\star + \lambda I)^{-1} - (L_\star^\top L_\star + \lambda I)(\Sigma_\star + \lambda I)^{-1}\|_{op}^{\frac{1}{2}} = \|(L^\top L - L_\star^\top L_\star)(\Sigma_\star + \lambda I)^{-1}\|_{op}^{\frac{1}{2}} \\
 &= \|(L^\top L - L^\top L_\star + L^\top L_\star - L_\star^\top L_\star)(\Sigma_\star + \lambda I)^{-1}\|_{op}^{\frac{1}{2}} = \|(L^\top (L - L_\star) + (L - L_\star)^\top L_\star)(\Sigma_\star + \lambda I)^{-1}\|_{op}^{\frac{1}{2}} \\
 &= \|((L - L_\star + L_\star)^\top (L - L_\star) + (L - L_\star)^\top L_\star)(\Sigma_\star + \lambda I)^{-1}\|_{op}^{\frac{1}{2}} \\
 &= \|((L - L_\star)^\top (L - L_\star) + L_\star^\top (L - L_\star) + (L - L_\star)^\top L_\star)(\Sigma_\star + \lambda I)^{-1}\|_{op}^{\frac{1}{2}} \\
 &\leq \|((L - L_\star)^\top (L - L_\star) + L_\star^\top (L - L_\star) + (L - L_\star)^\top L_\star)\|_{op}^{\frac{1}{2}} \|(\Sigma_\star + \lambda I)^{-1}\|_{op}^{\frac{1}{2}} \\
 &\leq \frac{1}{\lambda^{\frac{1}{2}}} \left(\|L - L_\star\|_{op} + \sqrt{2}\|L_\star\|_{op}^{\frac{1}{2}}\|(L - L_\star)\|_{op}^{\frac{1}{2}}\right) \\
 &\leq \frac{1}{\lambda^{\frac{1}{2}}} \left(\lambda^{\frac{1}{2}}\epsilon + \sqrt{2}\|X_\star\|_{op}^{\frac{1}{4}}\lambda^{\frac{1}{4}}\sqrt{\epsilon}\right) \\
 &= \epsilon + \sqrt{2}\sqrt{\epsilon}\|X_\star\|_{op}^{\frac{1}{4}}\lambda^{-\frac{1}{4}} = \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\lambda}^{\frac{1}{4}}).
 \end{aligned}$$

From the last inequality, we get that,

$$\|(\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}} \Delta_\lambda \mathbf{L}\|_{op} \leq \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}}). \quad (38)$$

The proof for the term $\|(\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}} \Delta_\lambda \mathbf{R}\|_{op}$ can be derived following similar steps. $\square$

Lemma A.2.5. *Let $\mathbf{L} \in \mathbb{R}^{m \times d}$, $\mathbf{R} \in \mathbb{R}^{n \times d}$ and denote $\Delta_\lambda \mathbf{R} = (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{1/2} - (\mathbf{R}_\star^\top \mathbf{R}_\star + \lambda \mathbf{I})^{1/2}$, $\Delta_\lambda \mathbf{L} = (\mathbf{L}^\top \mathbf{L} + \lambda \mathbf{I})^{1/2} - (\mathbf{L}_\star^\top \mathbf{L}_\star + \lambda \mathbf{I})^{1/2}$. Assume that $\max\{\|(\mathbf{L} - \mathbf{L}_\star)(\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}}\|_{op}, \|(\mathbf{R} - \mathbf{R}_\star)(\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}}\|_{op}\} \leq \epsilon$. Then it holds,*

$$\begin{aligned} \|(\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-\frac{1}{2}} (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_{op} &\leq \frac{1}{1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}})}, \\ \|(\mathbf{L}^\top \mathbf{L} + \lambda \mathbf{I})^{-\frac{1}{2}} (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_{op} &\leq \frac{1}{1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}})}. \end{aligned} \quad (39)$$

Proof: We have $(\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-\frac{1}{2}} (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}} = \left((\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{\frac{1}{2}} \right)^{-1}$. Hence,

$$\|(\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-\frac{1}{2}} (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_{op} = \frac{1}{\sigma_d \left((\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{\frac{1}{2}} \right)}. \quad (40)$$

By Weyl's inequality, we get

$$\begin{aligned} &\sigma_d \left((\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{\frac{1}{2}} \right) \\ &\geq \sigma_d \left((\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}} (\mathbf{R}_\star^\top \mathbf{R}_\star + \lambda \mathbf{I})^{\frac{1}{2}} \right) - \|(\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}} \Delta_\lambda \mathbf{R}\|_{op} \\ &= 1 - \|(\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}} \Delta_\lambda \mathbf{R}\|_{op}. \end{aligned} \quad (41)$$

From Lemma A.2.4, we have

$$\|(\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}} \Delta_\lambda \mathbf{R}\|_{op} \leq \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}}). \quad (42)$$

Hence, we conclude the proof. The proof for $\|(\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}} \Delta_\lambda \mathbf{R}\|_{op}$ can be similarly derived. $\square$

A.3. Proof of convergence of OPSA

Theorem A.3.1 (Convergence of Overparameterized Preconditioned Subgradient Algorithm). *Let $\mathcal{L}(\mathbf{X}) : \mathbb{R}^{m \times n} \mapsto \mathbb{R}$ be convex w.r.t $\mathbf{X}$, and assume that it satisfies the rank- d restricted L -Lipschitz continuity assumption and the rank- d restricted μ -sharpness condition, defined above. Assume also that*

$$\text{dist}(\mathbf{F}_0, \mathbf{F}_\star) \leq \lambda \epsilon \quad (43)$$

and let $\lambda = \frac{\|\mathbf{X}_\star\|_{op}}{\bar{\lambda}}$, $\sigma_r(\mathbf{X}_\star) = 1$, and $\chi = \frac{L}{\mu}$. Then for the Overparameterized Preconditioned Subgradient algorithm given in Algorithm 1, with the Polyak's stepsize defined in (24), we have,

$$\text{dist}(\mathbf{F}_t, \mathbf{F}_\star) \leq \rho(\chi, \epsilon, \lambda)^{\frac{t}{2}} \left(\frac{\|\mathbf{X}\|_{op}}{\bar{\lambda}} \epsilon \right)^{\frac{1}{2}}, \quad (44)$$

where $\rho(\chi, \epsilon, \lambda)$ is given by

$$\begin{aligned} \rho(\chi, \epsilon, \lambda) &= 1 - \frac{1}{2\chi^2} \sqrt{\frac{\sqrt{2}-1}{1+2\lambda}} \left(\sqrt{\frac{\sqrt{2}-1}{1+2\lambda}} \left(2 - \frac{1}{(1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}}))^2} \right) \right. \\ &\quad \left. - 2\chi\sqrt{2} \left(\frac{3}{2}\epsilon + 2 \frac{\sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}})}{1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}})} \right) \right). \end{aligned} \quad (45)$$

Proof: We denote $\mathbf{L} = \mathbf{L}_t \mathbf{Q}_t$, $\mathbf{R} = \mathbf{R}_t \mathbf{Q}_t^{-\top}$, $\Delta_{\mathbf{L}} = \mathbf{L} - \mathbf{L}_\star$, $\Delta_{\mathbf{R}} = \mathbf{R} - \mathbf{R}_\star$ and $\mathbf{S} = \mathbf{S}_t$.

We focus on the contraction of $\text{dist}(\mathbf{F}_{t+1}, \mathbf{F}_\star)$.

$$\text{dist}^2(\mathbf{F}_{t+1}, \mathbf{F}_\star) \leq \|(\mathbf{L}_{t+1} \mathbf{Q}_t - \mathbf{L}_\star)(\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_F^2 + \|(\mathbf{R}_{t+1} \mathbf{Q}_t^{-\top} - \mathbf{R}_\star)(\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_F^2. \quad (46)$$

Let $\mathbf{L} = \mathbf{L}_t \mathbf{Q}_t$, $\mathbf{R} = \mathbf{R}_t \mathbf{Q}_t^{-\top}$ and $\mathbf{S} = \mathbf{S}_t$ and first bound the term $\|(\mathbf{L}_{t+1} \mathbf{Q}_t - \mathbf{L}_\star)(\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_F^2$ as follows,

$$\begin{aligned} & \|(\mathbf{L}_{t+1} \mathbf{Q}_t - \mathbf{L}_\star)(\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_F^2 = \left\| \left(\mathbf{L} - \eta \mathbf{S} \mathbf{R} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-1} - \mathbf{L}_\star \right) (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}} \right\|_F^2 \\ &= \|\Delta_{\mathbf{L}} (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_F^2 - 2\eta \langle \mathbf{S}, \Delta_{\mathbf{L}} (\Sigma_\star + \lambda \mathbf{I}) (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-1} \mathbf{R}^\top \rangle \\ & \quad + \eta^2 \|\mathbf{S} \mathbf{R} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-1} (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_F^2 \\ &= \|\Delta_{\mathbf{L}} (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_F^2 - 2\eta \underbrace{\langle \mathbf{S}, \Delta_{\mathbf{L}} \mathbf{R}_\star^\top + \frac{1}{2} \Delta_{\mathbf{L}} \Delta_{\mathbf{R}}^\top \rangle}_{\mathcal{O}_1} \\ & \quad - 2\eta \underbrace{\langle \mathbf{S}, \Delta_{\mathbf{L}} (\Sigma_\star + \lambda \mathbf{I}) (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-1} \mathbf{R}^\top - \Delta_{\mathbf{L}} \mathbf{R}_\star^\top - \frac{1}{2} \Delta_{\mathbf{L}} \Delta_{\mathbf{R}}^\top \rangle}_{\mathcal{O}_2} \\ & \quad + \eta^2 \underbrace{\|\mathbf{S} \mathbf{R} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-1} (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_F^2}_{\mathcal{O}_3}. \end{aligned}$$

We focus on bounding the term $|\mathcal{O}_2|$,

$$\begin{aligned} & |\langle \mathbf{S}, \Delta_{\mathbf{L}} (\Sigma_\star + \lambda \mathbf{I}) (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-1} \mathbf{R}^\top - \Delta_{\mathbf{L}} \mathbf{R}_\star^\top - \frac{1}{2} \Delta_{\mathbf{L}} \Delta_{\mathbf{R}}^\top \rangle| \\ & \leq \|\mathbf{S}\|_F \|\Delta_{\mathbf{L}} (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}}\|_F \times \\ & \quad (\|\mathbf{R} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-1} (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}} - \mathbf{R}_\star (\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}}\|_{op} + \frac{1}{2} \|\Delta_{\mathbf{R}} (\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}}\|_{op}), \end{aligned} \quad (47)$$

where we have used the inequalities $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_{op} \|\mathbf{B}\|_F$ and $\|\mathbf{A} - \mathbf{B}\|_{op} \leq \|\mathbf{A}\|_{op} + \|\mathbf{B}\|_{op}$.

For the first term of (47) we have

$$\begin{aligned} & \|\mathbf{R} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-1} (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}} - \mathbf{R}_\star (\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}}\|_{op} \\ &= \|\mathbf{R} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-\frac{1}{2}} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-\frac{1}{2}} (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}} - \mathbf{R}_\star (\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}}\|_{op} \\ &= \|\mathbf{R} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-\frac{1}{2}} \left(\left((\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-\frac{1}{2}} - (\mathbf{R}_\star^\top \mathbf{R}_\star + \lambda \mathbf{I})^{-\frac{1}{2}} \right) (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}} + (\mathbf{R}_\star^\top \mathbf{R}_\star + \lambda \mathbf{I})^{-\frac{1}{2}} (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}} \right) \right. \\ & \quad \left. - \mathbf{R}_\star (\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}}\|_{op} \\ &= \|\mathbf{R} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-\frac{1}{2}} \left(\left((\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-\frac{1}{2}} - (\mathbf{R}_\star^\top \mathbf{R}_\star + \lambda \mathbf{I})^{-\frac{1}{2}} \right) (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}} + \mathbf{I} \right) - \mathbf{R}_\star (\Sigma_\star + \lambda \mathbf{I})^{-\frac{1}{2}}\|_{op} \\ &\leq \|\mathbf{R} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-\frac{1}{2}}\|_{op} \left\| \left((\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-\frac{1}{2}} - (\mathbf{R}_\star^\top \mathbf{R}_\star + \lambda \mathbf{I})^{-\frac{1}{2}} \right) (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}} \right\|_{op} \\ & \quad + \|\mathbf{R} (\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-\frac{1}{2}} - \mathbf{R}_\star (\mathbf{R}_\star^\top \mathbf{R}_\star + \lambda \mathbf{I})^{-\frac{1}{2}}\|_{op}. \end{aligned}$$

We now focus on bounding the term $\| \left((\mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I})^{-\frac{1}{2}} - (\mathbf{R}_\star^\top \mathbf{R}_\star + \lambda \mathbf{I})^{-\frac{1}{2}} \right) (\Sigma_\star + \lambda \mathbf{I})^{\frac{1}{2}} \|_{op}$. Let $\mathbf{A} = \mathbf{R}^\top \mathbf{R} + \lambda \mathbf{I}$

and $B = R_\star^\top R_\star + \lambda I$. We have

$$\begin{aligned}
 & \| (A^{-\frac{1}{2}} - B^{-\frac{1}{2}}) (\Sigma_\star + \lambda I)^{\frac{1}{2}} \|_{op} = \| (A^{-\frac{1}{2}} - B^{-\frac{1}{2}}) (A^{\frac{1}{2}} B^{\frac{1}{2}}) (A^{\frac{1}{2}} B^{\frac{1}{2}})^{-1} (\Sigma_\star + \lambda I)^{\frac{1}{2}} \|_{op} \\
 & = \| (B^{\frac{1}{2}} - A^{-\frac{1}{2}} A^{\frac{1}{2}} B^{\frac{1}{2}}) (A^{\frac{1}{2}} B^{\frac{1}{2}})^{-1} (\Sigma_\star + \lambda I)^{\frac{1}{2}} \|_{op} \\
 & = \| (B^{\frac{1}{2}} - B^{-\frac{1}{2}} A^{\frac{1}{2}} B^{\frac{1}{2}}) B^{-\frac{1}{2}} (A^{-\frac{1}{2}}) (\Sigma_\star + \lambda I)^{\frac{1}{2}} \|_{op} \\
 & = \| (I - B^{-\frac{1}{2}} A^{\frac{1}{2}}) (A^{-\frac{1}{2}}) (\Sigma_\star + \lambda I)^{\frac{1}{2}} \|_{op} \\
 & \leq \| B^{-\frac{1}{2}} (B^{\frac{1}{2}} - A^{\frac{1}{2}}) \|_{op} \| A^{-\frac{1}{2}} (\Sigma_\star + \lambda I)^{\frac{1}{2}} \|_{op} \\
 & = \frac{\sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\lambda^{\frac{1}{4}}})}{1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\lambda^{\frac{1}{4}}})}, \tag{48}
 \end{aligned}$$

where the last inequality follows by using Lemmata A.2.4 and A.2.5.

Next, we bound the term $\| R (R^\top R + \lambda I)^{-\frac{1}{2}} - R_\star (R_\star^\top R_\star + \lambda I)^{-\frac{1}{2}} \|_{op}$. We have

$$\begin{aligned}
 & \| R (R^\top R + \lambda I)^{-\frac{1}{2}} - R_\star (R_\star^\top R_\star + \lambda I)^{-\frac{1}{2}} \|_{op} \\
 & = \| (R - R_\star) (R^\top R + \lambda I)^{-\frac{1}{2}} + R_\star \left((R^\top R + \lambda I)^{-\frac{1}{2}} - (R_\star^\top R_\star + \lambda I)^{-\frac{1}{2}} \right) \|_{op} \\
 & \leq \| (R - R_\star) (R^\top R + \lambda I)^{-\frac{1}{2}} \|_{op} + \| R_\star \left((R^\top R + \lambda I)^{-\frac{1}{2}} - (R_\star^\top R_\star + \lambda I)^{-\frac{1}{2}} \right) \|_{op} \\
 & \leq \| R - R_\star \|_{op} \| (R^\top R + \lambda I)^{-\frac{1}{2}} \|_{op} + \| R_\star \left((R^\top R + \lambda I)^{-\frac{1}{2}} - (R_\star^\top R_\star + \lambda I)^{-\frac{1}{2}} \right) \|_{op} \\
 & \leq \epsilon + \frac{\sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\lambda^{\frac{1}{4}}})}{1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\lambda^{\frac{1}{4}}})}, \tag{49}
 \end{aligned}$$

where we have used the inequality $\| R - R_\star \|_{op} \leq \epsilon\sqrt{\lambda}$. Note that this result holds, since we assumed $\text{dist}(\mathbf{F}, \mathbf{F}_\star) \leq \lambda\epsilon$, hence we can show the following,

$$\| (L - L_\star) (\Sigma_\star + \lambda I)^{\frac{1}{2}} (\Sigma_\star + \lambda I)^{-\frac{1}{2}} \|_{op}^2 \sigma_r^2((\Sigma_\star + \lambda I)^{\frac{1}{2}}) \tag{50}$$

$$\begin{aligned}
 & + \| (R - R_\star) (\Sigma_\star + \lambda I)^{\frac{1}{2}} (\Sigma_\star + \lambda I)^{-\frac{1}{2}} \|_{op}^2 \sigma_r^2((\Sigma_\star + \lambda I)^{\frac{1}{2}}) < \lambda^2 \epsilon^2 \\
 & \rightarrow \| (L - L_\star) (\Sigma_\star + \lambda I)^{\frac{1}{2}} (\Sigma_\star + \lambda I)^{-\frac{1}{2}} \|_{op}^2 \lambda + \| (R - R_\star) (\Sigma_\star + \lambda I)^{\frac{1}{2}} (\Sigma_\star + \lambda I)^{-\frac{1}{2}} \|_{op}^2 \sigma_r^2((\Sigma_\star + \lambda I)^{\frac{1}{2}}) < \lambda^2 \epsilon^2 \\
 & \max(\| L - L_\star \|_{op}, \| R - R_\star \|_{op}) < \sqrt{\lambda} \epsilon. \tag{51}
 \end{aligned}$$

Combining the above bounds we get

$$\begin{aligned}
 & \| R (R^\top R + \lambda I)^{-1} (\Sigma_\star + \lambda I)^{\frac{1}{2}} - R_\star (\Sigma_\star + \lambda I)^{-\frac{1}{2}} \|_{op} \\
 & \leq \epsilon + 2 \frac{\sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\lambda^{\frac{1}{4}}})}{1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\lambda^{\frac{1}{4}}})}. \tag{52}
 \end{aligned}$$

We now focus on bounding $\mathcal{O}_3$,

$$\begin{aligned}
 & \| S R (R^\top R + \lambda I)^{-1} (\Sigma_\star + \lambda I)^{\frac{1}{2}} \|_F^2 \\
 & \leq \underbrace{\| S R (R^\top R + \lambda I)^{-\frac{1}{2}} \|_F^2}_{\mathcal{O}_{3a}} \times \underbrace{\| (R^\top R + \lambda I)^{-\frac{1}{2}} (\Sigma_\star + \lambda I)^{\frac{1}{2}} \|_{op}^2}_{\mathcal{O}_{3b}} \\
 & \leq \| S R (R^\top R + \lambda I)^{-\frac{1}{2}} \|_F^2 \frac{1}{(1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\lambda^{\frac{1}{4}}}))^2}. \tag{53}
 \end{aligned}$$

Using the derived inequalities above we have

$$\begin{aligned} \|(\mathbf{L}_{t+1}\mathbf{Q}_t - \mathbf{L}_\star)(\boldsymbol{\Sigma}_\star + \lambda\mathbf{I})^{\frac{1}{2}}\|_F^2 &\leq \|\Delta_{\mathbf{L}}(\boldsymbol{\Sigma}_\star + \lambda\mathbf{I})^{\frac{1}{2}}\|_F^2 - 2\eta\langle \mathbf{S}, \Delta_{\mathbf{L}}\mathbf{R}_\star^\top + \frac{1}{2}\Delta_{\mathbf{L}}\Delta_{\mathbf{R}}^\top \rangle \\ &\quad + 2\eta L \|\Delta_{\mathbf{L}}(\boldsymbol{\Sigma}_\star + \lambda\mathbf{I})^{\frac{1}{2}}\|_F \left(\frac{3}{2}\epsilon + 2 \frac{\sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}})}{1 - (\sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}}))} \right) \\ &\quad + \eta^2 \frac{1}{(1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}}))^2} \|\mathbf{S}\mathbf{R}(\mathbf{R}^\top\mathbf{R} + \lambda\mathbf{I})\|_F^2. \end{aligned}$$

Following the same steps for bounding the term $\|(\mathbf{R}_{t+1}\mathbf{Q}_t^\top - \mathbf{R}_\star)(\boldsymbol{\Sigma}_\star + \lambda\mathbf{I})^{\frac{1}{2}}\|_F^2$, we have

$$\begin{aligned} &\text{dist}^2(\mathbf{F}_{t+1}, \mathbf{F}_\star) \\ &\leq \|\Delta_{\mathbf{L}}(\boldsymbol{\Sigma}_\star + \lambda\mathbf{I})^{\frac{1}{2}}\|_F^2 + \|\Delta_{\mathbf{R}}(\boldsymbol{\Sigma}_\star + \lambda\mathbf{I})^{\frac{1}{2}}\|_F^2 - 2\eta\langle \mathbf{S}, \mathbf{X} - \mathbf{X}_\star \rangle \\ &\quad + 2\eta L \left(\frac{3}{2}\epsilon + 2 \frac{\sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}})}{1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}}))} \right) \left(\|\Delta_{\mathbf{L}}(\boldsymbol{\Sigma}_\star + \lambda\mathbf{I})^{\frac{1}{2}}\|_F + \|\Delta_{\mathbf{R}}(\boldsymbol{\Sigma}_\star + \lambda\mathbf{I})^{\frac{1}{2}}\|_F \right) \\ &\quad + \eta^2 \frac{1}{(1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}}))^2} \left(\|\mathbf{S}\mathbf{R}(\mathbf{R}^\top\mathbf{R} + \lambda\mathbf{I})^{-\frac{1}{2}}\|_F^2 + \|\mathbf{S}^\top\mathbf{L}(\mathbf{L}^\top\mathbf{L} + \lambda\mathbf{I})^{-\frac{1}{2}}\|_F^2 \right) \\ &\leq \text{dist}^2(\mathbf{F}_t, \mathbf{F}_\star) - 2\eta \left(\tilde{\mathcal{L}}(\mathbf{L}_t\mathbf{R}_t^\top) - \tilde{\mathcal{L}}(\mathbf{X}_\star) \right) + 2\eta L \sqrt{2} \left(\frac{3}{2}\epsilon + 2 \frac{\sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}})}{1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}}))} \right) \text{dist}(\mathbf{F}_t, \mathbf{F}_\star) \\ &\quad + \eta^2 \frac{1}{(1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}}))^2} \left(\|\mathbf{S}\mathbf{R}(\mathbf{R}^\top\mathbf{R} + \lambda\mathbf{I})^{-\frac{1}{2}}\|_F^2 + \|\mathbf{S}^\top\mathbf{L}(\mathbf{L}^\top\mathbf{L} + \lambda\mathbf{I})^{-\frac{1}{2}}\|_F^2 \right). \end{aligned}$$

Convergence with Polyak's stepsize. We use Polyak's stepsize defined as

$$\eta_t = \frac{\tilde{\mathcal{L}}(\mathbf{L}_t\mathbf{R}_t^\top) - \tilde{\mathcal{L}}(\mathbf{X}_\star)}{\|\mathbf{S}\mathbf{R}(\mathbf{R}^\top\mathbf{R} + \lambda\mathbf{I})^{-\frac{1}{2}}\|_F^2 + \|\mathbf{S}^\top\mathbf{L}(\mathbf{L}^\top\mathbf{L} + \lambda\mathbf{I})^{-\frac{1}{2}}\|_F^2}. \quad (54)$$

Due to convexity of $\mathcal{L}(\mathbf{X})$, Lemma A.2.2, and the assumption $\sigma_r(\mathbf{X}_\star) = 1$, we also have,

$$\langle \mathbf{S}, \mathbf{X} - \mathbf{X}_\star \rangle \geq \tilde{\mathcal{L}}(\mathbf{X}) - \tilde{\mathcal{L}}(\mathbf{X}_\star) \geq \mu \|\mathbf{X} - \mathbf{X}_\star\| \geq \mu \sqrt{\frac{(\sqrt{2}-1)}{1+2\lambda}} \text{dist}(\mathbf{F}, \mathbf{F}_\star).$$

We thus get

$$\begin{aligned} \text{dist}^2(\mathbf{F}_{t+1}, \mathbf{F}_\star) &\leq \text{dist}^2(\mathbf{F}_t, \mathbf{F}_\star) - \eta_t \left(2 - \frac{1}{(1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}}))^2} \right) (\tilde{\mathcal{L}}(\mathbf{X}) - \tilde{\mathcal{L}}(\mathbf{X}_\star)) \\ &\quad + 2\sqrt{2}\eta_t L \left(\frac{3}{2}\epsilon + 2 \frac{\sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}})}{1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}}))} \right) \text{dist}(\mathbf{F}_t, \mathbf{F}_\star) \\ &\leq \text{dist}^2(\mathbf{F}_t, \mathbf{F}_\star) - \eta_t \mu \left(\left(2 - \frac{1}{(1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}}))^2} \right) \sqrt{\frac{\sqrt{2}-1}{1+2\lambda}} \right. \\ &\quad \left. - 2\sqrt{2}\frac{L}{\mu} \left(\frac{3}{2}\epsilon + 2 \frac{\sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}})}{1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}^{\frac{1}{4}}}))} \right) \right) \text{dist}(\mathbf{F}_t, \mathbf{F}_\star). \end{aligned}$$

Using (Tong et al., 2021b, Lemma A.2.2 and Lemma 4), we can lower bound the step size as

$$\eta_t \geq \frac{\mu}{2L^2} \sqrt{\frac{\sqrt{2}-1}{1+2\lambda}} \text{dist}(\mathbf{F}, \mathbf{F}_\star).$$

We define $\chi = \frac{L}{\mu}$ and then we get

$$\begin{aligned} & \text{dist}^2(\mathbf{F}_{t+1}, \mathbf{F}_\star) \\ & \leq \left(1 - \frac{1}{2\chi^2} \sqrt{\frac{\sqrt{2}-1}{1+2\lambda}} \left(\sqrt{\frac{\sqrt{2}-1}{1+2\lambda}} \left(2 - \frac{1}{(1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}}^{\frac{1}{4}}))^2} \right) \right. \right. \\ & \quad \left. \left. - 2\chi\sqrt{2} \left(\frac{3}{2}\epsilon + 2 \frac{\sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}}^{\frac{1}{4}})}{1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}}^{\frac{1}{4}})} \right) \right) \right) \text{dist}^2(\mathbf{F}_t, \mathbf{F}_\star). \end{aligned}$$

We define $\rho(\chi, \epsilon, \lambda)$ as the contraction rate given by,

$$\begin{aligned} \rho(\chi, \epsilon, \lambda) &= 1 - \frac{1}{2\chi^2} \sqrt{\frac{\sqrt{2}-1}{1+2\lambda}} \left(\sqrt{\frac{\sqrt{2}-1}{1+2\lambda}} \left(2 - \frac{1}{(1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}}^{\frac{1}{4}}))^2} \right) \right. \\ & \quad \left. - 2\chi\sqrt{2} \left(\frac{3}{2}\epsilon + 2 \frac{\sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}}^{\frac{1}{4}})}{1 - \sqrt{\epsilon}(\sqrt{\epsilon} + \sqrt{2\bar{\lambda}}^{\frac{1}{4}})} \right) \right). \end{aligned}$$

This finishes the proof. $\square$

A.3.1. Proof of Theorem 5.4

Theorem 5.4 Let $\mathcal{L}(\mathbf{X}) : \mathbb{R}^{m \times n} \mapsto \mathbb{R}$ be convex w.r.t $\mathbf{X}$, and assume that it satisfies the rank- d restricted L -Lipschitz continuity assumption and the rank- d restricted μ -sharpness condition, defined above. Assume also that $\lambda = \frac{\|\mathbf{X}_\star\|_{op}}{\bar{\lambda}}$, $\bar{\lambda} = c\|\mathbf{X}_\star\|_{op}$, $c = 20$, and without loss of generality $\sigma_r(\mathbf{X}_\star) = 1$. Set $\epsilon = \frac{10^{-4}}{\chi\|\mathbf{X}_\star\|_{op}^{\frac{1}{2}}}$, which implies.

$$\text{dist}(\mathbf{F}_0, \mathbf{F}_\star) \leq \lambda\epsilon = \frac{10^{-4}}{\chi} \frac{\|\mathbf{X}_\star\|_{op}^{-\frac{1}{2}}}{c}, \quad (55)$$

where $\chi = \frac{L}{\mu}$. Then for the Overparameterized Preconditioned Subgradient Algorithm (OPSA) given in Algorithm 1, with the Polyak's step-size defined in (24), we have,

$$\begin{aligned} \text{dist}(\mathbf{F}_t, \mathbf{F}_\star) &\leq \left(1 - \frac{0.12}{\chi^2} \right)^{\frac{t}{2}} \frac{10^{-4} \|\mathbf{X}_\star\|_{op}^{-\frac{1}{2}}}{c\chi}, \\ \|\mathbf{L}_t \mathbf{R}_t^\top - \mathbf{X}_\star\|_F &\leq \left(1 - \frac{0.12}{\chi^2} \right)^{\frac{t}{2}} \frac{1.5 \times 10^{-4} \|\mathbf{X}_\star\|_{op}^{-\frac{1}{2}}}{c\chi}. \end{aligned}$$

Proof: By using the general expression for the rate $\rho(\chi, \epsilon, \lambda)$ provided in Theorem A.3.1 and for $\lambda = \frac{1}{c}$, and $\epsilon = \frac{10^{-4}}{\|\mathbf{X}_\star\|_{op}^{\frac{1}{2}}\chi}$.

We have,

$$(1 - \rho(\chi, \epsilon, \lambda))\chi^2 \geq 0.12, \quad (56)$$

which leads to the contraction rate showing up in the statement of the Theorem. By applying Lemma A.2.3, we can show the second inequality. $\square$

A.4. More Numerical Experiments

A.4.1. Partially Observed Video Background Subtraction

Video background subtraction with partial observation is one standard benchmark for robust matrix completion. Note that the RIP condition of matrix completion is only defined with the tangent space projection and the incoherence condition. Nevertheless, following the setup as in (Cai et al., 2025), we apply OPSA to the partially observed video background subtraction task on two real video datasets, namely *shoppingmall* and *restaurant*¹. The visual results are reported in Figures 8 and 9, compared against the fully observed RPCA results. Although the rank is overestimated and the observation is partially accessible, OPSA achieves crisp visual results.



Figure 8. Video background subtraction on *shoppingmall* video. Each row corresponds to a frame in the video. The first column is for the original frames. The next two columns are the “groundtruth” foreground and background provided by AccAltProj (Cai et al., 2019), a non-convex RPCA algorithm, with rank 2 and full observation. The last two columns are the foreground and background outputted by the proposed OPSA with $5\times$ overestimated rank $d = 10$ and 30% observation.

¹The datasets were originally provided by http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html; however, it is no longer available. The datasets are now available at <https://hqcai.org/datasets/shoppingmall.mat> and <https://hqcai.org/datasets/restaurant.mat>, respectively.



Figure 9. Video background subtraction on *restaurant* video. Each row corresponds to a frame in the video. The first column is for the original frames. The next two columns are the “groundtruth” foreground and background provided by AccAltProj (Cai et al., 2019), a non-convex RPCA algorithm, with rank 2 and full observation. The last two columns are the foreground and background outputted by the proposed OPISA with $5\times$ overestimated rank $d = 10$ and 30% observation.