The Annals of Statistics
2021, Vol. 49, No. 5, 2948–2971
https://doi.org/10.1214/21-AOS2066

BRIDGING CONVEX AND NONCONVEX OPTIMIZATION IN ROBUST PCA: NOISE, OUTLIERS AND MISSING DATA

BY YUXIN CHEN[1], JIANQING FAN[2,*], CONG MA[3] AND YULING YAN[2,†]

[1]*Department of Electrical and Computer Engineering, Princeton University, yuxin.chen@princeton.edu*
[2]*Department of Operations Research and Financial Engineering, Princeton University, *jqfan@princeton.edu; †yulingy@princeton.edu*
[3]*Department of Statistics, University of Chicago, congm@uchicago.edu*

This paper delivers improved theoretical guarantees for the convex programming approach in low-rank matrix estimation, in the presence of (1) random noise, (2) gross sparse outliers and (3) missing data. This problem, often dubbed as *robust principal component analysis (robust PCA)*, finds applications in various domains. Despite the wide applicability of convex relaxation, the available statistical support (particularly the stability analysis in the presence of random noise) remains highly suboptimal, which we strengthen in this paper. When the unknown matrix is well conditioned, incoherent and of constant rank, we demonstrate that a principled convex program achieves near-optimal statistical accuracy, in terms of both the Euclidean loss and the ℓ_∞ loss. All of this happens even when nearly a constant fraction of observations are corrupted by outliers with arbitrary magnitudes. The key analysis idea lies in bridging the convex program in use and an auxiliary nonconvex optimization algorithm, and hence the title of this paper.

1. Introduction. A diverse array of science and engineering applications (e.g., video surveillance, joint shape matching, graph clustering, covariance modeling, graphical models) involves estimation of low-rank matrices (Candès et al. (2011), Chandrasekaran, Parrilo and Willsky (2012), Chen, Guibas and Huang (2014), Chen et al. (2014), Chi, Lu and Chen (2019), Davenport and Romberg (2016), Fan, Liao and Mincheva (2013)). The imperfectness of data acquisition processes, however, presents several common yet critical challenges: (1) random noise: which reflects the uncertainty of the environment and/or the measurement processes; (2) outliers: which represent a sort of corruption that exhibits abnormal behavior and (3) incomplete data, namely, we might only get to observe a fraction of the matrix entries. Low-rank matrix estimation algorithms aimed at addressing these challenges have been extensively studied under the umbrella of *robust principal component analysis (Robust PCA)* (Candès et al. (2011), Chandrasekaran et al. (2011)), a terminology popularized by the seminal work (Candès et al. (2011)).

To formulate the above mentioned problem in a more precise manner, imagine that we seek to estimate an unknown low-rank matrix $\boldsymbol{L}^\star \in \mathbb{R}^{n_1 \times n_2}$. What we can obtain is a collection of partially observed and corrupted entries as follows:

$$M_{ij} = L^\star_{ij} + S^\star_{ij} + E_{ij}, \quad (i,j) \in \Omega_{\mathsf{obs}}, \tag{1.1}$$

where $\boldsymbol{S}^\star = [S^\star_{ij}]$ is a matrix consisting of outliers, $\boldsymbol{E} = [E_{ij}]$ represents the random noise, and we only observe entries over an index subset $\Omega_{\mathsf{obs}} \subseteq [n_1] \times [n_2]$ with $[n] := \{1, 2, \ldots, n\}$. The current paper assumes that $\boldsymbol{S}^\star$ is a relatively sparse matrix whose nonzero entries might have arbitrary magnitudes. This assumption has been commonly adopted in prior work to

Received January 2020; revised September 2020.
MSC2020 subject classifications. Primary 62F10; secondary 62B10.
Key words and phrases. Robust principal component analysis, convex relaxation, ℓ_∞ guarantees, leave-one-out analysis.

model gross outliers, while enabling reliable disentanglement of the outlier component and the low-rank component (Candès et al. (2011), Chandrasekaran et al. (2011), Chen et al. (2013), Li (2013)). In addition, we suppose that the entries $\{E_{ij}\}$ are independent zero-mean sub-Gaussian random variables, as commonly assumed in the statistics literature to model a large family of random noise. The aim is to reliably estimate $\boldsymbol{L}^\star$ given the grossly corrupted and possibly incomplete data (1.1). Ideally, this task should be accomplished without knowing the locations and magnitudes of the outliers $\boldsymbol{S}^\star$.

1.1. *A principled convex programming approach.* Focusing on the noiseless case with $\boldsymbol{E} = \boldsymbol{0}$, the papers by Candès et al. (2011), Chandrasekaran et al. (2011) delivered a positive and somewhat surprising message: both the low-rank component $\boldsymbol{L}^\star$ and the sparse component $\boldsymbol{S}^\star$ can be efficiently recovered with absolutely no error by means of a principled convex program

$$\underset{\boldsymbol{L},\boldsymbol{S}\in\mathbb{R}^{n_1\times n_2}}{\text{minimize}} \quad \|\boldsymbol{L}\|_* + \tau\|\boldsymbol{S}\|_1 \qquad \text{subject to} \quad \mathcal{P}_{\Omega_{\text{obs}}}(\boldsymbol{L}+\boldsymbol{S}-\boldsymbol{M}) = \boldsymbol{0}, \tag{1.2}$$

provided that certain "separation" and "incoherence" conditions on $(\boldsymbol{L}^\star, \boldsymbol{S}^\star, \Omega_{\text{obs}})$ hold[1] and that the regularization parameter τ is properly chosen. Here, $\|\boldsymbol{L}\|_*$ denotes the nuclear norm (i.e., the sum of the singular values) of $\boldsymbol{L}$, $\|\boldsymbol{S}\|_1 = \sum_{i,j}|S_{ij}|$ denotes the usual entrywise ℓ_1 norm, and $\mathcal{P}_{\Omega_{\text{obs}}}(\boldsymbol{M})$ denotes the Euclidean projection of a matrix $\boldsymbol{M}$ onto the subspace of matrices supported on Ω_{obs}. Given that the nuclear norm $\|\cdot\|_*$ (resp., the ℓ_1 norm $\|\cdot\|_1$) is the convex relaxation of the rank function $\mathsf{rank}(\cdot)$ (resp., the ℓ_0 counting norm $\|\cdot\|_0$), the rationale behind (1.2) is rather clear: it seeks a decomposition $(\boldsymbol{L}, \boldsymbol{S})$ of $\boldsymbol{M}$ by promoting the low-rank structure of $\boldsymbol{L}$ as well as the sparsity structure of $\boldsymbol{S}$.

Moving on to the more realistic noisy setting, a natural strategy is to solve the following regularized least-squares problem:

$$\underset{\boldsymbol{L},\boldsymbol{S}\in\mathbb{R}^{n_1\times n_2}}{\text{minimize}} \quad \frac{1}{2}\big\|\mathcal{P}_{\Omega_{\text{obs}}}(\boldsymbol{L}+\boldsymbol{S}-\boldsymbol{M})\big\|_{\text{F}}^2 + \lambda\|\boldsymbol{L}\|_* + \tau\|\boldsymbol{S}\|_1. \tag{1.3}$$

With the regularization parameters $\lambda, \tau > 0$ properly chosen, one hopes to strike a balance between enhancing the goodness of fit (by enforcing $\boldsymbol{L}+\boldsymbol{S}-\boldsymbol{M}$ to be small) and promoting the desired low-complexity structures (by regularizing both the nuclear norm of $\boldsymbol{L}$ and the ℓ_1 norm of $\boldsymbol{S}$). A natural and important question comes into our mind:

Where does the algorithm (1.3) *stand in terms of its statistical performance in the presence of random noise, sparse outliers and missing data?*

Unfortunately, however simple this program (1.3) might seem, the existing theoretical support remains far from satisfactory, as we shall discuss momentarily.

1.2. *Theory-practice gaps under random noise.* To assess the tightness of prior statistical guarantees for (1.3), we find it convenient to first look at a simple setting where (i) $n_1 = n_2 = n$, (ii) $\boldsymbol{E}$ consists of independent Gaussian components, namely, $E_{ij} \sim \mathcal{N}(0, \sigma^2)$ and (iii) there is no missing data. This simple scenario is sufficient to illustrate the suboptimality of prior theory.

Prior statistical guarantees. The paper Zhou et al. (2010) was the first to derive a sort of statistical performance guarantees for the above convex program. Under mild conditions,

[1]Clearly, if the low-rank matrix $\boldsymbol{L}^\star$ is also sparse, one cannot possibly separate $\boldsymbol{S}^\star$ from $\boldsymbol{L}^\star$. The same holds true if the matrix $\boldsymbol{S}^\star$ is simultaneously sparse and low-rank.

Zhou et al. (2010) demonstrated that any minimizer $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{S}})$ of (1.3) achieves[2]

$$\|\widehat{\boldsymbol{L}} - \boldsymbol{L}^\star\|_{\mathrm{F}} = O\big(n\|\boldsymbol{E}\|_{\mathrm{F}}\big) = O\big(\sigma n^2\big) \tag{1.4}$$

with high probability, where we have substituted in the well-known high-probability bound $\|\boldsymbol{E}\|_{\mathrm{F}} = O(\sigma n)$ under i.i.d. Gaussian noise. While this theory corroborates the potential stability of convex relaxation against both additive noise and sparse outliers, it remains unclear whether the estimation error bound (1.4) reflects the true performance of the convex program in use. In what follows, we shall compare it with an oracle error bound and collect some numerical evidence.

Comparisons with an oracle bound. Consider an idealistic scenario where an oracle informs us of the outlier matrix $\boldsymbol{S}^\star$. With the assistance of this oracle, the task of estimating $\boldsymbol{L}^\star$ reduces to a low-rank matrix denoising problem (Donoho and Gavish (2014)). By fixing $\boldsymbol{S}$ to be $\boldsymbol{S}^\star$ in (1.3), we arrive at a simplified convex program

$$\underset{\boldsymbol{L}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{L} - (\boldsymbol{L}^\star + \boldsymbol{E})\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{L}\|_*. \tag{1.5}$$

It is known that (e.g., Chen et al. (2020a), Donoho and Gavish (2014)): under mild conditions and with a properly chosen λ, the estimation error of (1.5) satisfies

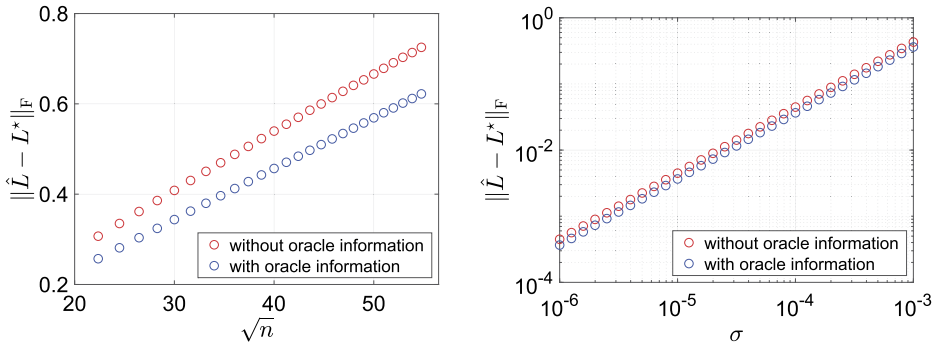
$$\|\widehat{\boldsymbol{L}} - \boldsymbol{L}^\star\|_{\mathrm{F}} = O(\sigma\sqrt{nr}), \tag{1.6}$$

where we abuse the notation and denote by $\widehat{\boldsymbol{L}}$ the minimizer of (1.5). The large gap between the above two bounds (1.4) and (1.6) is self-evident; in particular, if $r = O(1)$, the gap between these two bounds can be as large as an order of $n^{1.5}$.

A numerical example without oracles. One might naturally wonder whether the discrepancy between the two bounds (1.4) and (1.6) stems from the magical oracle information (i.e., $\boldsymbol{S}^\star$) which (1.3) does not have the luxury to know. To demonstrate that this is not the case, we conduct some numerical experiments to assess the importance of such oracle information. Generate $\boldsymbol{L}^\star = \boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}$, where $\boldsymbol{X}^\star, \boldsymbol{Y}^\star \in \mathbb{R}^{n\times r}$ are random orthonormal matrices. Each entry of $\boldsymbol{S}^\star$ is generated independently from a mixed distribution: with probability $1/10$, the entry is drawn from $\mathcal{N}(0, 10)$; otherwise, it is set to be zero. In other words, approximately 10% of the entries in $\boldsymbol{L}^\star$ are corrupted by large outliers. Throughout the experiments, we set $\lambda = 5\sigma\sqrt{n}$ and $\tau = 2\sigma\sqrt{\log n}$ with σ the standard deviation of each noise entry $\{E_{ij}\}$. Figure 1(a) fixes $r = 5$, $\sigma = 10^{-3}$ and examines the dependency of the Euclidean error $\|\widehat{\boldsymbol{L}} - \boldsymbol{L}^\star\|_{\mathrm{F}}$ on the size $\sqrt{n}$. Similarly, Figure 1(b) fixes $r = 5$, $n = 1000$ and displays the estimation error $\|\widehat{\boldsymbol{L}} - \boldsymbol{L}^\star\|_{\mathrm{F}}$ as the noise size σ varies in a log-log plot. As can be seen from Figure 1, the performance of the oracle-aided estimator (1.5) matches the theoretical prediction (1.6), namely, the numerical estimation error $\|\widehat{\boldsymbol{L}} - \boldsymbol{L}^\star\|_{\mathrm{F}}$ is proportional to both $\sqrt{n}$ and σ. Perhaps more intriguingly, even without the help of the oracle, the original estimator (1.3) performs quite well and behaves qualitatively similarly. In comparison with the bound (1.4) derived in the prior work (Zhou et al. (2010)), our numerical experiments suggest that the convex estimator (1.3) might perform much better than previously predicted.

All in all, there seems to be a large gap between the practical performance of (1.3) and the existing theoretical support. This calls for a new theory that better explains practice, which we pursue in the current paper. We remark in passing that statistical guarantees have been developed in Agarwal, Negahban and Wainwright (2012), Klopp, Lounici and Tsybakov (2017) for other convex estimators (i.e., the ones that are different from the convex estimator (1.3) considered herein). We shall compare our results with theirs later in Section 1.4.

[2]Mathematically, Zhou et al. (2010) investigated an equivalent constrained form of (1.3) and developed an upper bound on the corresponding estimation error.



(a) (b)

FIG. 1. *(a) Euclidean estimation errors of* (1.3) *and* (1.5) *versus the problem size* $\sqrt{n}$*, where we fix* $r = 5$*,* $\sigma = 10^{-3}$*; (b) Euclidean estimation errors of* (1.3) *and* (1.5) *versus the noise level* σ *in a log-log plot, where we fix* $n = 1000$*,* $r = 5$*. For both plots, we take* $\lambda = 5\sigma\sqrt{n}$ *and* $\tau = 2\sigma\sqrt{\log n}$*. The results are averaged over* 50 *independent trials.*

1.3. *Models, assumptions and notation.* As it turns out, the appealing empirical performance of the convex program (1.3) in the presence of both sparse outliers and zero-mean random noise can be justified in theory. Toward this end, we need to introduce several notations and model assumptions that will be used throughout. Let $\boldsymbol{U}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}$ be the singular value decomposition (SVD) of the unknown rank-r matrix $\boldsymbol{L}^\star \in \mathbb{R}^{n_1\times n_2}$, where $\boldsymbol{U}^\star \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{V}^\star \in \mathbb{R}^{n_2\times r}$ consist of orthonormal columns and $\boldsymbol{\Sigma}^\star = \mathsf{diag}\{\sigma_1^\star, \ldots, \sigma_r^\star\}$ is a diagonal matrix. Here, we let

$$\sigma_{\max} := \sigma_1^\star \geq \sigma_2^\star \geq \cdots \geq \sigma_r^\star =: \sigma_{\min} \quad \text{and} \quad \kappa := \sigma_{\max}/\sigma_{\min}$$

represent the singular values and the condition number of $\boldsymbol{L}^\star$, respectively. We denote by $\Omega^\star$ the support set of $\boldsymbol{S}^\star$, that is,

$$\Omega^\star := \{(i,j) \in \Omega_{\mathsf{obs}} : S_{ij}^\star \neq 0\}. \tag{1.7}$$

With this set of notation in place, we list below our key model assumptions.

ASSUMPTION 1.1 (Incoherence). The low-rank matrix $\boldsymbol{L}^\star$ with SVD $\boldsymbol{L}^\star = \boldsymbol{U}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}$ is assumed to be μ-incoherent in the sense that

$$\|\boldsymbol{U}^\star\|_{2,\infty} \leq \sqrt{\frac{\mu}{n_1}}\|\boldsymbol{U}^\star\|_F = \sqrt{\frac{\mu r}{n_1}} \quad \text{and} \quad \|\boldsymbol{V}^\star\|_{2,\infty} \leq \sqrt{\frac{\mu}{n_2}}\|\boldsymbol{V}^\star\|_F = \sqrt{\frac{\mu r}{n_2}}. \tag{1.8}$$

Here, $\|\boldsymbol{U}\|_{2,\infty}$ denotes the largest ℓ_2 norm of all rows of a matrix $\boldsymbol{U}$.

ASSUMPTION 1.2 (Random sampling). Each entry is observed independently with probability p, namely,

$$\mathbb{P}\{(i,j) \in \Omega_{\mathsf{obs}}\} = p. \tag{1.9}$$

ASSUMPTION 1.3 (Random locations of outliers). Each observed entry is independently corrupted by an outlier with probability ρ_{s}, namely,

$$\mathbb{P}\{(i,j) \in \Omega^\star \mid (i,j) \in \Omega_{\mathsf{obs}}\} = \rho_{\mathsf{s}}, \tag{1.10}$$

where $\Omega^\star \subseteq \Omega_{\mathsf{obs}}$ is the support of the outlier matrix $\boldsymbol{S}^\star$.

ASSUMPTION 1.4 (Random signs of outliers). The signs of the nonzero entries of $\boldsymbol{S}^\star$ are i.i.d. symmetric Bernoulli random variables (independent from the locations), namely,

$$\operatorname{sign}\big(S^\star_{ij}\big) \overset{\text{ind.}}{=} \begin{cases} 1, & \text{with probability } 1/2, \\ -1, & \text{else,} \end{cases} \qquad \text{for all } (i,j) \in \Omega^\star. \tag{1.11}$$

ASSUMPTION 1.5 (Random noise). The noise matrix $\boldsymbol{E} = [E_{ij}]$ is composed of independent symmetric[3] zero-mean sub-Gaussian random variables with sub-Gaussian norm at most $\sigma > 0$, that is, $\|E_{ij}\|_{\psi_2} \le \sigma$ (see Vershynin (2012), Definition 5.7, for precise definitions).

We take a moment to expand on our model assumptions. Assumption 1.1 is standard in the low-rank matrix recovery literature (Candès and Recht (2009), Candès et al. (2011), Chen (2015), Chi, Lu and Chen (2019)). If μ is small, then this assumption specifies that the singular spaces of $\boldsymbol{L}^\star$ is not sparse in the standard basis, thus ensuring that $\boldsymbol{L}^\star$ is not simultaneously low-rank and sparse. Assumption 1.3 requires the sparsity pattern of the outliers $\boldsymbol{S}^\star$ to be random, which precludes it from being simultaneously sparse and low-rank. In essence, Assumptions 1.1 and 1.3 are identifiability conditions, taken together as a sort of separation condition on $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$, which plays a crucial role in guaranteeing exact recovery in the noiseless case (i.e., $\boldsymbol{E} = \boldsymbol{0}$); see Candès et al. (2011) for more discussions on these conditions. Assumption 1.4 requires the signs of the outliers to be random, which has also been made in Wong and Lee (2017), Zhou et al. (2010).[4] We shall discuss in detail the crucial role of this random sign assumption (as opposed to deterministic sign patterns) in Section 1.6.

Finally, we introduce several notation to be used throughout. Denote by $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ the condition $|f(n)| \le Cg(n)$ for some constant $C > 0$ when n is sufficiently large; we use $f(n) \gtrsim g(n)$ to denote $f(n) \ge C|g(n)|$ for some constant $C > 0$ when n is sufficiently large; we also use $f(n) \asymp g(n)$ to indicate that $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$ hold simultaneously. The notation $f(n) \gg g(n)$ (resp., $f(n) \ll g(n)$) means that there exists a sufficiently large (resp., small) constant $c_1 > 0$ (resp., $c_2 > 0$) such that $f(n) \ge c_1 g(n)$ (resp., $f(n) \le c_2 g(n)$). For any subspace T, we denote by $\mathcal{P}_T(\boldsymbol{M})$ the Euclidean projection of a matrix $\boldsymbol{M}$ onto the subspace T, and let $\mathcal{P}_{T^\perp}(\boldsymbol{M}) := \boldsymbol{M} - \mathcal{P}_T(\boldsymbol{M})$. For any index set Ω, we denote by $\mathcal{P}_\Omega(\boldsymbol{M})$ the Euclidean projection of a matrix $\boldsymbol{M}$ onto the subspace of matrices supported on Ω, and define $\mathcal{P}_{\Omega^c}(\boldsymbol{M}) := \boldsymbol{M} - \mathcal{P}_\Omega(\boldsymbol{M})$. For any matrix $\boldsymbol{M}$, we let $\|\boldsymbol{M}\|$, $\|\boldsymbol{M}\|_{\mathrm{F}}$, $\|\boldsymbol{M}\|_*$, $\|\boldsymbol{M}\|_1$ and $\|\boldsymbol{M}\|_\infty$ denote its spectral norm, Frobenius norm, nuclear norm, entrywise ℓ_1 norm and entrywise ℓ_∞ norm, respectively.

1.4. *Main results.* Armed with the above model assumptions, we are positioned to present our improved statistical guarantees for convex relaxation (1.3) in the random noise setting. Without loss of generality, assume that

$$n_1 \ge n_2.$$

As we shall elucidate in Section 1.5 and Section 3, our theory is established by exploiting an intriguing and intimate connection between convex relaxation and nonconvex optimization, and hence the title of this paper.

For the sake of simplicity, we shall start by presenting our statistical guarantees when the rank r, the condition number κ and the incoherence parameter μ of $\boldsymbol{L}^\star$ are all bounded by

[3] In fact, we only require E_{ij} to be symmetric for all $(i,j) \in \Omega^\star$.

[4] Note that while the theorems in Wong and Lee (2017), Zhou et al. (2010) do not make explicit this random sign assumption, the proofs therein do rely on this assumption to guarantee the existence of certain approximate dual certificates.

some constants. Despite its simplicity, this setting subsumes as special cases a wide array of fundamentally important applications, including angular and phase synchronization (Singer (2011)) in computational biology, joint shape mapping problem (Chen, Guibas and Huang (2014), Huang and Guibas (2013)) in computer vision, and so on. All of these problems involve estimating a very well-conditioned matrix $\boldsymbol{L}^\star$ with a small rank.

THEOREM 1.6. *Suppose that Assumptions 1.1–1.5 hold, and that* $r, \kappa, \mu = O(1)$. *Take* $\lambda = C_\lambda \sigma \sqrt{n_1 p}$ *and* $\tau = C_\tau \sigma \sqrt{\log n_2}$ *in (1.3) for some large enough constants* $C_\lambda, C_\tau > 0$. *Assume that*

$$n_1 n_2 p \geq C_{\mathsf{sample}} n_1 \log^6 n_1, \qquad \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n_1}{p}} \leq \frac{c_{\mathsf{noise}}}{\sqrt{\log n_1}} \quad \text{and} \quad \rho_s \leq \frac{c_{\mathsf{outlier}}}{\log n_1} \tag{1.12}$$

for some sufficiently large constant $C_{\mathsf{sample}} > 0$ *and some sufficiently small constants* $c_{\mathsf{noise}}, c_{\mathsf{outlier}} > 0$. *Then with probability exceeding* $1 - O(n_2^{-3})$, *the following hold*:

1. *Any minimizer* $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$ *of the convex program (1.3) obeys*

$$\left\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\right\|_{\mathrm{F}} \leq C_{\mathsf{err}} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n_1}{p}} \left\|\boldsymbol{L}^\star\right\|_{\mathrm{F}}, \tag{1.13a}$$

$$\left\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\right\|_{\infty} \leq C_{\mathsf{err}} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n_1 \log n_1}{p}} \left\|\boldsymbol{L}^\star\right\|_{\infty}, \tag{1.13b}$$

$$\left\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\right\| \leq C_{\mathsf{err}} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n_1}{p}} \left\|\boldsymbol{L}^\star\right\| \tag{1.13c}$$

for some constant $C_{\mathsf{err}} > 0$.

2. *Letting* $\boldsymbol{L}_{\mathsf{cvx},r} := \arg\min_{\boldsymbol{L}:\mathsf{rank}(\boldsymbol{L}) \leq r} \|\boldsymbol{L} - \boldsymbol{L}_{\mathsf{cvx}}\|_{\mathrm{F}}$ *be the best rank-r approximation of* $\boldsymbol{L}_{\mathsf{cvx}}$, *we have*

$$\left\|\boldsymbol{L}_{\mathsf{cvx},r} - \boldsymbol{L}_{\mathsf{cvx}}\right\|_{\mathrm{F}} \leq \frac{1}{n_2^5} \cdot \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n_1}{p}} \left\|\boldsymbol{L}^\star\right\|_{\mathrm{F}}, \tag{1.14}$$

and the statistical guarantees (1.13) hold unchanged if $\boldsymbol{L}_{\mathsf{cvx}}$ *is replaced by* $\boldsymbol{L}_{\mathsf{cvx},r}$.

Before we embark on interpreting our statistical guarantees, let us first parse the required conditions (1.12) in Theorem 1.6. For simplicity, we assume that $n_1 = n_2 = n$.

- *Missing data*. Theorem 1.6 accommodates the case where a dominant fraction of entries are unobserved (more precisely, the sample size can be as low as an order of $n\,\mathrm{poly}\log n$). This is an appealing result since, even when there is no noise and no outlier (i.e., $\boldsymbol{E} = \boldsymbol{0}$ and $\rho_s = 0$), the minimal sample size required for exact matrix completion is at least on the order of $n \log n$ (Candès and Tao (2010)). In comparison, prior theory on robust PCA with both sparse outliers and dense additive noise is either based on full observations (Agarwal, Negahban and Wainwright (2012), Zhou et al. (2010)), or assumes the sampling rate p exceeds some universal constant (Wong and Lee (2017)). In other words, these prior results require the number of observed entries to exceed the order of n^2. The only exception is Klopp, Lounici and Tsybakov (2017), which also allows a significant amount of missing data (i.e., $p \gtrsim (\mathrm{poly}\log n)/n$).
- *Noise levels*. The noise condition, namely $\sigma\sqrt{n \log n / p} \lesssim \sigma_{\min}$, accommodates a wide range of noise levels. To see this, it is straightforward to check that this noise condition is equivalent to

$$\sigma \lesssim \sqrt{\frac{np}{\log n}} \left\|\boldsymbol{L}^\star\right\|_{\infty}$$

as long as $r, \mu, \kappa \asymp 1$. In other words, the entrywise noise level σ is allowed to be significantly larger than the maximum magnitude of the entries in the low-rank matrix $\boldsymbol{L}^\star$, as long as $p \gg (\log n)/n$.

- *Tolerable fraction of outliers*. The above theorem assumes that no more than a fraction $\rho_s \lesssim 1/\log n$ of observations are corrupted by outliers. In words, our theory allows *nearly* a constant proportion (up to a logarithmic order) of the entries of $\boldsymbol{L}^\star$ to be corrupted with arbitrary magnitudes.

Next, we move on to the interpretation of our statistical guarantees. Note that we still assume that $n_1 = n_2 = n$ for ease of presentation.

- *Near-optimal statistical guarantees*. Our first result (1.13a) gives an Euclidean estimation error bound of (1.3)

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}. \tag{1.15}$$

This cannot be improved even when an oracle has informed us of the outliers $\boldsymbol{S}^\star$ and the tangent space of $\boldsymbol{L}^\star$; see Candès and Plan (2010), Section III.B. We remark that under similar model assumptions, the paper Wong and Lee (2017) derived an estimation error bound for a constrained version of the convex program (1.3), which asserts that this convex estimator $\widetilde{\boldsymbol{L}}_{\mathsf{cvx}}$ satisfies[5]

$$\|\widetilde{\boldsymbol{L}}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}} \lesssim \sigma n^{1.5}, \tag{1.16}$$

with the proviso that p is at least on the constant order. The restriction on p arises from the dual certificate constructed in Candès et al. (2011), which is also used in the proof of Theorem 4 in Wong and Lee (2017). While this is suboptimal compared to our results in the setting considered herein, it is worth pointing out that the bound therein accommodates arbitrary noise matrix $\boldsymbol{E}$ (e.g., deterministic, adversary), and here in (1.16) we specialize their result to the random noise setting, namely the noise $\boldsymbol{E}$ obeys Assumption 1.5. In addition, under the full observation (i.e., $p = 1$) setting, the paper Agarwal, Negahban and Wainwright (2012) derived an estimation error bound for a convex program similar to (1.3), but with an additional constraint regularizing the spikiness of the low-rank component. Note that instead of imposing the incoherence condition as in Assumption 1.1, the prior work Agarwal, Negahban and Wainwright (2012) assumes a milder spikiness condition on $\boldsymbol{L}^\star$, which only constrains the maximum entry in the matrix $\boldsymbol{L}^\star$ is not too large. When $\{E_{ij}\}$ are i.i.d. drawn from $\mathcal{N}(0, \sigma^2)$ and when there is no missing data (i.e. $p = 1$), the Euclidean estimation error bound achievable by their estimator $\boldsymbol{L}_{\mathsf{cvx}}^{\mathsf{ANW}}$ reads

$$\|\boldsymbol{L}_{\mathsf{cvx}}^{\mathsf{ANW}} - \boldsymbol{L}^\star\|_{\mathrm{F}} \lesssim \sigma\sqrt{n}\max\{1, \sqrt{n\rho_{\mathsf{s}}\log n}\} + \|\boldsymbol{L}^\star\|_\infty n\sqrt{\rho_{\mathsf{s}}}, \tag{1.17}$$

which is suboptimal compared to our results. In particular, (i) the bound (1.17) does not vanish even as the noise level decreases to zero, and (ii) it becomes looser as ρ_{s} grows (e.g., if $\rho_{\mathsf{s}} \asymp 1/\log n$, the bound (1.17) is $O(\sqrt{n})$ larger than our bound). Moreover, the work Agarwal, Negahban and Wainwright (2012) did not account for missing data. Similar to Agarwal, Negahban and Wainwright (2012) (but with an additional spikiness condition on $\boldsymbol{S}^\star$), the paper Klopp, Lounici and Tsybakov (2017) derived an estimation error bound

[5]More specifically, Wong and Lee (2017), Theorem 4, studies the following convex program $\text{minimize}_{\boldsymbol{L},\boldsymbol{S}\in\mathbb{R}^{n\times n}} \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1$ s.t. $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}+\boldsymbol{S}-\boldsymbol{M})\|_{\mathrm{F}} \le \delta$. Here, the quantity δ needs to be larger than $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}+\boldsymbol{S}-\boldsymbol{M})\|_{\mathrm{F}}$. Under our setting, the minimum level of δ should be a high-probability upper bound on $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}}$, which is on the order of $\sigma n\sqrt{p}$. With this choice of δ, Wong and Lee (2017), Theorem 4, yields $\|\widetilde{\boldsymbol{L}}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}} \le [2 + 8\sqrt{n}(1+\sqrt{8/p})]\delta \lesssim \sigma n^{1.5}$.

TABLE 1
Comparison of our statistical guarantee and prior theory

	Euclidean estimation error	Accounting for missing data
Zhou et al. (2010)	σn^2	no
Agarwal, Negahban and Wainwright (2012)	$\sigma\sqrt{n}\max\{\sqrt{r}, \sqrt{n\rho_{\mathrm{s}}\log n}\}$ $+ \|\boldsymbol{L}^\star\|_\infty n\sqrt{\rho_{\mathrm{s}}}$	no
Wong and Lee (2017)	$\sigma n^{1.5}$	yes ($p \gtrsim 1$)
Klopp, Lounici and Tsybakov (2017)	$\max\{\sigma, \|\boldsymbol{L}^\star\|_\infty, \|\boldsymbol{S}^\star\|_\infty\}\cdot$ $\sqrt{(n\log n)/p}\max\{1, \sqrt{np\rho_{\mathrm{s}}}\}$	yes ($p \gtrsim (\mathrm{poly}\log n)/n$)
This paper	$\sigma\sqrt{nr/p}$	yes ($p \gtrsim \kappa^4\mu^2 r^2(\mathrm{poly}\log n)/n$)

for a constrained convex program, with a new constraint regularizing the spikiness of the sparse outliers. Their Euclidean estimation error bound reads

$$\|\boldsymbol{L}_{\mathsf{cvx}}^{\mathsf{KLT}} - \boldsymbol{L}^\star\|_{\mathrm{F}} \lesssim \max\{\sigma, \|\boldsymbol{L}^\star\|_\infty, \|\boldsymbol{S}^\star\|_\infty\}\sqrt{\frac{n\log n}{p}}\max\{1, \sqrt{np\rho_{\mathrm{s}}}\}, \tag{1.18}$$

which is also suboptimal compared to our results. In particular, (1) their error bound degrades as the magnitude $\|\boldsymbol{S}^\star\|_\infty$ of the outlier increases; (2) when there is no missing data (i.e., $p = 1$), their bound might be off by a factor as large as $O(\sqrt{n})$. It is worth emphasizing that the theory developed in these prior works is developed to accommodate a broader range of matrices. For example, both Agarwal, Negahban and Wainwright (2012) and Klopp, Lounici and Tsybakov (2017) study the set of entrywise bounded low-rank matrices (without assuming the incoherence condition); Agarwal, Negahban and Wainwright (2012) even allows $\boldsymbol{L}^\star$ to be approximately low rank. To ease comparison, Table 1 displays a summary of our results versus prior statistical guarantees when specialized to the settings considered herein.

- *Entrywise and spectral norm error control*. Moving beyond Euclidean estimation errors, our theory also provides statistical guarantees measured by two other important metrics: the entrywise ℓ_∞ norm (cf. (1.13b)) and the spectral norm (cf. (1.13c)). In particular, our entrywise error bound (1.13b) in reads

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_\infty \lesssim \sigma\sqrt{\frac{\log n}{np}} \tag{1.19}$$

as long as $r, \kappa, \mu \asymp 1$, which is about $O(n)$ times small than the Euclidean loss (1.15) modulo some logarithmic factor. This uncovers an appealing "delocalization" behavior of the estimation errors, namely, the estimation errors of $\boldsymbol{L}^\star$ are fairly spread out across all entries. This can also be viewed as an "implicit regularization" phenomenon: the convex program automatically controls the spikiness of the low-rank solution, without the need of explicitly regularizing it (e.g., adding a constraint $\|\boldsymbol{L}\|_\infty \le \alpha$ as adopted in the prior work Agarwal, Negahban and Wainwright (2012), Klopp, Lounici and Tsybakov (2017)). See Figure 2 for the numerical evidence for the relative entrywise and spectral norm error of $\boldsymbol{L}_{\mathsf{cvx}}$.
- *Approximate low-rank structure of the convex estimator* $\boldsymbol{L}_{\mathsf{cvx}}$. Last but not least, Theorem 1.6 ensures that the convex estimate $\boldsymbol{L}_{\mathsf{cvx}}$ is nearly rank-r, so that a rank-r approximation of $\boldsymbol{L}_{\mathsf{cvx}}$ is extremely accurate. In other words, the convex program automatically

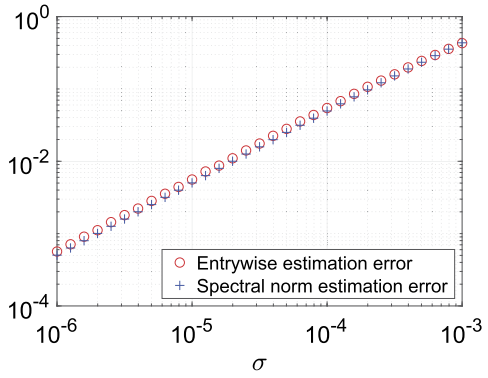


FIG. 2. *The relative estimation error of* $\boldsymbol{L}_{\mathsf{cvx}}$ *measured by both* $\|\cdot\|_\infty$ (*i.e.*, $\|\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^\star\|_\infty/\|\boldsymbol{L}^\star\|_\infty$) *and* $\|\cdot\|$ (*i.e.*, $\|\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^\star\|/\|\boldsymbol{L}^\star\|$) *versus the standard deviation* σ *of the noise in a log-log plot. The results are reported for* $n=1000$, $r=5$, $p=0.2$, $\rho_s=0.1$, $\lambda=5\sigma\sqrt{np}$, $\tau=2\sigma\sqrt{\log n}$, *and are averaged over* 50 *independent trials. In addition, the data generating process is similar to that in Figure* 1.

adapts to the true rank of $\boldsymbol{L}^\star$ without having any prior knowledge about r. As we shall see shortly, this is a crucial observation underlying the intimate connection between convex relaxation and a certain nonconvex approach.

Moving beyond the setting with $r,\kappa,\mu\asymp 1$, we have developed theoretical guarantees that allow r,κ,μ to grow with the problem dimension n_1,n_2. The result is this.

THEOREM 1.7. *Suppose that Assumptions* 1.1–1.5 *hold and that* $n_1\ge n_2$. *Take* $\lambda=C_\lambda\sigma\sqrt{n_1 p}$ *and* $\tau=C_\tau\sigma\sqrt{\log n_2}$ *in* (1.3) *for some large enough constants* $C_\lambda, C_\tau>0$. *Assume that*

$$n_1 n_2 p \ge C_{\mathsf{sample}}\kappa^4\mu^2 r^2 n_1\log^6 n_1,$$

$$\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n_1}{p}}\le\frac{c_{\mathsf{noise}}}{\sqrt{\kappa^4\mu r\log n_1}},\quad \text{and}\quad \rho_s\le\frac{c_{\mathsf{outlier}}}{\kappa^3\mu r\log n_1} \tag{1.20}$$

for some sufficiently large constant $C_{\mathsf{sample}}>0$ *and some sufficiently small constants* $c_{\mathsf{noise}}, c_{\mathsf{outlier}}>0$. *Then with probability exceeding* $1-O(n_2^{-3})$, *the following hold*:

1. *Any minimizer* $(\boldsymbol{L}_{\mathsf{cvx}},\boldsymbol{S}_{\mathsf{cvx}})$ *of the convex program* (1.3) *obeys*

$$\|\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^\star\|_{\mathrm{F}}\le C_{\mathsf{err}}\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n_1}{p}}\|\boldsymbol{L}^\star\|_{\mathrm{F}}, \tag{1.21a}$$

$$\|\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^\star\|_{\infty}\le C_{\mathsf{err}}\sqrt{\kappa^3\mu r}\cdot\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n_1\log n_1}{p}}\|\boldsymbol{L}^\star\|_{\infty}, \tag{1.21b}$$

$$\|\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^\star\|\le C_{\mathsf{err}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n_1}{p}}\|\boldsymbol{L}^\star\| \tag{1.21c}$$

for some constant $C_{\mathsf{err}}>0$.

2. *Letting* $\boldsymbol{L}_{\mathsf{cvx},r}:=\arg\min_{\boldsymbol{L}:\mathsf{rank}(\boldsymbol{L})\le r}\|\boldsymbol{L}-\boldsymbol{L}_{\mathsf{cvx}}\|_{\mathrm{F}}$ *be the best rank-*r *approximation of* $\boldsymbol{L}_{\mathsf{cvx}}$, *we have*

$$\|\boldsymbol{L}_{\mathsf{cvx},r}-\boldsymbol{L}_{\mathsf{cvx}}\|_{\mathrm{F}}\le\frac{1}{n_2^5}\cdot\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n_1}{p}}\|\boldsymbol{L}^\star\|_{\mathrm{F}}, \tag{1.22}$$

and the statistical guarantees (1.21) *hold unchanged if* $\boldsymbol{L}_{\mathsf{cvx}}$ *is replaced by* $\boldsymbol{L}_{\mathsf{cvx},r}$.

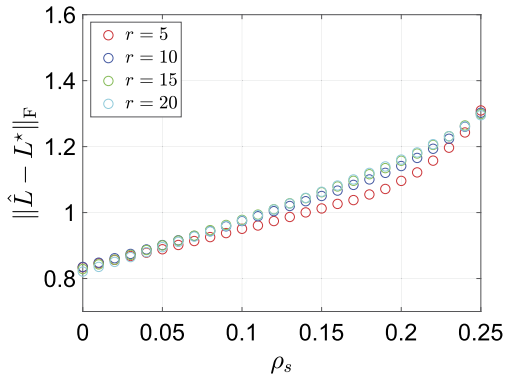


FIG. 3. *Euclidean estimation errors of* (1.3) *versus* ρ_{s} *under four different ranks* $r = 5, 10, 15, 20$. *The results are reported for* $n = 1000$, $p = 0.04r$, $\sigma = 10^{-3}$, $\lambda = 5\sigma\sqrt{np}$, $\tau = 2\sigma\sqrt{\log n}$, *and are averaged over* 50 *independent trials. In addition, the data generating process is similar to that in Figure* 1.

Similar to Theorem 1.6, our general theory (i.e., Theorem 1.7) provides the estimation error of the convex estimator $\boldsymbol{L}_{\mathsf{cvx}}$ in three different norms (i.e., the Euclidean, entrywise and operator norms), and reveals the near low-rankness of the convex estimator (cf. (1.22)) as well as the implicit regularization phenomenon (cf. (1.21b)).

Finally, we make note of several aspects of our general theory that call for further improvement. For instance, when there is no missing data and $n_1 = n_2 = n$, the rank r of the unknown matrix $\boldsymbol{L}^\star$ needs to satisfy $r \lesssim \sqrt{n}$. On the positive side, our result allows r to grow with the problem dimension n. However, prior results in the noiseless case (Candès et al. (2011), Li (2013)) allow r to grow almost linearly with n. This unsatisfactory aspect arises from the suboptimal analysis (in terms of the dependency on r) of a tightly related nonconvex estimation algorithm (to be elaborated on later), which to the best of our knowledge, has not been resolved in the nonconvex low-rank matrix recovery literature (Chen, Liu and Li (2020), Ma et al. (2020)). See Section 2 for more discussions about this point. Moreover, when $\boldsymbol{E} = \boldsymbol{0}$, it is known that ρ_s can be as large as a constant even when the rank r is allowed to grow with the dimension n (Chen et al. (2013), Li (2013)). Our current theory, however, fails to cover the case with $\rho_s \asymp 1$ in the presence of noise. We demonstrate through numerical experiments that the dependence of ρ_{s} on r might indeed by suboptimal in our current theory. More specifically, Figure 3 depicts the numerical Euclidean estimation errors w.r.t. the corruption probability ρ_{s} as we vary the rank while fixing the sampling ratio. It can be seen that the estimation error curves corresponding to different ranks align very well with each other, thus suggesting the capability of convex relaxation in tolerating a constant fraction ρ_{s} of outliers.

1.5. *A peek at our technical approach.* Before delving into the proof details, we immediately highlight our key technical ideas and novelties. For simplicity, we assume $n_1 = n_2 = n$ throughout this section.

Connections between convex and nonconvex optimization. Instead of directly analyzing the convex program (1.3), we turn attention to a seemingly different, but in fact closely related, nonconvex program

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r},\boldsymbol{S}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \frac{1}{2}\big\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{S} - \boldsymbol{M})\big\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\big(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2\big) + \tau\|\boldsymbol{S}\|_1. \tag{1.23}$$

This idea is inspired by an interesting numerical finding (cf. Figure 4) that the solution to the convex program (1.3), and an estimate obtained by attempting to solve the nonconvex formulation (1.23), are exceedingly close in our experiments. If such an intimate connection

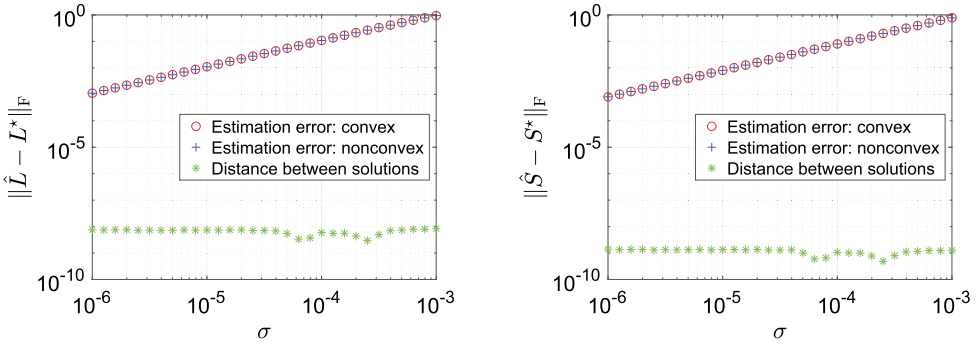


FIG. 4. (*a*) *The relative estimation errors of both* $\boldsymbol{L}_{\mathsf{cvx}}$ *(the convex estimator* ((1.3))*) and* $\boldsymbol{L}_{\mathsf{ncvx}}$ *(the estimate returned by the nonconvex approach tailored to* ((1.23))*) and the relative distance between them versus the standard deviation* σ *of the noise.* (*b*) *The relative estimation errors of both* $\boldsymbol{S}_{\mathsf{cvx}}$ *(the convex estimator in* ((1.3))*) and* $\boldsymbol{S}_{\mathsf{ncvx}}$ *(the estimate returned by the nonconvex approach tailored to* ((1.23))*) and the relative distance between them versus the standard deviation* σ *of the noise. The results are reported for* $n = 1000$, $r = 5$, $p = 0.2$, $\rho_s = 0.1$, $\lambda = 5\sigma\sqrt{np}$, $\tau = 2\sigma\sqrt{\log n}$ *and are averaged over* 50 *independent trials.*

can be formalized, then it suffices to analyze the statistical performance of the nonconvex approach instead.[6] Fortunately, recent advances in nonconvex low-rank factorization (see Chi, Lu and Chen (2019) for an overview) provide powerful tools for analyzing nonconvex low-rank estimation, allowing us to derive the desired statistical guarantees that can then be transferred to the convex approach. Of course, this is merely a high-level picture of our proof strategy, and we defer the details to Section 3.

It is worth emphasizing that our key idea—that is, bridging convex and nonconvex optimization—is drastically different from previous technical approaches for analyzing convex estimators (e.g., (1.3)). As it turns out, these prior approaches, which include constructing dual certificates and/or exploiting restricted strong convexity, have their own deficiencies in analyzing (1.3) and fall short of explaining the effectiveness of (1.3) in the random noise setting. For instance, constructing dual certificates in the noisy case is notoriously challenging given that we do not have closed-form expressions for the primal solutions (so that it is difficult to invoke the powerful dual construction strategies like the golfing scheme (Gross (2011)) developed for the noiseless case). If we directly utilize the dual certificates constructed for the noiseless case, we would end up with an overly conservative bound like (1.4), which is exactly why the results in Wong and Lee (2017), Zhou et al. (2010) are suboptimal. On the other hand, while it is viable to show certain strong convexity of (1.3) when restricted to some highly local sets and directions, it is unclear how (1.3) forces its solution to stay within the desired set and follow the desired directions, without adding further (and often unnecessary) constraints to (1.3).

Nonconvex low-rank estimation with nonsmooth loss functions. It is worth noting that a similar connection between convex and nonconvex optimization has been pointed out by Chen et al. (2020a) toward understanding the power of convex relaxation for noisy matrix completion. Due to the absence of sparse outliers in the noisy matrix completion problem, the nonconvex loss function considered therein is smooth in nature, which greatly simplifies

[6]On the surface, the convex program (1.3) and the nonconvex one (1.23) are closely related: the convex solution $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$ coincides with that of the nonconvex program (1.23) if $\boldsymbol{L}_{\mathsf{cvx}}$ is rank-r. This is an immediate consequence of the algebraic identity $\|\boldsymbol{Z}\|_* = \inf_{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}:\boldsymbol{X}\boldsymbol{Y}^\top=\boldsymbol{Z}}(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2)$ (Mazumder, Hastie and Tibshirani (2010), Srebro and Shraibman (2005)). However, it is difficult to know *a priori* the rank of the convex solution. Hence such a connection does not prove useful in establishing the statistical properties of the convex estimator.

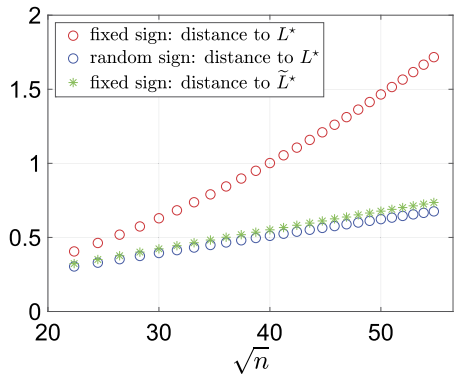


FIG. 5. *The red (resp., blue) line displays the Euclidean estimation error of* ((1.3)) *versus* $\sqrt{n}$ *under fixed (resp., random) sign patterns of* $\boldsymbol{S}^\star$. *The green line displays the Euclidean distance between* $\boldsymbol{L}_{\mathsf{cvx}}$ *and* $\widetilde{\boldsymbol{L}}^\star$ *under fixed sign patterns of* $\boldsymbol{S}^\star$. *The results are reported for* $r = 5$, $p = 1$, $\sigma = 10^{-3}$ *and* $\rho_{\mathsf{s}} = 1/\log n$, *with* $\lambda = 5\sigma\sqrt{np}$ *and* $\tau = 2\sigma\sqrt{\log n}$ *and are averaged over* 50 *independent trials. For the random sign setting, the nonzero entries of* $\boldsymbol{S}^\star$ *are independently generated as* $z \cdot 5\sigma$, *where* z *follows a Rademacher distribution. For the fixed sign setting, each nonzero entry of* $\boldsymbol{S}^\star$ *equals to* 5σ.

both the algorithmic and theoretical development. By contrast, the nonsmoothness inherent in (1.23) makes it particularly challenging to achieve the two desiderata mentioned above, namely, connecting the convex and nonconvex solutions and establishing the optimality of the nonconvex solution. In fact, to establish the connection between convex and nonconvex solutions, we put forward a novel two-step analysis strategy. Specifically, we first develop a crude upper bound on the Euclidean estimation error leveraging the idea of approximate dual certificates; see Theorem 3.1. While this crude upper bound is far from optimal, it serves as an important starting point toward formalizing the intimate relation between the convex solution $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$ and the nonconvex solution $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S})$, since it is challenging to establish $\boldsymbol{XY} \approx \boldsymbol{L}_{\mathsf{cvx}}$ and $\boldsymbol{S} \approx \boldsymbol{S}_{\mathsf{cvx}}$ simultaneously without the aid of a crude bound. Second, in establishing the optimality of the nonconvex solution, the nonsmoothness nature of the nonconvex loss prevents us from applying the vanilla gradient descent scheme (as has been done in Chen et al. (2020a). To address this issue, we develop an alternating minimization scheme—which alternates between gradient updates on $(\boldsymbol{X}, \boldsymbol{Y})$ and minimization of $\boldsymbol{S}$—aimed at minimizing the nonsmooth nonconvex loss function (1.23); see Algorithm 1 for details. As it turns out, such a simple algorithm allows us to track the proximity of the convex and nonconvex solutions and establish the optimality of the nonconvex solution all at once.

1.6. *Random signs of outliers.* The careful reader might wonder whether it is possible to remove the random sign assumption on $\boldsymbol{S}^\star$ (namely, Assumption 1.4) without compromising our statistical guarantees. After all, the results of Candès et al. (2011), Chandrasekaran et al. (2011), Li (2013) derived for the noise-free case do not rely on such a random sign assumption at all.[7] Unfortunately, removal of such a condition might be problematic in general, as illustrated by the following example.

An example with nonrandom signs. Suppose that (i) $n_1 = n_2 = n$, (ii) each nonzero entry of $\boldsymbol{S}^\star$ obeys $S^\star_{ij} = c_0\sigma$, (iii) $\rho_{\mathsf{s}} = c_1/\log n$ for some sufficiently small constant $c_1 > 0$, and (iv) there is no missing data (i.e., $p = 1$). In such a scenario, the data matrix can be decomposed

[7]Notably, in the noisy setting, prior theory (Wong and Lee (2017), Zhou et al. (2010)) also implicitly assumes this random sign condition, while Agarwal, Negahban and Wainwright (2012), Klopp, Lounici and Tsybakov (2017) do not require this condition.

as

$$\boldsymbol{M} = \boldsymbol{L}^\star + \boldsymbol{S}^\star + \boldsymbol{E} = \underbrace{\boldsymbol{L}^\star + \mathbb{E}[\boldsymbol{S}^\star]}_{=:\widetilde{\boldsymbol{L}}^\star} + \underbrace{\boldsymbol{S}^\star - \mathbb{E}[\boldsymbol{S}^\star] + \boldsymbol{E}}_{=:\widetilde{\boldsymbol{E}}}.$$

Two observations are worth noting: (1) given that $\mathbb{E}[\boldsymbol{S}^\star] = c_0\rho_{\mathsf{s}}\sigma\mathbf{1}\mathbf{1}^\top$ with $\mathbf{1}$ the all-one vector, the rank of the matrix $\widetilde{\boldsymbol{L}}^\star = \boldsymbol{L}^\star + \mathbb{E}[\boldsymbol{S}^\star]$ is at most $r+1$; (2) $\widetilde{\boldsymbol{E}}$ is a zero-mean random matrix consisting of independent entries with sub-Gaussian norm $O(\sigma)$. In other words, the decomposition $\boldsymbol{M} = \widetilde{\boldsymbol{L}}^\star + \widetilde{\boldsymbol{E}}$ corresponds to a case with random noise but no outliers. Consequently, we can invoke Theorem 1.6 to conclude that (assuming $r = O(1)$ and $\widetilde{\boldsymbol{L}}^\star$ is incoherent with condition number $O(1)$): any minimizer $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$ of (1.3) obeys

$$\begin{aligned}\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star - \rho_{\mathsf{s}}\sigma\mathbf{1}\mathbf{1}^\top\|_{\mathrm{F}} = \|\boldsymbol{L}_{\mathsf{cvx}} - \widetilde{\boldsymbol{L}}^\star\|_{\mathrm{F}} &\lesssim \frac{\sigma}{\sigma_{\min}(\widetilde{\boldsymbol{L}}^\star)}\sqrt{n}\|\widetilde{\boldsymbol{L}}^\star\|_{\mathrm{F}} \\ &\lesssim \sigma\sqrt{nr}\frac{\sigma_{\max}(\widetilde{\boldsymbol{L}}^\star)}{\sigma_{\min}(\widetilde{\boldsymbol{L}}^\star)} \lesssim \sigma\sqrt{n}\end{aligned}$$

with high probability. Here, the last step follows since $\widetilde{\boldsymbol{L}}^\star$ is of constant rank and condition number. This, however, leads to a lower bound on the estimation error

$$\begin{aligned}\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}} &\geq \|c_0\rho_{\mathsf{s}}\sigma\mathbf{1}\mathbf{1}^\top\|_{\mathrm{F}} - \|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star - \rho_{\mathsf{s}}\sigma\mathbf{1}\mathbf{1}^\top\|_{\mathrm{F}} = \sigma\big(c_0\rho_{\mathsf{s}}n - O(\sqrt{n})\big) \\ &= \big(1 - o(1)\big)\frac{c_0c_1\sigma n}{\log n},\end{aligned}$$

which can be $O(\sqrt{n}/\log n)$ times larger than the desired estimation error $O(\sigma\sqrt{n})$. Numerical experiments under the above setting (with $c_0 = 5$ and $c_1 = 1$) also suggest that (i) the estimation error under the fixed sign setting might be orderwise larger than that under the random sign setting; and (ii) under the fixed sign setting, the estimator (1.3) approximately recovers $\widetilde{\boldsymbol{L}}^\star$ instead of $\boldsymbol{L}^\star$; see Figure 5.

The take-away message is this: when the entries of $\boldsymbol{S}^\star$ are of nonrandom signs, it might sometimes be possible to decompose $\boldsymbol{S}^\star$ into (1) a low-rank bias component with a large Euclidean norm, and (2) a random fluctuation component whose typical size does not exceed that of $\boldsymbol{E}$. If this is the case, then the convex program (1.3) might mistakenly treat the bias component as a part of the low-rank matrix $\boldsymbol{L}^\star$, thus dramatically hampering its estimation accuracy.

2. Prior art. Principal component analysis (PCA) (Fan et al. (2018), Jolliffe (1986), Pearson (1901)) is one of the most widely used statistical methods for dimension reduction in data analysis. However, PCA is known to be quite sensitive to adversarial outliers—even a single corrupted data point can make PCA completely off. This motivated the investigation of robust PCA, which aims at making PCA robust to gross adversarial outliers. As formulated in Candès et al. (2011), Chandrasekaran et al. (2011), this is closely related to the problem of disentangling a low-rank matrix $\boldsymbol{L}^\star$ and a sparse outlier matrix $\boldsymbol{S}^\star$ (with unknown locations and magnitudes) from a superposition of them. Consequently, robust PCA can be viewed as an outlier-robust extension of the low-rank matrix estimation/completion tasks (Candès and Recht (2009), Chi, Lu and Chen (2019), Keshavan, Montanari and Oh (2010)). In a similar vein, robust PCA has also been extensively studied in the context of structured covariance estimation under approximate factor models (Fan, Fan and Lv (2008), Fan, Liao and Mincheva (2013), Fan, Wang and Zhong (2017), Fan, Wang and Zhong (2019)), where the population covariance of certain random sample vectors is a mixture of a low-rank matrix and a sparse matrix, corresponding to the factor component and the idiosyncratic component, respectively.

Focusing on the convex relaxation approach, Candès et al. (2011), Chandrasekaran et al. (2011) started by considering the noiseless case with no missing data (i.e., $\boldsymbol{E} = \boldsymbol{0}$ and $p = 1$) and demonstrated that, under mild conditions, convex relaxation succeeds in exactly decomposing both $\boldsymbol{L}^\star$ and $\boldsymbol{S}^\star$ from the data matrix $\boldsymbol{L}^\star + \boldsymbol{S}^\star$. More specifically, Chandrasekaran et al. (2011) adopted a deterministic model without assuming any probabilistic structure on the outlier matrix $\boldsymbol{S}^\star$. As shown in Chandrasekaran et al. (2011) and several subsequent works (Chen et al. (2013), Hsu, Kakade and Zhang (2011)), convex relaxation is guaranteed to work as long as the fraction of outliers in each row/column does not exceed $O(1/r)$. In contrast, Candès et al. (2011) proposed a random model by assuming that $\boldsymbol{S}^\star$ has random support (cf. Assumption 1.3); under this model, exact recovery is guaranteed even if a constant fraction of the entries of $\boldsymbol{S}^\star$ are nonzero with arbitrary magnitudes. Following the random location model proposed in Candès et al. (2011), the paper Ganesh et al. (2010) showed that, in the absence of noise, convex programming can provably tolerate a dominant fraction of outliers, provided that the signs of the nonzero entries of $\boldsymbol{S}^\star$ are randomly generated (cf. Assumption 1.4). Later, the papers Chen et al. (2013), Li (2013) extended these results to the case when most entries of the matrix are unseen; even in the presence of highly incomplete data, convex relaxation still succeeds when a constant proportion of the observed entries are arbitrarily corrupted. It is worth noting that the results of Chen et al. (2013) accommodated both models proposed in Chandrasekaran et al. (2011) and Candès et al. (2011), while the results of Li (2013) focused on the latter model.

The literature on robust PCA with not only sparse outliers but also dense noise—namely, when the measurements take the form $\boldsymbol{M} = \mathcal{P}_{\Omega_{\text{obs}}}(\boldsymbol{L}^\star + \boldsymbol{S}^\star + \boldsymbol{E})$ —is relatively scarce. Agarwal, Negahban and Wainwright (2012), Zhou et al. (2010) were among the first to present a general theory for robust PCA with dense noise, which was further extended in Klopp, Lounici and Tsybakov (2017), Wong and Lee (2017). As we mentioned before, the first three (Agarwal, Negahban and Wainwright (2012), Wong and Lee (2017), Zhou et al. (2010)) accommodated arbitrary noise with the last one (Klopp, Lounici and Tsybakov (2017)) focusing on the random noise. As we have discussed in Section 1.4, the statistical guarantees provided in these papers are highly suboptimal when it comes to the random noise setting considered herein. The paper Chen and Chi (2014) extended the robust PCA results to the case where the truth is not only low-rank but also of Hankel structure. The results therein, however, suffered from the same suboptimality issue.

Moving beyond convex relaxation methods, another line of work proposed nonconvex approaches for robust PCA (Cai, Cai and Wei (2019), Cherapanamjeri, Gupta and Jain (2017), Drusvyatskiy, Ioffe and Lewis (2021), Gu, Wang and Liu (2016), Li et al. (2019), Netrapalli et al. (2014), Yi et al. (2016), Zhang, Wang and Gu (2018)), largely motivated by the recent success of nonconvex methods in low-rank matrix factorization (Cai and Wei (2018), Candès, Li and Soltanolkotabi (2015), Charisopoulos et al. (2019), Chen and Candès (2017), Chen and Candès (2018), Chen and Wainwright (2015), Chen et al. (2019a), Chi, Lu and Chen (2019), Jain, Netrapalli and Sanghavi (2013), Keshavan, Montanari and Oh (2010), Ma et al. (2020), Netrapalli, Jain and Sanghavi (2015), Sun and Luo (2016), Wang, Giannakis and Eldar (2018), Wei et al. (2016), Zhang, Chi and Liang (2016), Zheng and Lafferty (2016)). Following the deterministic model of Chandrasekaran et al. (2011), the paper Netrapalli et al. (2014) proposed an alternating projection/minimization scheme to seek a low-rank and sparse decomposition of the observed data matrix. In the noiseless setting, that is, $\boldsymbol{E} = \boldsymbol{0}$, this alternating minimization scheme provably disentangles the low-rank and sparse matrix from their superposition under mild conditions. In addition, Netrapalli et al. (2014) extended their result to the arbitrary noise case where the size of the noise is extremely small, namely, $\|\boldsymbol{E}\|_\infty \ll \sigma_{\min}/n$. When the noise $\{E_{ij}\} \sim \mathcal{N}(0, \sigma^2)$, this is equivalent to the condition $\sigma \ll \sigma_{\min}/(n\sqrt{\log n})$. Comparing this with our noise condition $\sigma \ll \sigma_{\min}/(\sqrt{n \log n})$

(cf. (1.12)) when $r, \mu, \kappa \asymp 1$, one sees that our theoretical guarantees cover a wider range of noise levels. Similarly, Yi et al. (2016) applied regularized gradient descent on a smooth nonconvex loss function which enjoys provable convergence guarantees to $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ under the noiseless and partial observation setting. A recent paper Drusvyatskiy, Ioffe and Lewis (2021) considered the nonsmooth nonconvex formulation for robust PCA and established rigorously the convergence of subgradient-type methods in the rank-1 setting, that is, $r = 1$. However, the extension to more general rank remains out of reach.

It is worth noting that noisy matrix completion problem (Candès and Plan (2010), Chen et al. (2020a)) is subsumed as a special case by the model studied in this paper (namely, it is a special case with $\boldsymbol{S}^\star = \boldsymbol{0}$). Statistical optimality under the random noise setting (cf. Assumption 1.5)—including the convex relaxation approach (Chen et al. (2020a), Klopp (2014), Koltchinskii, Lounici and Tsybakov (2011), Negahban and Wainwright (2012)) and the nonconvex approach (Chen, Liu and Li (2020), Ma et al. (2020))—has been extensively studied. Focusing on arbitrary deterministic noise, Candès and Plan (2010) established the stability of the convex approach, whose resulting estimation error bound is similar to the one established for robust PCA with noise in Zhou et al. (2010)) (see (1.4)). The paper Krahmer and Stöger (2021) later confirmed that the estimation error bound established in Candès and Plan (2010) is the best one can hope for in the arbitrary noise setting for matrix completion, although it might be highly suboptimal if we restrict attention to random noise.

Finally, there is also a large literature considering robust PCA under different settings and/or from different perspectives. For instance, the computational efficiency in solving the convex optimization problem (1.3) and its variants has been studied in the optimization literature (e.g., Goldfarb, Ma and Scheinberg (2013), Ma and Aybat (2018), Shen, Wen and Zhang (2014), Tao and Yuan (2011)). The problem has also been investigated under a streaming/online setting (Feng, Xu and Yan (2013), Guo, Qiu and Vaswani (2014), Qiu and Vaswani (2010), Qiu et al. (2014), Vaswani and Narayanamurthy (2018), Zhan et al. (2016)). These are beyond the scope of the current paper.

3. Architecture of the proof. In this section, we give an outline for proving Theorem 1.7. The proof of Theorem 1.6 follows immediately as it is a special case of Theorem 1.7. For simplicity of presentation, our proof sets $n_1 = n_2 = n$. It is straightforward to obtain the proof for the general rectangular case via minor modification.

The main ingredient of the proof lies in establishing an intimate link between convex and nonconvex optimization. Unless otherwise noted, we shall set the regularization parameters as

$$\lambda = C_\lambda \sigma \sqrt{np} \quad \text{and} \quad \tau = C_\tau \sigma \sqrt{\log n} \tag{3.1}$$

throughout. In addition, the soft thresholding operator at level τ is defined such that

$$\mathcal{S}_\tau(x) := \mathsf{sign}(x) \max\big(|x| - \tau, 0\big). \tag{3.2}$$

For any matrix $\boldsymbol{X}$, the matrix $\mathcal{S}_\tau(\boldsymbol{X})$ is obtained by applying the soft thresholding operator $\mathcal{S}_\tau(\cdot)$ to each entry of $\boldsymbol{X}$ separately. Additionally, we define the true low-rank factors as follows:

$$\boldsymbol{X}^\star := \boldsymbol{U}^\star \big(\boldsymbol{\Sigma}^\star\big)^{1/2} \quad \text{and} \quad \boldsymbol{Y}^\star := \boldsymbol{V}^\star \big(\boldsymbol{\Sigma}^\star\big)^{1/2}, \tag{3.3}$$

where $\boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$ is the SVD of the true low-rank matrix $\boldsymbol{L}^\star$.

3.1. *Crude estimation error bounds for convex relaxation.* We start by delivering a crude upper bound on the Euclidean estimation error, built upon the (approximate) duality certificate previously constructed in Chen et al. (2013). The proof is postponed to Appendix 4.

THEOREM 3.1. *Consider any given $\lambda > 0$ and set $\tau \asymp \lambda\sqrt{(\log n)/np}$. Suppose that Assumptions 1.1–1.4 hold, and that*

$$n^2 p \geq C\mu^2 r^2 n \log^6 n \quad \text{and} \quad \rho_{\mathsf{s}} \leq c$$

hold for some sufficiently large (resp. small) constant $C > 0$ (resp., $c > 0$). Then with probability at least $1 - O(n^{-10})$, any minimizer $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$ of the convex program (1.3) satisfies

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}}^2 + \|\boldsymbol{S}_{\mathsf{cvx}} - \boldsymbol{S}^\star\|_{\mathrm{F}}^2 \lesssim \lambda^2 n^5 \log^3 n + \frac{n}{\lambda^2}\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}}^4. \tag{3.4}$$

It is worth noting that the above theorem holds true for an arbitrary noise matrix $\boldsymbol{E}$. When specialized to the case with independent sub-Gaussian noise, this crude bound admits a simpler expression as follows.

COROLLARY 3.2. *Take $\lambda = C_\lambda \sigma\sqrt{np}$ and $\tau = C_\tau \sigma\sqrt{\log n}$ for some universal constant $C_\lambda, C_\tau > 0$. Under the assumptions of Theorem 3.1 and Assumption 1.5, we have—with probability exceeding $1 - O(n^{-10})$—that*

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}} \lesssim \sigma n^3 \log^{3/2} n \quad \text{and} \quad \|\boldsymbol{S}_{\mathsf{cvx}} - \boldsymbol{S}^\star\|_{\mathrm{F}} \lesssim \sigma n^3 \log^{3/2} n. \tag{3.5}$$

PROOF. This corollary follows immediately by combining Theorem 3.1 and Lemma 3.3 below. □

LEMMA 3.3. *Suppose that Assumption 1.5 holds and that $n^2 p > C_1 n \log^2 n$ for some sufficiently large constant $C_1 > 0$. Then with probability exceeding $1 - O(n^{-10})$, one has*

$$\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\| \lesssim \sigma\sqrt{np} \quad \text{and} \quad \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}} \lesssim \sigma n\sqrt{p}.$$

While the above results often lose a polynomial factor in n, namely, the optimal error bound, it serves as an important starting point that paves the way for subsequent analytical refinement.

3.2. *Approximate stationary points of the nonconvex formulation.* Instead of analyzing the convex estimator directly, we take a detour by considering the following nonconvex optimization problem

$$\begin{aligned} &\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r},\boldsymbol{S}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad F(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S}) \\ &:= \underbrace{\frac{1}{2p}\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{S} - \boldsymbol{M})\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{Y}\|_{\mathrm{F}}^2}_{=:f(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{S})} + \frac{\tau}{p}\|\boldsymbol{S}\|_1. \end{aligned} \tag{3.6}$$

Here, $f(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{S})$ is a function of $\boldsymbol{X}$ and $\boldsymbol{Y}$ with $\boldsymbol{S}$ frozen, which contains the smooth component of the loss function $F(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S})$. As it turns out, the solution to convex relaxation (1.3) is exceedingly close to an estimate $(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S})$ obtained by a nonconvex algorithm aimed at solving (3.6)—to be detailed in Section 3.3. This fundamental connection between the two algorithmic paradigms provides a powerful framework that allows us to understand convex relaxation by studying nonconvex optimization.

In what follows, we set out to develop the afforementioned intimate connection. Before proceeding, we first state the following conditions concerned with the interplay between the noise size, the estimation accuracy of the nonconvex estimate $(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S})$, and the regularization parameters.

CONDITION 3.4. *The regularization parameters* λ *and* $\tau \asymp \lambda\sqrt{(\log n)/np}$ *satisfy*

- $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\| < \lambda/16$ *and* $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_\infty \le \tau/4$;
- $\|\boldsymbol{S} - \boldsymbol{S}^\star\| < \lambda/16$ *and* $\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{L}^\star\|_\infty \le \tau/4$;
- $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{L}^\star) - p(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{L}^\star)\| < \lambda/8$.

As an interpretation, the above condition says that: (1) the regularization parameters are not too small compared to the size of the noise, so as to ensure that we enforce a sufficiently large degree of regularization; (2) the estimate represented by the point $(\boldsymbol{X}\boldsymbol{Y}^\top, \boldsymbol{S})$ is sufficiently close to the truth. At this point, whether this condition is meaningful or not remains far from clear; we shall return to justify its feasibility shortly.

In addition, we need another condition concerning the injectivity of $\mathcal{P}_{\Omega^\star}$ w.r.t. a certain tangent space. For a rank-r matrix $\boldsymbol{L}$ with singular value decomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n\times r}$, the tangent space of the set of rank-r matrices at the point $\boldsymbol{L}$ is given by

$$\{\boldsymbol{U}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{V}^\top | \boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times r}\}.$$

Again, the validity of this condition will be discussed momentarily.

CONDITION 3.5 (Injectivity). *Let* T *be the tangent space of the set of rank-*r *matrices at the point* $\boldsymbol{X}\boldsymbol{Y}^\top$. *Assume that there exist a constants* $c_{\mathrm{inj}} > 0$ *such that for all* $\boldsymbol{H} \in T$, *one has*

$$p^{-1}\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{H})\|_{\mathrm{F}}^2 \ge \frac{c_{\mathrm{inj}}}{\kappa}\|\boldsymbol{H}\|_{\mathrm{F}}^2 \quad \text{and} \quad p^{-1}\|\mathcal{P}_{\Omega^\star}(\boldsymbol{H})\|_{\mathrm{F}}^2 \le \frac{c_{\mathrm{inj}}}{4\kappa}\|\boldsymbol{H}\|_{\mathrm{F}}^2.$$

With the above conditions in place, we are ready to make precise the intimate link between convex relaxation and a candidate nonconvex solution. The proof is deferred to Appendix 5.

THEOREM 3.6. *Suppose that* $n \ge \kappa$ *and* $\rho_{\mathsf{s}} \le c/\kappa$ *for some sufficiently small constant* $c > 0$. *Assume that there exists a triple* $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S})$ *such that*

$$\|\nabla f(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{S})\|_{\mathrm{F}} \le \frac{1}{n^{20}}\frac{\lambda}{p}\sqrt{\sigma_{\min}}, \quad \text{and} \quad \boldsymbol{S} = \mathcal{P}_{\Omega_{\mathsf{obs}}}(\mathcal{S}_\tau(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top)). \tag{3.7}$$

Further, assume that any singular value of $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *lies in* $[\sqrt{\sigma_{\min}/2}, \sqrt{2\sigma_{\max}}]$. *If the solution* $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$ *to the convex program* (1.3) *admits the following crude error bound*

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}} \lesssim \sigma n^4, \tag{3.8}$$

then under Conditions 3.4–3.5 *we have*

$$\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{L}_{\mathsf{cvx}}\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^5} \quad \text{and} \quad \|\boldsymbol{S} - \boldsymbol{S}_{\mathsf{cvx}}\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^5}.$$

This theorem is a deterministic result, focusing on some sort of "approximate stationary points" of $F(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S})$. To interpret this, observe that in view of (3.7), one has $\nabla f(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{S}) \approx \boldsymbol{0}$, and $\boldsymbol{S}$ minimizes $F(\boldsymbol{X}, \boldsymbol{Y}, \cdot)$ for any fixed $\boldsymbol{X}$ and $\boldsymbol{Y}$. If one can identify such an approximate stationary point that is sufficiently close to the truth (so that it satisfies Condition 3.4), then under mild conditions our theory asserts that

$$\boldsymbol{X}\boldsymbol{Y}^\top \approx \boldsymbol{L}_{\mathsf{cvx}} \quad \text{and} \quad \boldsymbol{S} \approx \boldsymbol{S}_{\mathsf{cvx}}.$$

This would in turn formalize the intimate relation between the solution to convex relaxation and an approximate stationary point of the nonconvex formulation. The existence of such approximate stationary points will be verified shortly in Section 3.3.

The careful reader might immediately remark that this theorem does not say anything explicit about the minimizer of the nonconvex optimization problem (3.6); rather, it only pays

Algorithm 1 Alternating minimization method for solving the nonconvex problem (3.6)

Suitable initialization: $\boldsymbol{X}^0 = \boldsymbol{X}^\star$, $\boldsymbol{Y}^0 = \boldsymbol{Y}^\star$, $\boldsymbol{S}^0 = \boldsymbol{S}^\star$.

Gradient updates: **for** $t = 0, 1, \ldots, t_0 - 1$ **do**

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t; \boldsymbol{S}^t) = \boldsymbol{X}^t - \frac{\eta}{p}\big[\mathcal{P}_{\Omega_{\text{obs}}}(\boldsymbol{X}^t \boldsymbol{Y}^{t\top} + \boldsymbol{S}^t - \boldsymbol{M})\boldsymbol{Y}^t + \lambda \boldsymbol{X}^t\big]; \tag{3.9a}$$

$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t; \boldsymbol{S}^t) = \boldsymbol{Y}^t - \frac{\eta}{p}\big\{\big[\mathcal{P}_{\Omega_{\text{obs}}}(\boldsymbol{X}^t \boldsymbol{Y}^{t\top} + \boldsymbol{S}^t - \boldsymbol{M})\big]^\top \boldsymbol{X}^t + \lambda \boldsymbol{Y}^t\big\}; \tag{3.9b}$$

$$\boldsymbol{S}^{t+1} = \mathcal{S}_\tau\big[\mathcal{P}_{\Omega_{\text{obs}}}(\boldsymbol{M} - \boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top})\big]. \tag{3.9c}$$

attention to a special class of approximate stationary points of the nonconvex formulation. This arises mainly due to a technical consideration: it seems more difficult to analyze the nonconvex optimizer directly than to study certain approximate stationary points. Fortunately, our theorem indicates that any approximate stationary point obeying the above conditions serves as an extremely tight approximation of the convex estimate and, therefore, it suffices to identify and analyze any such points.

3.3. *Constructing an approximate stationary point via nonconvex algorithms.* By virtue of Theorem 3.6, the key to understanding convex relaxation is to construct an approximate stationary point of the nonconvex problem (3.6) that enjoys desired statistical properties. For this purpose, we resort to the iterative algorithm (Algorithm 1) to solve the nonconvex program (3.6).

In a nutshell, Algorithm 1 alternates between one iteration of gradient updates (w.r.t. the decision matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$) and optimization of the nonsmooth problem w.r.t. $\boldsymbol{S}$ (with $\boldsymbol{X}$ and $\boldsymbol{Y}$ frozen).[8] For the sake of simplicity, we initialize this algorithm from the ground truth $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star, \boldsymbol{S}^\star)$, but our analysis framework might be extended to accommodate other more practical initialization (e.g., the one obtained by a spectral method Chen et al. (2020b)).

The following theorem makes precise the statistical guarantees of the above nonconvex optimization algorithm; the proof is deferred to Appendix 6. Here and throughout, we define

$$\boldsymbol{H}^t := \mathop{\arg\min}_{\boldsymbol{R} \in \mathcal{O}^{r\times r}} \big(\|\boldsymbol{X}^t \boldsymbol{R} - \boldsymbol{X}^\star\|_{\text{F}}^2 + \|\boldsymbol{Y}^t \boldsymbol{R} - \boldsymbol{Y}^\star\|_{\text{F}}^2\big)^{1/2}, \tag{3.10}$$

where $\mathcal{O}^{r\times r}$ denotes the set of $r \times r$ orthonormal matrices.

THEOREM 3.7. *Instate the assumptions of Theorem 1.7 and define*

$$\delta_n := \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}.$$

Take $t_0 = n^{47}$ *and* $\eta \asymp 1/(n\kappa^3\sigma_{\max})$ *in Algorithm 1. With probability at least* $1 - O(n^{-3})$, *the iterates* $\{(\boldsymbol{X}^t, \boldsymbol{Y}^t, \boldsymbol{S}^t)\}_{0\le t\le t_0}$ *of Algorithm 1 satisfy*

$$\max\{\|\boldsymbol{X}^t\boldsymbol{H}^t - \boldsymbol{X}^\star\|_{\text{F}}, \|\boldsymbol{Y}^t\boldsymbol{H}^t - \boldsymbol{Y}^\star\|_{\text{F}}\} \lesssim \delta_n \|\boldsymbol{X}^\star\|_{\text{F}}, \tag{3.11a}$$

$$\max\{\|\boldsymbol{X}^t\boldsymbol{H}^t - \boldsymbol{X}^\star\|, \|\boldsymbol{Y}^t\boldsymbol{H}^t - \boldsymbol{Y}^\star\|\} \lesssim \delta_n \|\boldsymbol{X}^\star\|, \tag{3.11b}$$

[8] Note that for any given $\boldsymbol{X}$ and $\boldsymbol{Y}$, the solution to minimize$_{\boldsymbol{S}} F(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S})$ is given precisely by $\mathcal{S}_\tau(\mathcal{P}_{\Omega_{\text{obs}}}(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top))$.

$$\begin{aligned}&\max\{\|\boldsymbol{X}^t\boldsymbol{H}^t-\boldsymbol{X}^\star\|_{2,\infty},\|\boldsymbol{Y}^t\boldsymbol{H}^t-\boldsymbol{Y}^\star\|_{2,\infty}\}\\&\lesssim\kappa\sqrt{\log n}\delta_n\max\{\|\boldsymbol{X}^\star\|_{2,\infty},\|\boldsymbol{Y}^\star\|_{2,\infty}\},\end{aligned}\tag{3.11c}$$

$$\|\boldsymbol{S}^t-\boldsymbol{S}^\star\|\lesssim\sigma\sqrt{np}.\tag{3.11d}$$

In addition, with probability at least $1-O(n^{-3})$, *one has*

$$\min_{0\le t<t_0}\|\nabla f(\boldsymbol{X}^t,\boldsymbol{Y}^t;\boldsymbol{S}^t)\|_{\mathrm{F}}\le\frac{1}{n^{20}}\frac{\lambda}{p}\sqrt{\sigma_{\min}}.\tag{3.12}$$

In short, the bounds (3.11a)–(3.11c) reveal that the entire sequence $\{\boldsymbol{X}^t,\boldsymbol{Y}^t\}_{t=0}^{t_0}$ stays sufficiently close to the truth (measured by $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|$, and more importantly, $\|\cdot\|_{2,\infty}$), the inequality (3.11d) demonstrates the goodness of fit of $\{\boldsymbol{S}^t\}_{0\le t\le t_0}$ in terms of the spectral norm accuracy, whereas the last bound (3.12) indicates that there is at least one point in the sequence $\{\boldsymbol{X}^t,\boldsymbol{Y}^t,\boldsymbol{S}^t\}_{0\le t\le t_0}$ that can serve as an approximate stationary point of the nonconvex formulation.

We shall also gather a few immediate consequences of Theorem 3.7 as follows, which contain basic properties that will be useful throughout.

COROLLARY 3.8. *Instate the assumptions of Theorem 3.7. Suppose that the sample size obeys* $n^2p\gg\kappa^4\mu^2r^2n\log^4 n$, *the noise satisfies* $\delta_n\ll1/\sqrt{\kappa^4\mu r\log n}$, *the outlier fraction satisfies* $\rho_{\mathsf{s}}\ll1/(\kappa^3\mu r\log n)$. *With probability at least* $1-O(n^{-3})$, *the iterates of Algorithm 1 satisfy*

$$\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top}-\boldsymbol{L}^\star\|_{\mathrm{F}}\lesssim\kappa\delta_n\|\boldsymbol{L}^\star\|_{\mathrm{F}},\tag{3.13a}$$

$$\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top}-\boldsymbol{L}^\star\|_{\infty}\lesssim\sqrt{\kappa^3\mu r\log n}\delta_n\|\boldsymbol{L}^\star\|_{\infty},\tag{3.13b}$$

$$\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top}-\boldsymbol{L}^\star\|\lesssim\delta_n\|\boldsymbol{L}^\star\|\tag{3.13c}$$

simultaneously for all $t\le t_0$.

PROOF. See Chen et al. (2020a), Appendix D.12. □

3.4. *Proof of Theorem 1.7.* Define

$$t_*:=\arg\min_{0\le t<t_0}\|\nabla f(\boldsymbol{X}^t,\boldsymbol{Y}^t;\boldsymbol{S}^t)\|_{\mathrm{F}};\tag{3.14}$$

$$(\boldsymbol{X}_{\mathsf{ncvx}},\boldsymbol{Y}_{\mathsf{ncvx}},\boldsymbol{S}_{\mathsf{ncvx}}):=(\boldsymbol{X}^{t_*}\boldsymbol{H}^{t_*},\boldsymbol{Y}^{t_*}\boldsymbol{H}^{t_*},\boldsymbol{S}^{t_*}).\tag{3.15}$$

Theorem 3.7 and Corollary 3.8 have established appealing statistical performance of the nonconvex solution $(\boldsymbol{X}_{\mathsf{ncvx}},\boldsymbol{Y}_{\mathsf{ncvx}},\boldsymbol{S}_{\mathsf{ncvx}})$. To transfer this desired statistical property to that of $(\boldsymbol{L}_{\mathsf{cvx}},\boldsymbol{S}_{\mathsf{cvx}})$, it remains to show that the nonconvex estimator $(\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top,\boldsymbol{S}_{\mathsf{ncvx}})$ is extremely close to the convex estimator $(\boldsymbol{L}_{\mathsf{cvx}},\boldsymbol{S}_{\mathsf{cvx}})$. Toward this end, we intend to invoke Theorem 3.6; therefore, it boils down to verifying the conditions therein.

1. The small gradient condition (cf. (3.7)) holds automatically under (3.12).
2. By virtue of the spectral norm bound (3.11b), one has

$$\|\boldsymbol{X}_{\mathsf{ncvx}}-\boldsymbol{X}^\star\|=\|\boldsymbol{X}^{t_*}\boldsymbol{H}^{t_*}-\boldsymbol{X}^\star\|\lesssim\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\|\boldsymbol{L}^\star\|\le\frac{\sqrt{\sigma_{\min}}}{10},$$

as long as $\sigma\sqrt{\kappa n/p}\ll\sigma_{\min}$. This together with the Weyl inequality verifies the constraints on the singular values of $(\boldsymbol{X}_{\mathsf{ncvx}},\boldsymbol{Y}_{\mathsf{ncvx}})$.

3. The crude error bounds are valid in view of Theorem 3.1.

4. Regarding Condition 3.4 and Condition 3.5, Lemma 3.3 and standard inequalities about sub-Gaussian random variables imply that $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\| < \lambda/16$ and $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_\infty \le \tau/4$. In addition, the bounds (3.11d) and (3.13b) ensure the second assumption $\|\boldsymbol{S}_{\mathsf{ncvx}} - \boldsymbol{S}^\star\| \le \lambda/16$ and $\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{L}^\star\|_\infty \le \tau/4$ in Condition 3.4. We are left with the last assumption in Condition 3.4 and Condition 3.5, which are guaranteed to hold in view of the following lemma (see Appendix 3 for the proof).

LEMMA 3.9. *Instate the notation and assumptions of Theorem 1.7. Then with probability exceeding* $1 - O(n^{-10})$, *we have*

$$\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star) - p(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star)\| < \lambda/8, \tag{3.16a}$$

$$\frac{1}{p}\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{H})\|_{\mathrm{F}}^2 \ge \frac{1}{32\kappa}\|\boldsymbol{H}\|_{\mathrm{F}}^2, \quad \forall \boldsymbol{H} \in T, \tag{3.16b}$$

$$p^{-1}\|\mathcal{P}_{\Omega^\star}(\boldsymbol{H})\|_{\mathrm{F}}^2 \le \frac{1}{128\kappa}\|\boldsymbol{H}\|_{\mathrm{F}}^2, \quad \forall \boldsymbol{H} \in T \tag{3.16c}$$

simultaneously for all $(\boldsymbol{X}, \boldsymbol{Y})$ *obeying*

$$\|\boldsymbol{X} - \boldsymbol{X}^\star\|_{2,\infty} \le C_\infty \kappa \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\max\{\|\boldsymbol{X}^\star\|_{2,\infty}, \|\boldsymbol{Y}^\star\|_{2,\infty}\}; \tag{3.17a}$$

$$\|\boldsymbol{Y} - \boldsymbol{Y}^\star\|_{2,\infty} \le C_\infty \kappa \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\max\{\|\boldsymbol{X}^\star\|_{2,\infty}, \|\boldsymbol{Y}^\star\|_{2,\infty}\}. \tag{3.17b}$$

Here, T *denotes the tangent space of the set of rank-*r *matrices at the point* $\boldsymbol{X}\boldsymbol{Y}^\top$, *and* $C_\infty > 0$ *is an absolute constant.*

Armed with the above conditions, we can readily invoke Theorem 3.6 to reach

$$\|\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top - \boldsymbol{L}_{\mathsf{cvx}}\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^5} \quad \text{and} \quad \|\boldsymbol{S}_{\mathsf{ncvx}} - \boldsymbol{S}_{\mathsf{cvx}}\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^5}$$

with high probability. This taken collectively with Corollary 3.8 gives

$$\begin{aligned}\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}} &\le \|\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top - \boldsymbol{L}_{\mathsf{cvx}}\|_{\mathrm{F}} + \|\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top - \boldsymbol{L}^\star\|_{\mathrm{F}} \\ &\lesssim \frac{\sigma}{n^5} + \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\|\boldsymbol{L}^\star\|_{\mathrm{F}} \\ &\asymp \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\|\boldsymbol{L}^\star\|_{\mathrm{F}}.\end{aligned}$$

Similar arguments lead to the advertised high-probability bounds

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_\infty \lesssim \sqrt{\kappa^3\mu r}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\|\boldsymbol{L}^\star\|_\infty,$$

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\| \lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\|\boldsymbol{L}^\star\|.$$

Finally, given that $\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top$ is a rank-r matrix, the rank-r approximation $\boldsymbol{L}_{\mathsf{cvx},r} := \arg\min_{\boldsymbol{Z}:\mathrm{rank}(\boldsymbol{Z})\le r}\|\boldsymbol{Z} - \boldsymbol{L}_{\mathsf{cvx}}\|_{\mathrm{F}}$ of $\boldsymbol{L}_{\mathsf{cvx}}$ necessarily satisfies

$$\|\boldsymbol{L}_{\mathsf{cvx},r} - \boldsymbol{L}_{\mathsf{cvx}}\|_{\mathrm{F}} \le \|\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top - \boldsymbol{L}_{\mathsf{cvx}}\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^5} \le \frac{1}{n^5}\cdot\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\|\boldsymbol{L}^\star\|,$$

which establishes (1.22). In view of the triangle inequality, the properties (1.21) hold unchanged if $\boldsymbol{L}_{\mathsf{cvx}}$ is replaced by $\boldsymbol{L}_{\mathsf{cvx},r}$.

4. Discussion. This paper investigates the unreasonable effectiveness of convex programming in estimating an unknown low-rank matrix from grossly corrupted data. We develop an improved theory that confirms the optimality of convex relaxation in the presence of random noise, gross sparse outliers and missing data. In particular, our results significantly improve upon the prior statistical guarantees (Zhou et al. (2010)) under random noise, while further allowing for missing data. Our theoretical analysis is built upon an appealing connection between convex and nonconvex optimization, which has not been established previously.

Having said this, our current work leaves open several important issues that call for further investigation. To begin with, the conditions (1.20) stated in the main theorem are likely suboptimal in terms of the dependency on both the rank r and the condition number κ. For example, we shall keep in mind that in the noise-free setting, the sample size can be as low as $O(nr\text{poly}\log n)$ and the tolerable outlier fraction can be as large as a constant (Chen et al. (2013), Li (2013)), both of which exhibit more favorable scalings w.r.t. r and κ compared to our current condition (1.20). Moving forward, our analysis ideas suggest a possible route for analyzing convex relaxation for other structured estimation problems under both random noise and outliers, including but not limited to sparse PCA (the case with a simultaneously low-rank and sparse matrix) (Cai, Ma and Wu (2013)), low-rank Hankel matrix estimation (the case involving a low-rank Hankel matrix) (Chen and Chi (2014)), and blind deconvolution[9] (the case that aims to recover a low-rank matrix from structured Fourier measurements) (Ahmed, Recht and Romberg (2014)). Last but not least, we would like to point out that it is possible to design a similar debiasing procedure as in Chen et al. (2019b) for correcting the bias in the convex estimator, which further allows uncertainty quantification and statistical inference on the unknown low-rank matrix of interest.

Acknowledgments. Author names are sorted alphabetically. Y. Chen is the corresponding author.

Funding. Y. Chen is supported in part by the AFOSR YIP award FA9550-19-1-0030, by the ONR Grant N00014-19-1-2120, by the ARO Grants W911NF-20-1-0097 and W911NF-18-1-0303, by NSF Grants CCF-1907661, IIS-1900140, IIS-2100158, and DMS-2014279 and by the Princeton SEAS innovation award. J. Fan is supported in part by the NSF Grants DMS-1662139 and DMS-1712591, the ONR Grant N00014-19-1-2120 and the NIH Grant 2R01-GM072611-14.

SUPPLEMENTARY MATERIAL

Additional proofs (DOI: 10.1214/21-AOS2066SUPP; .pdf). Additional proofs of the results in the paper can be found in the Supplementary Material (Chen et al. (2021)).

REFERENCES

AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist.* **40** 1171–1197. MR2985947 https://doi.org/10.1214/12-AOS1000

AHMED, A., RECHT, B. and ROMBERG, J. (2014). Blind deconvolution using convex programming. *IEEE Trans. Inf. Theory* **60** 1711–1732. MR3168432 https://doi.org/10.1109/TIT.2013.2294644

CAI, H., CAI, J.-F. and WEI, K. (2019). Accelerated alternating projections for robust principal component analysis. *J. Mach. Learn. Res.* **20** Paper No. 20, 33. MR3911427

CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. MR3161458 https://doi.org/10.1214/13-AOS1178

[9]Our ongoing work Chen et al. (2020c) is pursuing this direction.

CAI, J.-F. and WEI, K. (2018). Solving systems of phaseless equations via Riemannian optimization with optimal sampling complexity. Preprint. arXiv:1809.02773.

CANDÈS, E. J., LI, X. and SOLTANOLKOTABI, M. (2015). Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Trans. Inf. Theory* **61** 1985–2007. MR3332993 https://doi.org/10.1109/TIT.2015.2399924

CANDÈS, E. and PLAN, Y. (2010). Matrix completion with noise. *Proc. IEEE* **98** 925–936.

CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240 https://doi.org/10.1007/s10208-009-9045-5

CANDÈS, E. J. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theory* **56** 2053–2080. MR2723472 https://doi.org/10.1109/TIT.2010.2044061

CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37. MR2811000 https://doi.org/10.1145/1970392.1970395

CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067 https://doi.org/10.1214/11-AOS949

CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P. A. and WILLSKY, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.* **21** 572–596. MR2817479 https://doi.org/10.1137/090761793

CHARISOPOULOS, V., DAVIS, D., DÍAZ, M. and DRUSVYATSKIY, D. (2019). Composite optimization for robust blind deconvolution. Preprint. arXiv::1901.01624.

CHEN, Y. (2015). Incoherence-optimal matrix completion. *IEEE Trans. Inf. Theory* **61** 2909–2923. MR3342311 https://doi.org/10.1109/TIT.2015.2415195

CHEN, Y. and CANDÈS, E. J. (2017). Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Comm. Pure Appl. Math.* **70** 822–883. MR3628877 https://doi.org/10.1002/cpa.21638

CHEN, Y. and CANDÈS, E. J. (2018). The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Comm. Pure Appl. Math.* **71** 1648–1714. MR3847751 https://doi.org/10.1002/cpa.21760

CHEN, Y. and CHI, Y. (2014). Robust spectral compressed sensing via structured matrix completion. *IEEE Trans. Inf. Theory* **60** 6576–6601. MR3265040 https://doi.org/10.1109/TIT.2014.2343623

CHEN, Y., GUIBAS, L. J. and HUANG, Q. (2014). Near-optimal joint optimal matching via convex relaxation. In *International Conference on Machine Learning* (*ICML*) 100–108.

CHEN, J., LIU, D. and LI, X. (2020). Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *IEEE Trans. Inform. Theory* **66** 5806–5841. MR4158648 https://doi.org/10.1109/TIT.2020.2992234

CHEN, Y. and WAINWRIGHT, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. arXiv:1509.03025.

CHEN, Y., JALALI, A., SANGHAVI, S. and CARAMANIS, C. (2013). Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory* **59** 4324–4337.

CHEN, Y., JALALI, A., SANGHAVI, S. and XU, H. (2014). Clustering partially observed graphs via convex optimization. *J. Mach. Learn. Res.* **15** 2213–2238. MR3231602

CHEN, Y., CHI, Y., FAN, J. and MA, C. (2019a). Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Math. Program.* **176** 5–37. MR3960803 https://doi.org/10.1007/s10107-019-01363-6

CHEN, Y., FAN, J., MA, C. and YAN, Y. (2019b). Inference and uncertainty quantification for noisy matrix completion. *Proc. Natl. Acad. Sci. USA* **116** 22931–22937. MR4036123 https://doi.org/10.1073/pnas.1910053116

CHEN, Y., CHI, Y., FAN, J., MA, C. and YAN, Y. (2020a). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM J. Optim.* **30** 3098–3121. MR4167625 https://doi.org/10.1137/19M1290000

CHEN, Y., CHI, Y., FAN, J. and MA, C. (2020b). Spectral methods for data science: A statistical perspective. Preprint. arXiv::2012.08496.

CHEN, Y., FAN, J., WANG, B. and YAN, Y. (2020c). Convex and nonconvex optimization are both minimax-optimal for noisy blind deconvolution. Preprint. arXiv::2008.01724.

CHEN, Y., FAN, J., MA, C. and YAN, Y. (2021). Supplement to "Bridging convex and nonconvex optimization in robust PCA: Noise, outliers and missing data." https://doi.org/10.1214/21-AOS2066SUPP

CHERAPANAMJERI, Y., GUPTA, K. and JAIN, P. (2017). Nearly optimal robust matrix completion. In *Proceedings of the* 34*th International Conference on Machine Learning* **70** 797–805. JMLR.org.

CHI, Y., LU, Y. M. and CHEN, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Trans. Signal Process.* **67** 5239–5269. MR4016283 https://doi.org/10.1109/TSP.2019.2937282

DAVENPORT, M. A. and ROMBERG, J. (2016). An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing* **10** 608–622.

DONOHO, D. and GAVISH, M. (2014). Minimax risk of matrix denoising by singular value thresholding. *Ann. Statist.* **42** 2413–2440. MR3269984 https://doi.org/10.1214/14-AOS1257

DRUSVYATSKIY, D., IOFFE, A. D. and LEWIS, A. S. (2021). Nonsmooth optimization using Taylor-like models: Error bounds, convergence, and termination criteria. *Math. Program.* **185** 357–383. MR4201717 https://doi.org/10.1007/s10107-019-01432-w

FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197. MR2472991 https://doi.org/10.1016/j.jeconom.2008.09.017

FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. MR3091653 https://doi.org/10.1111/rssb.12016

FAN, J., WANG, W. and ZHONG, Y. (2017). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *J. Mach. Learn. Res.* **18** Paper No. 207, 42. MR3827095

FAN, J., WANG, W. and ZHONG, Y. (2019). Robust covariance estimation for approximate factor models. *J. Econometrics* **208** 5–22. MR3906959 https://doi.org/10.1016/j.jeconom.2018.09.003

FAN, J., SUN, Q., ZHOU, W.-X. and ZHU, Z. (2018). Principal component analysis for big data. Preprint. arXiv::1801.01602.

FENG, J., XU, H. and YAN, S. (2013). Online robust PCA via stochastic optimization. In *Advances in Neural Information Processing Systems* 404–412.

GANESH, A., WRIGHT, J., LI, X., CANDES, E. J. and MA, Y. (2010). Dense error correction for low-rank matrices via principal component pursuit. In 2010 *IEEE International Symposium on Information Theory* 1513–1517. IEEE.

GOLDFARB, D., MA, S. and SCHEINBERG, K. (2013). Fast alternating linearization methods for minimizing the sum of two convex functions. *Math. Program.* **141** 349–382. MR3097290 https://doi.org/10.1007/s10107-012-0530-2

GROSS, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory* **57** 1548–1566. MR2815834 https://doi.org/10.1109/TIT.2011.2104999

GU, Q., WANG, Z. W. and LIU, H. (2016). Low-rank and sparse structure pursuit via alternating minimization. In *Artificial Intelligence and Statistics* 600–609.

GUO, H., QIU, C. and VASWANI, N. (2014). An online algorithm for separating sparse and low-dimensional signal sequences from their sum. *IEEE Trans. Signal Process.* **62** 4284–4297. MR3260427 https://doi.org/10.1109/TSP.2014.2331612

HSU, D., KAKADE, S. M. and ZHANG, T. (2011). Robust matrix decomposition with sparse corruptions. *IEEE Trans. Inform. Theory* **57** 7221–7234. MR2883652 https://doi.org/10.1109/TIT.2011.2158250

HUANG, Q. and GUIBAS, L. (2013). Consistent shape maps via semidefinite programming. *Computer Graphics Forum* **32** 177–186.

JAIN, P., NETRAPALLI, P. and SANGHAVI, S. (2013). Low-rank matrix completion using alternating minimization (extended abstract). In *STOC'*13*—Proceedings of the* 2013 *ACM Symposium on Theory of Computing* 665–674. ACM, New York. MR3210828 https://doi.org/10.1145/2488608.2488693

JOLLIFFE, I. T. (1986). *Principal Component Analysis. Springer Series in Statistics*. Springer, New York. MR0841268 https://doi.org/10.1007/978-1-4757-1904-8

KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from a few entries. *IEEE Trans. Inform. Theory* **56** 2980–2998. MR2683452 https://doi.org/10.1109/TIT.2010.2046205

KLOPP, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **20** 282–303. MR3160583 https://doi.org/10.3150/12-BEJ486

KLOPP, O., LOUNICI, K. and TSYBAKOV, A. B. (2017). Robust matrix completion. *Probab. Theory Related Fields* **169** 523–564. MR3704775 https://doi.org/10.1007/s00440-016-0736-y

KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869 https://doi.org/10.1214/11-AOS894

KRAHMER, F. and STÖGER, D. (2021). On the convex geometry of blind deconvolution and matrix completion. *Comm. Pure Appl. Math.* **74** 790–832. MR4221934

LI, X. (2013). Compressed sensing and matrix completion with constant proportion of corruptions. *Constr. Approx.* **37** 73–99. MR3010211 https://doi.org/10.1007/s00365-012-9176-9

LI, Y., MA, C., CHEN, Y. and CHI, Y. (2019). Nonconvex matrix factorization from rank-one measurements. In *The* 22*nd International Conference on Artificial Intelligence and Statistics* 1496–1505.

MA, S. and AYBAT, N. S. (2018). Efficient optimization algorithms for robust principal component analysis and its variants. *Proceedings of the IEEE* **106** 1411–1426.

MA, C., WANG, K., CHI, Y. and CHEN, Y. (2020). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Found. Comput. Math.* **20** 451–632. MR4099988 https://doi.org/10.1007/s10208-019-09429-9

MAZUMDER, R., HASTIE, T. and TIBSHIRANI, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11** 2287–2322. MR2719857

NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13** 1665–1697. MR2930649

NETRAPALLI, P., JAIN, P. and SANGHAVI, S. (2015). Phase retrieval using alternating minimization. *IEEE Trans. Signal Process.* **63** 4814–4826. MR3385838 https://doi.org/10.1109/TSP.2015.2448516

NETRAPALLI, P., NIRANJAN, U., SANGHAVI, S., ANANDKUMAR, A. and JAIN, P. (2014). Non-convex robust PCA. In *Advances in Neural Information Processing Systems* 1107–1115.

PEARSON, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2** 559–572.

QIU, C. and VASWANI, N. (2010). Real-time robust principal components' pursuit. In 2010 48*th Annual Allerton Conference on Communication, Control, and Computing* (*Allerton*) 591–598. IEEE.

QIU, C., VASWANI, N., LOIS, B. and HOGBEN, L. (2014). Recursive robust PCA or recursive sparse recovery in large but structured noise. *IEEE Trans. Inform. Theory* **60** 5007–5039. MR3245369 https://doi.org/10.1109/TIT.2014.2331344

SHEN, Y., WEN, Z. and ZHANG, Y. (2014). Augmented Lagrangian alternating direction method for matrix separation based on low-rank factorization. *Optim. Methods Softw.* **29** 239–263. MR3175484 https://doi.org/10.1080/10556788.2012.700713

SINGER, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.* **30** 20–36. MR2737931 https://doi.org/10.1016/j.acha.2010.02.001

SREBRO, N. and SHRAIBMAN, A. (2005). Rank, trace-norm and max-norm. In *Learning Theory. Lecture Notes in Computer Science* **3559** 545–560. Springer, Berlin. MR2203286 https://doi.org/10.1007/11503415_37

SUN, R. and LUO, Z.-Q. (2016). Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inform. Theory* **62** 6535–6579. MR3565131 https://doi.org/10.1109/TIT.2016.2598574

TAO, M. and YUAN, X. (2011). Recovering low-rank and sparse components of matrices from incomplete and noisy observations. *SIAM J. Optim.* **21** 57–81. MR2765489 https://doi.org/10.1137/100781894

VASWANI, N. and NARAYANAMURTHY, P. (2018). Static and dynamic robust PCA and matrix completion: A review. *Proceedings of the IEEE* **106** 1359–1379.

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

WANG, G., GIANNAKIS, G. B. and ELDAR, Y. C. (2018). Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Trans. Inform. Theory* **64** 773–794. MR3762591 https://doi.org/10.1109/TIT.2017.2756858

WEI, K., CAI, J.-F., CHAN, T. F. and LEUNG, S. (2016). Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM J. Matrix Anal. Appl.* **37** 1198–1222. MR3543156 https://doi.org/10.1137/15M1050525

WONG, R. K. W. and LEE, T. C. M. (2017). Matrix completion with noisy entries and outliers. *J. Mach. Learn. Res.* **18** Paper No. 147, 25. MR3763781

YI, X., PARK, D., CHEN, Y. and CARAMANIS, C. (2016). Fast algorithms for robust PCA via gradient descent. In *NIPS* 4152–4160.

ZHAN, J., LOIS, B., GUO, H. and VASWANI, N. (2016). Online (and offline) robust PCA: Novel algorithms and performance guarantees. In *Artificial Intelligence and Statistics* 1488–1496.

ZHANG, H., CHI, Y. and LIANG, Y. (2016). Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow. In *International Conference on Machine Learning* 1022–1031.

ZHANG, X., WANG, L. W. and GU, Q. (2018). A unified framework for nonconvex low-rank plus sparse matrix recovery. In *International Conference on Artificial Intelligence and Statistics*.

ZHENG, Q. and LAFFERTY, J. (2016). Convergence analysis for rectangular matrix completion using Burer–Monteiro factorization and gradient descent. arXiv:1605.07051.

ZHOU, Z., LI, X., WRIGHT, J., CANDÈS, E. and MA, Y. (2010). Stable principal component pursuit. In *International Symposium on Information Theory* 1518–1522.