SIAM J. OPTIM.
Vol. 30, No. 4, pp. 3098–3121

NOISY MATRIX COMPLETION: UNDERSTANDING STATISTICAL GUARANTEES FOR CONVEX RELAXATION VIA NONCONVEX OPTIMIZATION*

YUXIN CHEN†, YUEJIE CHI‡, JIANQING FAN§, CONG MA§, AND YULING YAN§

Abstract. This paper studies noisy low-rank matrix completion: given partial and noisy entries of a large low-rank matrix, the goal is to estimate the underlying matrix faithfully and efficiently. Arguably one of the most popular paradigms to tackle this problem is convex relaxation, which achieves remarkable efficacy in practice. However, the theoretical support of this approach is still far from optimal in the noisy setting, falling short of explaining its empirical success. We make progress towards demystifying the practical efficacy of convex relaxation vis-à-vis random noise. When the rank and the condition number of the unknown matrix are bounded by a constant, we demonstrate that the convex programming approach achieves near-optimal estimation errors—in terms of the Euclidean loss, the entrywise loss, and the spectral norm loss—for a wide range of noise levels. All of this is enabled by bridging convex relaxation with the nonconvex Burer–Monteiro approach, a seemingly distinct algorithmic paradigm that is provably robust against noise. More specifically, we show that an approximate critical point of the nonconvex formulation serves as an extremely tight approximation of the convex solution, thus allowing us to transfer the desired statistical guarantees of the nonconvex approach to its convex counterpart.

Key words. matrix completion, minimaxity, stability, convex relaxation, nonconvex optimization, Burer-Monteiro approach

AMS subject classifications. 90C25, 90C26

DOI. 10.1137/19M1290000

1. Introduction.

Suppose we are interested in a large low-rank data matrix, but only get to observe a highly incomplete subset of its entries. Can we hope to estimate the underlying data matrix in a reliable manner? This problem, often dubbed as *low-rank matrix completion*, spans a diverse array of science and engineering applications (e.g., collaborative filtering [81], localization [85], system identification [70], magnetic resonance parameter mapping [98], joint alignment [21]), and has inspired a flurry of research activities in the past decade. In the statistics literature, matrix completion also falls under the category of factor models with a large amount of missing data, which finds numerous statistical applications such as controlling false

*Received by the editors September 27, 2019; accepted for publication (in revised form) July 2, 2020; published electronically October 28, 2020. Proofs in this paper can be found at https://arxiv.org/pdf/1902.07698.pdf.

https://doi.org/10.1137/19M1290000

Funding: The first author is supported in part by AFOSR YIP award FA9550-19-1-0030, by ARO grant W911NF-18-1-0303, by ONR grant N00014-19-1-2120, by NSF grants CCF-1907661 and IIS-1900140, and by the Princeton SEAS innovation award. The second author is supported in part by ONR under grants N00014-18-1-2142 and N00014-19-1-2404, by ARO under grant W911NF-18-1-0303, and by NSF under grants CAREER ECCS-1818571 and CCF-1806154, and by the Princeton SEAS innovation award. The third author is supported in part by NSF grants DMS-1662139 and DMS-1712591, ONR grant N00014-19-1-2120, and NIH grant R01-GM072611-12. This work was done in part while the first author was visiting the Kavli Institute for Theoretical Physics (supported in part by NSF grant PHY-1748958).

†Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (yuxin.chen@princeton.edu).

‡Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA (yuejiechi@cmu.edu).

§Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA (jqfan@princeton.edu, congm@princeton.edu, yulingy@princeton.edu).

discovery rates for dependence data [36, 37, 39, 40], factor-adjusted variable selection [41, 63], principal component regression [3, 44, 58, 78], and large covariance matrix estimation [42, 43]. Recent years have witnessed the development of many tractable algorithms that come with statistical guarantees, with convex relaxation being one of the most popular paradigms [14, 15, 46]. See [25, 34] for an overview of this topic.

This paper focuses on noisy low-rank matrix completion, assuming that the revealed entries are corrupted by a certain amount of noise. Setting the stage, consider the task of estimating a rank-r data matrix $\boldsymbol{M}^\star = [M_{ij}^\star]_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$,[1] and suppose that this needs to be performed on the basis of a subset of noisy entries

$$M_{ij} = M_{ij}^\star + E_{ij}, \qquad (i,j) \in \Omega, \tag{1}$$

where $\Omega \subseteq \{1,\dots,n\} \times \{1,\dots,n\}$ denotes a set of indices, and E_{ij} stands for the additive noise at the location (i,j). As we shall elaborate shortly, solving noisy matrix completion via convex relaxation, while practically exhibiting excellent stability (in terms of the estimation errors against noise), is far less understood theoretically compared to the noiseless setting.

1.1. Convex relaxation: Limitations of prior results. Naturally, one would search for a low-rank solution that best fits the observed entries. One choice is the regularized least-squares formulation given by

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \qquad \frac{1}{2}\sum_{(i,j)\in\Omega}\big(Z_{ij}-M_{ij}\big)^2 + \lambda\,\mathsf{rank}(\boldsymbol{Z}), \tag{2}$$

where $\lambda > 0$ is some regularization parameter. In words, this approach optimizes certain trade-offs between the goodness of fit (through the squared loss expressed in the first term of (2)) and the low-rank structure (through the rank function in the second term of (2)). Due to computational intractability of rank minimization, we often resort to convex relaxation in order to obtain computationally feasible solutions. One notable example is the following convex program:

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \qquad g(\boldsymbol{Z}) \triangleq \frac{1}{2}\sum_{(i,j)\in\Omega}\big(Z_{ij}-M_{ij}\big)^2 + \lambda\,\|\boldsymbol{Z}\|_*\,, \tag{3}$$

where $\|\boldsymbol{Z}\|_*$ denotes the nuclear norm (i.e., the sum of singular values) of $\boldsymbol{Z}$—a convex surrogate for the rank function. A significant portion of existing theory supports the use of this paradigm in the noiseless setting: when E_{ij} vanishes for all $(i,j) \in \Omega$, the solution to (3) is known to be faithful (i.e., the estimation error becomes zero) even under near-minimal sample complexity [13, 14, 15, 20, 51, 79].

By contrast, the performance of convex relaxation remains largely unclear when it comes to noisy settings (which are often more practically relevant). Candès and Plan [13] first studied the stability of an equivalent variant[2] of (3) against noise. The estimation error $\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\|_{\mathrm{F}}$ derived therein, of the solution $\boldsymbol{Z}_{\mathsf{cvx}}$ to (3), is significantly larger than the oracle lower bound. This does not explain well the effectiveness

[1] It is straightforward to rephrase our discussions to a general rectangular matrix of size $n_1 \times n_2$. The current paper sets $n = n_1 = n_2$ throughout for simplicity of presentation.

[2] Technically, [13] deals with the constrained version of (3), which is equivalent to the Lagrangian form as in (3) with a proper choice of the regularization parameter.

of (3) in practice. In fact, the numerical experiments reported in [13] already indicated that the performance of convex relaxation is far better than their theoretical bounds. This discrepancy between numerical performance and existing theoretical bounds gives rise to the following natural yet challenging questions: *Where does the convex program* (3) *stand in terms of its stability vis-à-vis additive noise? Can we establish statistical performance guarantees that match its practical effectiveness?*

We note in passing that several other convex relaxation formulations have been thoroughly analyzed for noisy matrix completion, most notably by Negahban and Wainwright [74] and by Koltchinskii, Lounici, and Tsybakov [64]. These works have significantly advanced our understanding of the power of convex relaxation. However, the estimators studied therein, particularly the one in [64], are quite different from the one (3) considered here; as a consequence, the analysis therein does not lead to improved statistical guarantees of (3). Moreover, the performance guarantees provided for these variants are also suboptimal when restricted to the class of "incoherent" or "delocalized" matrices, unless the magnitudes of the noise are fairly large. See section 1.4 for more detailed discussions as well as numerical comparisons of these algorithms.

1.2. A detour: Nonconvex optimization. While the focus of the current paper is convex relaxation, we take a moment to discuss a seemingly distinct algorithmic paradigm: nonconvex optimization, which turns out to be remarkably helpful in understanding convex relaxation. Inspired by the Burer–Monteiro approach [7], the nonconvex scheme starts by representing the rank-r decision matrix (or parameters) $\boldsymbol{Z}$ as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{Y}^\top$ via low-rank factors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}$, and proceeds by solving the following nonconvex (regularized) least-squares problem [60]

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad \frac{1}{2}\sum_{(i,j)\in\Omega}\left[\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)_{ij} - M_{ij}\right]^2 + \mathsf{reg}(\boldsymbol{X},\boldsymbol{Y}). \tag{4}$$

Here, $\mathsf{reg}(\cdot,\cdot)$ denotes a certain regularization term that promotes additional structural properties.

To see its intimate connection with the convex program (3), we make the following observation: if the solution to (3) has rank r, then it must coincide with the solution to

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad \frac{1}{2}\sum_{(i,j)\in\Omega}\left[\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)_{ij} - M_{ij}\right]^2 + \underbrace{\frac{\lambda}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{Y}\|_{\mathrm{F}}^2}_{\mathsf{reg}(\boldsymbol{X},\boldsymbol{Y})}. \tag{5}$$

This can be easily verified by recognizing the elementary fact that

$$\|\boldsymbol{Z}\|_* = \inf_{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}:\boldsymbol{X}\boldsymbol{Y}^\top=\boldsymbol{Z}} \left\{\tfrac{1}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \tfrac{1}{2}\|\boldsymbol{Y}\|_{\mathrm{F}}^2\right\} \tag{6}$$

for any rank-r matrix $\boldsymbol{Z}$ [73, 86]. Note, however, that it is very challenging to predict when the key assumption in establishing this connection—namely, the rank-r assumption of the solution to the convex program (3)—can possibly hold (and, in particular, whether it can hold under minimal sample complexity requirement).

Despite the nonconvexity of (4), simple first-order optimization methods, in conjunction with proper initialization, are often effective in solving (4). Partial examples include gradient descent on manifold [60, 61, 95], gradient descent [71, 87], and projected gradient descent [31, 100]. Apart from their practical efficiency, the nonconvex

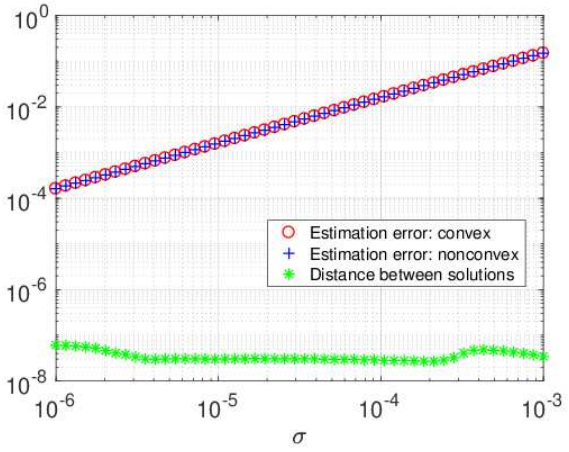

FIG. 1. *The relative estimation errors of both* $\boldsymbol{Z}_{\mathsf{cvx}}$ *(the estimate of the convex program* (3)*) and* $\boldsymbol{Z}_{\mathsf{ncvx}}$ *(the estimate returned by the nonconvex approach tailored to* (5)*) and the relative distance between them versus the standard deviation* σ *of the noise. The results are reported for* $n = 1000$*,* $r = 5$*,* $p = 0.2$*,* $\lambda = 5\sigma\sqrt{np}$*, and are averaged over* 20 *independent trials.*

optimization approach is also appealing in theory. To begin with, algorithms tailored to (4) often enable exact recovery in the noiseless setting. Perhaps more importantly, for a wide range of noise settings, the nonconvex approach achieves appealing estimation accuracy [31, 71], which could be significantly better than those bounds derived for convex relaxation discussed earlier. See [25, 33] for a summary of recent results. The appealing estimation accuracy together with the lower computational cost makes nonconvex approaches suitable for large-scale problems. Having said that, the convex approach is also widely used in practice. The convex programming is often observed to enjoy better stability against model mismatch (e.g., nonuniform sampling of the matrix entries [32]), which makes it appealing for moderate-size problems. In contrast, the theoretical understanding of the effectiveness of the convex method does not explain well its empirical performance; see section 1.1. This motivates us to take a closer inspection of the underlying connection between the two contrasting algorithmic frameworks in the hope that one can utilize the existing theory for the nonconvex approach to improve the stability analysis of the convex relaxation approach.

1.3. Empirical evidence: Convex and nonconvex solutions are often close. In order to obtain a better sense of the relationships between convex and nonconvex approaches, we begin by comparing the estimates returned by the two approaches via numerical experiments. Fix $n = 1000$ and $r = 5$. We generate $\boldsymbol{M}^\star = \boldsymbol{X}^\star \boldsymbol{Y}^{\star\top}$, where $\boldsymbol{X}^\star, \boldsymbol{Y}^\star \in \mathbb{R}^{n\times r}$ are random orthonormal matrices. Each entry $M^\star_{ij}$ of $\boldsymbol{M}^\star$ is observed with probability $p = 0.2$ independently, and then corrupted by an independent Gaussian noise $E_{ij} \sim \mathcal{N}(0, \sigma^2)$. Throughout the experiments, we set $\lambda = 5\sigma\sqrt{np}$. The convex program (3) is solved by the proximal gradient method [76], whereas we attempt solving the nonconvex formulation (5) by gradient descent with spectral initialization (see [33] for details). Let $\boldsymbol{Z}_{\mathsf{cvx}}$ (resp., $\boldsymbol{Z}_{\mathsf{ncvx}} = \boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top$) be the solution returned by the convex program (3) (resp., the nonconvex program (5)). Figure 1 displays the relative estimation errors of both methods ($\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\|_{\mathrm{F}}/\|\boldsymbol{M}^\star\|_{\mathrm{F}}$ and $\|\boldsymbol{Z}_{\mathsf{ncvx}} - \boldsymbol{M}^\star\|_{\mathrm{F}}/\|\boldsymbol{M}^\star\|_{\mathrm{F}}$) as well as the relative distance $\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{Z}_{\mathsf{ncvx}}\|_{\mathrm{F}}/\|\boldsymbol{M}^\star\|_{\mathrm{F}}$ between the two estimates. The results are averaged over 20 independent trials.

Interestingly, the distance between the convex and the nonconvex solutions seems extremely small (e.g., $\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{Z}_{\mathsf{ncvx}}\|_{\mathrm{F}}/\|\boldsymbol{M}^{\star}\|_{\mathrm{F}}$ is typically below 10^{-7}); in comparison, the relative estimation errors of both $\boldsymbol{Z}_{\mathsf{cvx}}$ and $\boldsymbol{Z}_{\mathsf{ncvx}}$ are substantially larger. In other words, the estimate returned by the nonconvex approach serves as a remarkably accurate approximation of the convex solution. Given that the nonconvex approach is often guaranteed to achieve intriguing statistical guarantees vis-à-vis random noise [71], this suggests that the convex program is equally stable—a phenomenon that was not captured by prior theory [13]. *Can we leverage existing theory for the nonconvex scheme to improve the statistical analysis of the convex relaxation approach?*

Before continuing, we remark that the above numerical connection between convex relaxation (3) and nonconvex optimization (5) has already been observed multiple times in prior literature [45, 61, 73, 80, 86]. Nevertheless, all prior observations on this connection were either completely empirical, or provided in a way that does not lead to improved statistical error bounds of the convex paradigm (3). In fact, the difficulty in rigorously justifying the above numerical observations has been noted in the literature; see, e.g., [61].[3]

1.4. Models and main results. The numerical experiments reported in section 1.3 suggest an alternative route for analyzing convex relaxation for noisy matrix completion. If one can formally justify the proximity between the convex and the nonconvex solutions, then it is possible to propagate the appealing stability guarantees from the nonconvex scheme to the convex approach. As it turns out, this simple idea leads to significantly enhanced statistical guarantees for the convex program (3), which we formally present in this subsection.

1.4.1. Models and assumptions. Before proceeding, we introduce a few model assumptions that play a crucial role in our theory.

Assumption 1.

(a) Random sampling: Each index (i,j) belongs to the index set Ω independently with probability p.

(b) Random noise: The noise matrix $\boldsymbol{E}=[E_{ij}]_{1\le i,j\le n}$ is composed of independent and identically distributed (i.i.d.) zero-mean sub-Gaussian random variables with sub-Gaussian norm at most $\sigma>0$, i.e., $\|E_{ij}\|_{\psi_2}\le\sigma$ (see [93, Definition 5.7]).

In addition, let $\boldsymbol{M}^{\star}=\boldsymbol{U}^{\star}\boldsymbol{\Sigma}^{\star}\boldsymbol{V}^{\star\top}$ be the singular value decomposition (SVD) of $\boldsymbol{M}^{\star}$, where $\boldsymbol{U}^{\star},\boldsymbol{V}^{\star}\in\mathbb{R}^{n\times r}$ consist of orthonormal columns and $\boldsymbol{\Sigma}^{\star}=\mathsf{diag}(\sigma_1^{\star},\sigma_2^{\star},\ldots,\sigma_r^{\star})\in\mathbb{R}^{r\times r}$ is a diagonal matrix obeying $\sigma_{\max}\triangleq\sigma_1^{\star}\ge\sigma_2^{\star}\ge\cdots\ge\sigma_r^{\star}\triangleq\sigma_{\min}$. Denote by $\kappa\triangleq\sigma_{\max}/\sigma_{\min}$ the condition number of $\boldsymbol{M}^{\star}$. We impose the following incoherence condition on $\boldsymbol{M}^{\star}$, which is known to be crucial for reliable recovery of $\boldsymbol{M}^{\star}$ [14, 20].

Definition 1.1. *A rank-r matrix $\boldsymbol{M}^{\star}\in\mathbb{R}^{n\times n}$ with SVD $\boldsymbol{M}^{\star}=\boldsymbol{U}^{\star}\boldsymbol{\Sigma}^{\star}\boldsymbol{V}^{\star\top}$ is said to be μ-incoherent if*

$$\|\boldsymbol{U}^{\star}\|_{2,\infty}\le\sqrt{\frac{\mu}{n}}\,\|\boldsymbol{U}^{\star}\|_{\mathrm{F}}=\sqrt{\frac{\mu r}{n}}\qquad\text{and}\qquad\|\boldsymbol{V}^{\star}\|_{2,\infty}\le\sqrt{\frac{\mu}{n}}\,\|\boldsymbol{V}^{\star}\|_{\mathrm{F}}=\sqrt{\frac{\mu r}{n}}.$$

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

[3] The seminal work [61] by Keshavan, Montanari and Oh stated that "*In view of the identity* (6) *it might be possible to use the results in this paper to prove stronger guarantees on the nuclear norm minimization approach. Unfortunately this implication is not immediate ... Trying to establish such an implication, and clarifying the relation between the two approaches is nevertheless a promising research direction.*"

Remark 1.2. It is worth noting that several other conditions on the low-rank matrix have been proposed in the noisy setting. Examples include the spikiness condition [74] and the bounded ℓ_∞ norm condition [64]. However, these conditions alone are often unable to ensure identifiability of the true matrix even in the absence of noise.

1.4.2. Theoretical guarantees: When both the rank and the condition number are constants. With these in place, we are positioned to present our improved statistical guarantees for convex relaxation. For convenience of presentation, we shall begin with a simple yet fundamentally important class of settings when the rank r and the condition number κ are both fixed constants. As it turns out, this class of problems arises in a variety of engineering applications. For example, in a fundamental problem in cryo-EM called angular synchronization [84], one needs to deal with rank-2 or rank-3 matrices with $\kappa = 1$; in a joint shape mapping problem that arises in computer graphics [29, 54], the matrix under consideration has low rank and a condition number equal to 1; and in structure, from motion in computer vision [90], one often seeks to estimate a matrix with $r \le 3$ and a small condition number. Encouragingly, our theory delivers near-optimal statistical guarantees for such practically important scenarios.

THEOREM 1.3. *Let $\boldsymbol{M}^\star$ be rank-r and μ-incoherent with a condition number κ, where the rank and the condition number satisfy $r, \kappa = O(1)$.*[4] *Suppose that Assumption 1 holds and take $\lambda = C_\lambda \sigma \sqrt{np}$ in (3) for some large enough constant $C_\lambda > 0$. Assume the sample size obeys $n^2 p \ge C\mu^2 n \log^3 n$ for some sufficiently large constant $C > 0$, and the noise satisfies $\sigma \lesssim \sqrt{\frac{np}{\mu^3 \log n}} \|\boldsymbol{M}^\star\|_\infty$ for some sufficiently small constant $c > 0$. Then with probability exceeding $1 - O(n^{-3})$,*

1. *any minimizer $\boldsymbol{Z}_{\mathsf{cvx}}$ of (3) obeys*

(7a)
$$\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \left\|\boldsymbol{M}^\star\right\|_{\mathrm{F}}; \quad \left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\| \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \left\|\boldsymbol{M}^\star\right\|,$$

(7b)
$$\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\infty} \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\mu n \log n}{p}} \left\|\boldsymbol{M}^\star\right\|_{\infty};$$

2. *letting $\boldsymbol{Z}_{\mathsf{cvx},r} \triangleq \arg\min_{\boldsymbol{Z}:\mathsf{rank}(\boldsymbol{Z}) \le r} \|\boldsymbol{Z} - \boldsymbol{Z}_{\mathsf{cvx}}\|_{\mathrm{F}}$ be the best rank-r approximation of $\boldsymbol{Z}_{\mathsf{cvx}}$, we have*[5]

$$\|\boldsymbol{Z}_{\mathsf{cvx},r} - \boldsymbol{Z}_{\mathsf{cvx}}\|_{\mathrm{F}} \le \frac{1}{n^3} \cdot \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \left\|\boldsymbol{M}^\star\right\|, \tag{8}$$

and the error bounds in (7) continue to hold if $\boldsymbol{Z}_{\mathsf{cvx}}$ is replaced by $\boldsymbol{Z}_{\mathsf{cvx},r}$.

To explain the applicability of the above theorem, we first remark on the conditions required for this theorem to hold; for simplicity, we assume that $\mu = O(1)$.

[4] Here and throughout, $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ means $|f(n)|/|g(n)| \le C$ for some constant $C > 0$ when n is sufficiently large; $f(n) \gtrsim g(n)$ means $|f(n)|/|g(n)| \ge C$ for some constant $C > 0$ when n is sufficiently large; and $f(n) \asymp g(n)$ if and only if $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$. In addition, $\|\cdot\|_\infty$ denotes the entrywise ℓ_∞ norm, whereas $\|\cdot\|$ is the spectral norm.

[5] The factor $1/n^3$ in (8) can be replaced by $1/n^c$ for an arbitrarily large fixed constant $c > 0$ (e.g., $c = 100$).

- *Sample complexity.* To begin with, the sample size needs to exceed the order of $n\text{poly}\log n$, which is information-theoretically optimal up to some logarithmic term [15].
- *Noise size.* We then turn attention to the noise requirement, i.e.,

$$\sigma \lesssim \sqrt{np/(\log n)}\,\|\boldsymbol{M}^\star\|_\infty .$$

  Note that under the sample size condition $n^2 p \geq Cn\log^3 n$, the size of the noise in each entry is allowed to be substantially larger than the maximum entry in the matrix. In other words, the signal-to-noise ratio w.r.t. each observed entry could be very small. According to prior literature (e.g., [61, Theorem 1.1] and [71, Theorem 2]), such noise conditions are typically required for spectral methods to perform noticeably better than random guessing.
- *Regularization parameter.* In the end, we remark on the choice of the regularization parameter λ. As in most regularized estimators, we need to pick the regularization parameter λ large enough so as to suppress the noise $\boldsymbol{E}$ (which controls the variance), and small enough so as not to shrink the signal $\boldsymbol{M}^\star$ too much (which controls the estimation bias). It turns out that setting $\lambda \asymp \sigma\sqrt{np}$ achieves the desired bias-variance trade-off; see Lemma 3.

Further, Theorem 1.3 has several important implications about the power of convex relaxation. The discussions below again concentrate on the case where $\mu = O(1)$.

- *Near-optimal stability guarantees.* Our results reveal that the Euclidean error of any convex optimizer $\boldsymbol{Z}_{\mathsf{cvx}}$ of (3) obeys

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim \sigma\sqrt{n/p}, \tag{9}$$

  implying that the performance of convex relaxation degrades gracefully as the signal-to-noise ratio decreases. This result matches the oracle lower bound derived in [13, eq. (III.13)], which also improves upon their statistical guarantee. Specifically, Candès and Plan [13] provided a stability guarantee in the presence of arbitrary bounded noise. When applied to the random noise model assumed here, their results yield $\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim \sigma n^{3/2}$, which could be $O(\sqrt{n^2 p})$ times more conservative than our bound (9).
- *Nearly low-rank structure of the convex solution.* In light of (8), the optimizer of the convex program (3) is almost, if not exactly, rank-r. When the true rank r is known a priori, it is not uncommon for practitioners to return the rank-r approximation of $\boldsymbol{Z}_{\mathsf{cvx}}$. Our theorem formally justifies that there is no loss of statistical accuracy—measured in terms of either $\|\cdot\|_{\mathrm{F}}$ or $\|\cdot\|_\infty$—when performing the rank-r projection operation.
- *Entrywise and spectral norm error control.* Moving beyond the Euclidean loss, our theory uncovers that the estimation errors of the convex optimizer are fairly spread out across all entries, thus implying near-optimal entrywise error control. This is a stronger form of error bounds, as an optimal Euclidean estimation accuracy alone does not preclude the possibility of the estimation errors being spiky and localized. Furthermore, the spectral norm error of the convex optimizer is also well-controlled. See Figure 2 for the numerical support.
- *Implicit regularization.* As a byproduct of the entrywise error control, this result indicates that the additional constraint $\|\boldsymbol{Z}\|_\infty \leq \alpha$ suggested by [74] is automatically satisfied and is hence unnecessary. In other words, the convex approach implicitly controls the spikiness of its entries, without resorting to explicit regularization. This is also confirmed by the numerical experiments

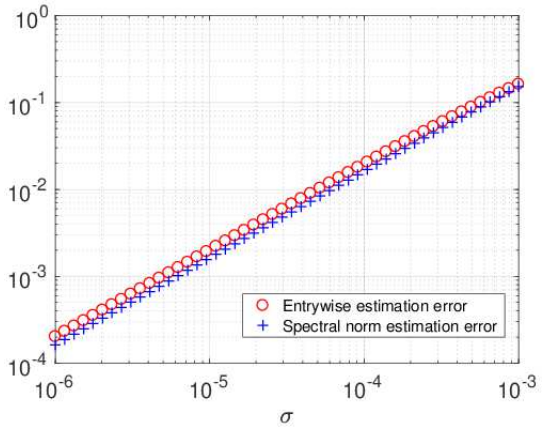

FIG. 2. *The relative estimation error of* $\boldsymbol{Z}_{\mathsf{cvx}}$ *measured by both* $\|\cdot\|_\infty$ *(i.e.,* $\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\|_\infty/\|\boldsymbol{M}^\star\|_\infty$*) and* $\|\cdot\|$ *(i.e.,* $\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\|/\|\boldsymbol{M}^\star\|$*) versus the standard deviation* σ *of the noise. The results are reported for* $n = 1000$*,* $r = 5$*,* $p = 0.2$*,* $\lambda = 5\sigma\sqrt{np}$*, and are averaged over* 20 *independent trials.*

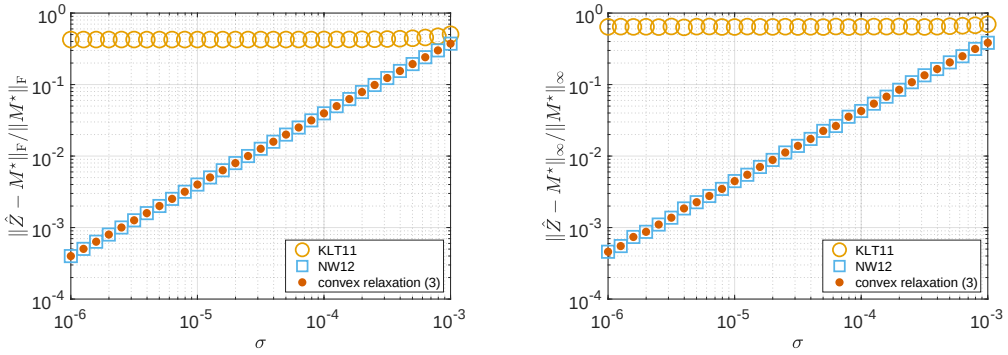

FIG. 3. *The relative estimation errors of* $\hat{\boldsymbol{Z}}$*, measured in terms of* ℓ_{F} *and* ℓ_∞*, versus the standard deviation* σ *of the noise. Here* $\hat{\boldsymbol{Z}}$ *can be either the modified convex estimator in* [64]*, the constrained convex estimator in* [74]*, or the vanilla convex estimator* (3)*. The results are reported for* $n = 1000$*,* $r = 5$*,* $p = 0.2$*, and are averaged over* 20 *Monte Carlo trials. For the modified convex estimator in* [64]*, we choose the regularization parameter* λ *therein to be* $1.5\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{1/(n^3p)}$*, as suggested by their theory. For the constrained one in* [74]*, the regularization parameter* λ *is set to be* $5\sigma\sqrt{np}$ *and the constraint* α *is set to be* $\|\boldsymbol{M}^\star\|_\infty$*. Both choices are recommended by* [74]*. As for* (3)*, we set* $\lambda = 5\sigma\sqrt{np}$*.*

reported in Figure 3, where we see that the estimation error of (3) and that of the constrained version considered in [74] are nearly identical.

- *Statistical guarantees for fast iterative optimization methods.* Various iterative algorithms have been developed to solve the nuclear norm regularized least-squares problem (3) up to an arbitrarily prescribed accuracy, examples including SVT (or proximal gradient methods) [8], FPC [72], SOFT–IMPUTE [73], FISTA [5, 89], to name just a few. Our theory immediately provides statistical guarantees for these algorithms. As we shall make precise in section 2, any point $\boldsymbol{Z}$ with $g(\boldsymbol{Z}) \le g(\boldsymbol{Z}_{\mathsf{cvx}}) + \varepsilon$ (where $g(\cdot)$ is defined in (3)) enjoys the same

error bounds as in (7) (with $\boldsymbol{Z}_{\mathsf{cvx}}$ replaced by $\boldsymbol{Z}$ in (7)), provided that $\varepsilon > 0$ is sufficiently small. In other words, when these convex optimization algorithms converge w.r.t. the objective value, they are guaranteed to return a statistically reliable estimate.

To better understand our contributions, we take a moment to discuss two important but different convex programs studied in [74] and [64]. To begin with, under a spikiness assumption on the low-rank matrix, Negahban and Wainwright [74] proposed to enforce an extra entrywise constraint $\|\boldsymbol{Z}\|_\infty \le \alpha$ when solving (3), in order to explicitly control the spikiness of the estimate. When applied to our model with $r, \kappa, \mu \asymp 1$, their results read (up to some logarithmic factor)

$$\left\|\hat{\boldsymbol{Z}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim \max\left\{\sigma, \|\boldsymbol{M}^\star\|_\infty\right\}\sqrt{n/p}, \tag{10}$$

where $\hat{\boldsymbol{Z}}$ is the estimate returned by their modified convex algorithm. While this matches the optimal bound when $\sigma \gtrsim \|\boldsymbol{M}^\star\|_\infty$, it becomes suboptimal when $\sigma \ll \|\boldsymbol{M}^\star\|_\infty$ (under our models). Moreover, as we have already discussed, the extra spikiness constraint becomes unnecessary in the regime considered herein. This also means that our result complements existing theory about the convex program in [74] by demonstrating its minimaxity for an additional range of noise. Another work by Koltchinskii, Lounici, and Tsybakov [64] investigated a completely different convex algorithm, which is effectively a spectral method (namely, one round of soft singular value thresholding on a rescaled zero-padded data matrix). The algorithm is shown to be minimax optimal over the class of low-rank matrices with bounded ℓ_∞ norm (note that this is very different from the set of incoherent matrices studied here). When specialized to our model, their error bound is the same as (10) (modulo some log factor), which also becomes suboptimal as σ decreases. The advantage of the convex program (3) is shown in Figure 3.

1.4.3. Theoretical guarantees: Extensions to more general settings. So far we have presented results when the true matrix has bounded rank and condition number, i.e., $r, \kappa = O(1)$. Our theory actually accommodates a significantly broader range of scenarios, where the rank and the condition number are both allowed to grow with the dimension n.

THEOREM 1.4. *Let $\boldsymbol{M}^\star$ be rank-r and μ-incoherent with a condition number κ. Suppose Assumption 1 holds and take $\lambda = C_\lambda \sigma\sqrt{np}$ in (3) for some large enough constant $C_\lambda > 0$. Assume the sample size obeys $n^2 p \ge C\kappa^4\mu^2 r^2 n \log^3 n$ for some sufficiently large constant $C > 0$, and the noise satisfies $\sigma\sqrt{\frac{n}{p}} \le c\frac{\sigma_{\min}}{\sqrt{\kappa^4\mu r \log n}}$ for some sufficiently small constant $c > 0$. Then with probability exceeding $1 - O(n^{-3})$,*

1. *any minimizer $\boldsymbol{Z}_{\mathsf{cvx}}$ of (3) obeys*

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{M}^\star\right\|_{\mathrm{F}}, \tag{11a}$$

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\infty} \lesssim \sqrt{\kappa^3\mu r}\cdot\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{M}^\star\right\|_{\infty}, \tag{11b}$$

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\| \lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{M}^\star\right\|; \tag{11c}$$

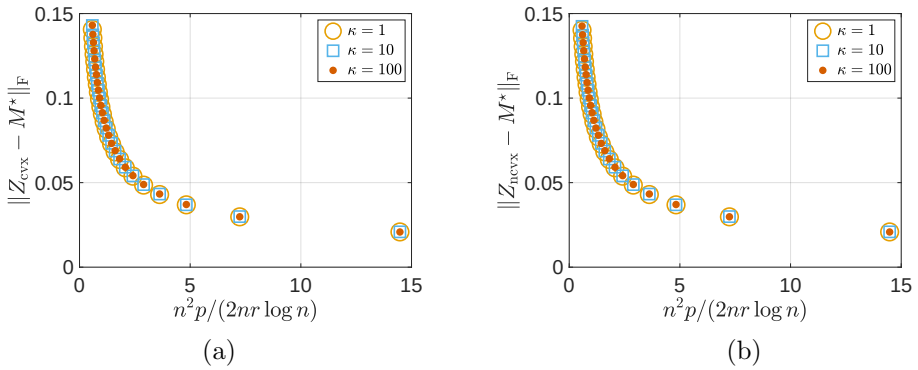

FIG. 4. *The estimation errors of* $\boldsymbol{Z}_{\mathsf{cvx}}$ *and* $\boldsymbol{Z}_{\mathsf{ncvx}}$*, measured in terms of* ℓ_{F}*, versus the rescaled sample complexity* $n^2p/(2nr\log n)$*. The results are reported for* $n = 1000$*,* $p = 0.2$*,* $\sigma = 10^{-4}$*, and are averaged over* 20 *Monte Carlo trials. The rank* r *is varied from* 1 *to* 25 *and the condition number* κ *is chosen from* $\{1, 10, 100\}$*.*

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

$$\|\boldsymbol{Z}_{\mathsf{cvx},r} - \boldsymbol{Z}_{\mathsf{cvx}}\|_{\mathrm{F}} \le \frac{1}{n^3} \cdot \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \left\|\boldsymbol{M}^\star\right\|, \tag{12}$$

and the error bounds in (11) *continue to hold if* $\boldsymbol{Z}_{\mathsf{cvx}}$ *is replaced by* $\boldsymbol{Z}_{\mathsf{cvx},r}$*.*

Remark 1.5 (the noise condition). The incoherence condition (cf. Definition 1.1) guarantees that the largest entry $\|\boldsymbol{M}^\star\|_\infty$ of the matrix $\boldsymbol{M}^\star$ is no larger than $\kappa\mu r\sigma_{\min}/n$. As a result, the noise condition stated in Theorem 1.4 covers all scenarios obeying

$$\sigma \lesssim \sqrt{\frac{np}{\kappa^6\mu^3 r^3 \log n}} \left\|\boldsymbol{M}^\star\right\|_\infty .$$

Therefore, the typical size of the noise is allowed to be much larger than the size of the largest entry of $\boldsymbol{M}^\star$, provided that $p \gg \frac{\kappa^6\mu^3 r^3 \log n}{n}$. In particular, when $r, \kappa = O(1)$, this recovers the noise condition in Theorem 1.3.

Notably, the sample size condition for noisy matrix completion (i.e., $n^2p \ge C\kappa^4\mu^2 r^2 n \log^3 n$) is more stringent than that in the noiseless setting (i.e., $n^2p \asymp nr\log^2 n$), and our statistical guarantees are likely suboptimal with respect to the dependency on r and κ. It turns out that both convex and nonconvex methods work well numerically even when the number of samples n^2p is on the order of $nr\log n$, which is much smaller than the required sample complexity $\kappa^4 r^2 n\log^3 n$ in Theorem 1.4; see Figure 4 for an illustration. From a technical point of view, this suboptimality is mainly due to the analysis of nonconvex optimization, a key ingredient of our analysis of convex relaxation. In fact, the state-of-the-art nonconvex analysis [31, 61, 71] requires the sample size to be much larger than the optimal one (e.g., $n^2p \gg n\mathsf{poly}(r)\mathsf{poly}(\kappa)$) even in the noiseless setting. It would certainly be interesting, and in fact important, to see whether it is possible to develop a theory with optimal dependency on r and κ. We leave this for future investigation.

It is also instrumental to compare our sample complexity and error bounds with those in the prior literature [10, 62, 64, 74]. See Table 1 for a comparison when the

TABLE 1
Comparisons of sample complexities and Euclidean estimation errors when $\mu = O(1)$.

	Sample complexity	Euclidean estimation error
[64]	$n \log^2 n$	$\max\left(\sigma, \|\boldsymbol{M}^\star\|_\infty\right) \sqrt{(nr \log n)/p}$
[74]	$n \log n$	$\max\left(\sigma, 1/n\right) \alpha^\star \sqrt{(nr \log n)/p}$
[62]	$n \log^3 n$	$\max\left(\sigma, \|\boldsymbol{M}^\star\|_\infty\right) \sqrt{(nr \log n)/p}$
[10]	n	$\sqrt{\max\left(\sigma, \|\boldsymbol{M}^\star\|_\infty\right)}\, \|\boldsymbol{M}^\star\|_{\max}\, n^{3/4}/p^{1/4}$
Ours	$\kappa^4 r^2 n \log^3 n$	$\kappa^2 \sigma \sqrt{nr/p}$

incoherence parameter μ is $O(1)$. Here

$$\alpha^\star \triangleq n\frac{\|\boldsymbol{M}^\star\|_\infty}{\|\boldsymbol{M}^\star\|_{\mathrm{F}}} \qquad \text{and} \qquad \|\boldsymbol{M}^\star\|_{\max} \triangleq \min_{\boldsymbol{M}^\star = \boldsymbol{U}\boldsymbol{V}^\top} \|\boldsymbol{U}\|_{2,\infty} \|\boldsymbol{V}\|_{2,\infty}.$$

Indeed, all the papers [10, 62, 64, 74] studied convex relaxation approaches for noisy matrix completion. However, there are two main differences from our analysis here. The first is regarding the convex method itself. All four papers [10, 62, 64, 74] require extra knowledge about the underlying low-rank matrix for the convex approach, e.g., the ℓ_∞ norm of the low-rank matrix in [10, 62, 74] and the sharpness constant in [74]. This results in different convex formulations from what we analyze here. The second major difference lies in the performance guarantees. In both the well-conditioned and the highly ill-conditioned regimes, the estimation error provided in [10, 62, 64, 74] does not vanish as the size of the noise decreases to zero. This is in stark contrast to our theory that guarantees the stability of the convex approach for noisy matrix completion.

Despite the above suboptimality issue, implications similar to those of Theorem 1.3 hold for this general setting. To begin with, the nearly low-rank structure of the convex solution is preserved (cf. (12)). In addition, the estimation error of the convex estimate is spread out across entries (cf. (11b)), thus uncovering an implicit regularization phenomenon underlying convex relaxation (which implicitly regularizes the spikiness constraint on the solution). Last but not least, the upper bounds (11) and (12) continue to hold for approximate minimizers of the convex program (3), thus yielding statistical guarantees for numerous iterative algorithms aimed at minimizing (3).

1.5. Numerical experiments. This subsection collects numerical supports for our theoretical findings in section 1.4.

Entrywise and spectral norm error of the convex approach. The experimental setting is similar to that in producing Figure 1. For completeness, we repeat it here. Fix $n = 1000$ and $r = 5$. We generate $\boldsymbol{M}^\star = \boldsymbol{X}^\star \boldsymbol{Y}^{\star\top}$, where $\boldsymbol{X}^\star, \boldsymbol{Y}^\star \in \mathbb{R}^{n\times r}$ are random orthonormal matrices. Each entry $M_{ij}^\star$ of $\boldsymbol{M}^\star$ is observed with probability $p = 0.2$ independently, and then corrupted by an independent Gaussian noise $E_{ij} \sim \mathcal{N}(0, \sigma^2)$. Throughout the experiments, we set $\lambda = 5\sigma\sqrt{np}$. The convex program (3) is solved by the proximal gradient method [76]. Figure 2 displays the relative estimation errors of the convex approach (3) in both the ℓ_∞ norm and the spectral norm. As can be seen, both forms of estimation errors scale linearly with the noise level, corroborating our theory.

Comparisons with other convex approaches. Utilizing the same experimental setting as before, we compare the numerical performance of (3) with two important but different convex programs studied in [74] and [64]. For the modified convex estimator in [64], we choose the regularization parameter λ therein to be $1.5\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\cdot\sqrt{1/(n^3p)}$, as suggested by their theory. For the constrained one in [74], the regularization parameter λ is set to be $5\sigma\sqrt{np}$ and the constraint α is set to be $\|\boldsymbol{M}^\star\|_\infty$. Both choices are recommended by [74]. Figure 3 displays the relative Euclidean and entrywise estimation error of the three convex programs. As can be seen, the estimation error of this thresholding-based spectral algorithm [64] does not decrease as the noise shrinks, and its performance seems uniformly outperformed by that of convex relaxation (3) and the constrained estimator in [74]. In fact, this is part of our motivation to pursue an improved theoretical understanding of the formulation (3).

Numerical sample complexity of both convex and nonconvex methods.. Theorem 1.4 requires the sample complexity n^2p to exceed $\kappa^4 r^2 n\log^3 n$, which we believe could be improved. To justify this, under a similar setting to that of Figure 1, we plot the estimation errors of $\boldsymbol{Z}_{\mathsf{cvx}}$ and $\boldsymbol{Z}_{\mathsf{ncvx}}$ versus the rescaled sample complexity $n^2p/(2nr\log n)$ in Figure 4. It can be seen that the estimation errors degrade gracefully as the rescaled sample size gets smaller and the performance does not change w.r.t. the condition number κ.

2. Strategy and novelty. In this section, we introduce the strategy for proving our main theorem, i.e., Theorem 1.4. Theorem 1.3 follows immediately. Informally, the main technical difficulty stems from the lack of closed-form expressions for the primal solution to (3), which in turn makes it difficult to construct a dual certificate. This is in stark contrast to the noiseless setting, where one clearly anticipates the ground truth $\boldsymbol{M}^\star$ to be the primal solution; in fact, this is precisely why the analysis for the noisy case is significantly more challenging. Our strategy, as we shall detail below, mainly entails invoking an iterative nonconvex algorithm to "approximate" such a primal solution.

Before continuing, we introduce a few more notations. Let $\mathcal{P}_\Omega(\cdot): \mathbb{R}^{n\times n}\mapsto\mathbb{R}^{n\times n}$ represent the projection onto the subspace of matrices supported on Ω, namely,

$$[\mathcal{P}_\Omega(\boldsymbol{Z})]_{ij} = \begin{cases} Z_{ij} & \text{for } (i,j)\in\Omega, \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

for any matrix $\boldsymbol{Z}\in\mathbb{R}^{n\times n}$. For a rank-$r$ matrix $\boldsymbol{M}$ with SVD $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, denote by T the tangent space of the rank-r manifold at $\boldsymbol{M}$, i.e.,

$$T = \left\{\boldsymbol{U}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{V}^\top \mid \boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{n\times r}\right\}. \tag{14}$$

Correspondingly, let $\mathcal{P}_T(\cdot)$ be the orthogonal projection onto the subspace T, that is,

$$\mathcal{P}_T(\boldsymbol{Z}) = \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{Z}\boldsymbol{V}\boldsymbol{V}^\top \tag{15}$$

for any matrix $\boldsymbol{Z}\in\mathbb{R}^{n\times n}$. In addition, let $T^\perp$ and $\mathcal{P}_{T^\perp}(\cdot)$ denote the orthogonal complement of T and the projection onto $T^\perp$, respectively. Regarding the ground truth, we denote

$$\boldsymbol{X}^\star = \boldsymbol{U}^\star(\boldsymbol{\Sigma}^\star)^{1/2} \qquad \text{and} \qquad \boldsymbol{Y}^\star = \boldsymbol{V}^\star(\boldsymbol{\Sigma}^\star)^{1/2}. \tag{16}$$

The nonconvex problem (5) is equivalent to

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \qquad f(\boldsymbol{X},\boldsymbol{Y}) \triangleq \frac{1}{2p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{Y}\|_{\mathrm{F}}^2, \tag{17}$$

where we have inserted an extra factor $1/p$ (compared to (5)) to simplify the presentation of the analysis later on.

2.1. Exact duality. In order to analyze the convex program (3), it is natural to start with the first-order optimality condition. Specifically, suppose that $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$ is a (primal) solution to (3) with SVD $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$.[6] As before, let T be the tangent space of $\boldsymbol{Z}$, and let $T^\perp$ be the orthogonal complement of T. Then the first-order optimality condition for (3) reads there exists a matrix $\boldsymbol{W} \in T^\perp$ (called a dual certificate) such that

$$\frac{1}{\lambda}\mathcal{P}_\Omega\big(\boldsymbol{M}-\boldsymbol{Z}\big) = \boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W}, \tag{18a}$$

$$\|\boldsymbol{W}\| \le 1. \tag{18b}$$

This condition is not only necessary to certify the optimality of $\boldsymbol{Z}$, but also "almost sufficient" in guaranteeing the uniqueness of the solution $\boldsymbol{Z}$; see https://arxiv.org/pdf/1902.07698.pdf for in-depth discussions.

The challenge then boils down to identifying such a primal-dual pair $(\boldsymbol{Z}, \boldsymbol{W})$ satisfying the optimality condition (18). For the noise-free case, the primal solution is clearly $\boldsymbol{Z} = \boldsymbol{M}^\star$ if exact recovery is to be expected; the dual certificate can then be either constructed exactly by the least-squares solution to a certain underdetermined linear system [14, 15], or produced approximately via a clever golfing scheme pioneered by Gross [51]. For the noisy case, however, it is often difficult to hypothesize on the primal solution $\boldsymbol{Z}$, as it depends on the random noise in a complicated way. In fact, the lack of a suitable guess of $\boldsymbol{Z}$ (and hence $\boldsymbol{W}$) was the major hurdle that prior works faced when carrying out the duality analysis.

2.2. A candidate primal solution via nonconvex optimization. Motivated by the numerical experiment in section 1.3, we propose to examine whether the optimizer of the nonconvex problem (5) stays close to the solution to the convex program (3). Towards this, suppose that $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}$ form a critical point of (5) with $\mathsf{rank}(\boldsymbol{X}) = \mathsf{rank}(\boldsymbol{Y}) = r$.[7] Then the first-order condition reads

$$\frac{1}{\lambda}\mathcal{P}_\Omega\big(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{Y}^\top\big)\boldsymbol{Y} = \boldsymbol{X}, \tag{19a}$$

$$\frac{1}{\lambda}\left[\mathcal{P}_\Omega\big(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{Y}^\top\big)\right]^\top \boldsymbol{X} = \boldsymbol{Y}. \tag{19b}$$

To develop some intuition about the connection between (18) and (19), let us take a look at the case with $r = 1$. Denote $\boldsymbol{X} = \boldsymbol{x}$ and $\boldsymbol{Y} = \boldsymbol{y}$ and assume that the two rank-1 factors are "balanced," namely, $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2 \neq 0$. It then follows from (19) that $\lambda^{-1}\mathcal{P}_\Omega(\boldsymbol{M}-\boldsymbol{x}\boldsymbol{y}^\top)$ has a singular value 1, whose corresponding left and right singular vectors are $\boldsymbol{x}/\|\boldsymbol{x}\|_2$ and $\boldsymbol{y}/\|\boldsymbol{y}\|_2$, respectively. In other words, one can express

$$\frac{1}{\lambda}\mathcal{P}_\Omega\big(\boldsymbol{M}-\boldsymbol{x}\boldsymbol{y}^\top\big) = \frac{1}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2}\boldsymbol{x}\boldsymbol{y}^\top + \boldsymbol{W}, \tag{20}$$

where $\boldsymbol{W}$ is orthogonal to the tangent space of $\boldsymbol{x}\boldsymbol{y}^\top$; this is precisely the condition (18a). It remains to argue that (18b) is valid as well. Towards this end, the

[6] Here and below, we use $\boldsymbol{Z}$ (rather than $\boldsymbol{Z}_{\mathsf{cvx}}$) for notational simplicity, whenever it is clear from the context.

[7] Once again, we abuse the notation $(\boldsymbol{X}, \boldsymbol{Y})$ (instead of using $(\boldsymbol{X}_{\mathsf{ncvx}}, \boldsymbol{Y}_{\mathsf{ncvx}})$) for notational simplicity, whenever it is clear from the context.

first-order condition (19) alone is insufficient, as there might be nonglobal critical points (e.g., saddle points) that are unable to approximate the convex solution well. Fortunately, as long as the candidate $\boldsymbol{x}\boldsymbol{y}^\top$ is not far away from the ground truth $\boldsymbol{M}^\star$, one can guarantee $\|\boldsymbol{W}\| < 1$ as required in (18b).

The above informal argument about the link between the convex and the nonconvex problems can be formalized. To begin with, we introduce the following conditions on the regularization parameter λ.

Condition 1 (regularization parameter). The regularization parameter λ satisfies
(a) (Relative to noise) $\|\mathcal{P}_\Omega(\boldsymbol{E})\| < \lambda/8$;
(b) (Relative to nonconvex solution) $\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star) - p(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star)\| < \lambda/8$.

Remark 2.1. Condition 1 requires that the regularization parameter λ should dominate a certain norm of the noise, as well as of the deviation of $\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star$ from its mean $p(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star)$; as will be seen shortly, the latter condition can be met when $(\boldsymbol{X}, \boldsymbol{Y})$ is sufficiently close to $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$.

With the above condition in place, the following result demonstrates that a critical point $(\boldsymbol{X}, \boldsymbol{Y})$ of the nonconvex problem (5) readily translates to the unique minimizer of the convex program (3). This lemma is established in https://arxiv.org/pdf/1902.07698.pdf.

LEMMA 2.2 (exact nonconvex versus convex optimizers). *Suppose that* $(\boldsymbol{X}, \boldsymbol{Y})$ *is a critical point of* (5) *satisfying* $\mathsf{rank}(\boldsymbol{X}) = \mathsf{rank}(\boldsymbol{Y}) = r$, *and the sampling operator* $\mathcal{P}_\Omega$ *is injective when restricted to the elements of the tangent space* T *of* $\boldsymbol{X}\boldsymbol{Y}^\top$, *namely,*

$$\mathcal{P}_\Omega(\boldsymbol{H}) = \boldsymbol{0} \quad \Longleftrightarrow \quad \boldsymbol{H} = \boldsymbol{0} \quad \text{for all } \boldsymbol{H} \in T. \tag{21}$$

Under Condition 1, *the point* $\boldsymbol{Z} \triangleq \boldsymbol{X}\boldsymbol{Y}^\top$ *is the unique minimizer of* (3).

In order to apply Lemma 2.2, one needs to locate a critical point of (5) that is sufficiently close to the truth, for which one natural candidate is the global optimizer of (5). The caveat, however, is the lack of theory characterizing directly the properties of the optimizer of (5). Instead, what is available in prior theory is the characterization of some iterative sequence (e.g., gradient descent iterates) aimed at solving (5). It is unclear from prior theory whether the iterative algorithm under study (e.g., gradient descent) converges to the global optimizer in the presence of noise. This leads to technical difficulty in justifying the proximity between the nonconvex optimizer and the convex solution via Lemma 2.2.

2.3. Approximate nonconvex optimizers. Fortunately, perfect knowledge of the nonconvex optimizer is not pivotal. Instead, an approximate solution to the nonconvex problem (5) (or equivalently (17)) suffices to serve as a reasonably tight approximation of the convex solution. More precisely, we desire two factors $(\boldsymbol{X}, \boldsymbol{Y})$ that result in nearly zero (rather than exactly zero) gradients:

$$\nabla_{\boldsymbol{X}} f(\boldsymbol{X}, \boldsymbol{Y}) \approx \boldsymbol{0} \qquad \text{and} \qquad \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}, \boldsymbol{Y}) \approx \boldsymbol{0},$$

where $f(\cdot, \cdot)$ is the nonconvex objective function as defined in (17). This relaxes the condition discussed in Lemma 2.2 (which only applies to critical points of (5) as opposed to approximate critical points). As it turns out, such points can be found via gradient descent tailored to (5). The sufficiency of the near-zero gradient condition is made possible by slightly strengthening the injectivity assumption (21), which is stated below.

Condition 2 (injectivity). Let T be the tangent space of $\boldsymbol{X}\boldsymbol{Y}^\top$. There is a quantity $c_{\text{inj}} > 0$ such that

$$p^{-1}\left\|\mathcal{P}_\Omega\left(\boldsymbol{H}\right)\right\|_{\text{F}}^2 \geq c_{\text{inj}}\left\|\boldsymbol{H}\right\|_{\text{F}}^2 \qquad \text{for all } \boldsymbol{H} \in T. \tag{22}$$

The following lemma states quantitatively how an approximate nonconvex optimizer serves as an excellent proxy of the convex solution, which we establish in https://arxiv.org/pdf/1902.07698.pdf.

LEMMA 2.3 (approximate nonconvex versus convex optimizers). *Suppose that* $(\boldsymbol{X}, \boldsymbol{Y})$ *obeys*

$$\left\|\nabla f\left(\boldsymbol{X}, \boldsymbol{Y}\right)\right\|_{\text{F}} \leq c\frac{\sqrt{c_{\text{inj}}p}}{\kappa} \cdot \frac{\lambda}{p}\sqrt{\sigma_{\min}} \tag{23}$$

for some sufficiently small constant $c > 0$. *Further assume that any singular value of* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *lies in* $[\sqrt{\sigma_{\min}/2}, \sqrt{2\sigma_{\max}}]$. *Then under Conditions* 1 *and* 2, *any minimizer* $\boldsymbol{Z}_{\mathsf{cvx}}$ *of* (3) *satisfies*

$$\left\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{Z}_{\mathsf{cvx}}\right\|_{\text{F}} \lesssim \frac{\kappa}{c_{\text{inj}}}\frac{1}{\sqrt{\sigma_{\min}}}\left\|\nabla f\left(\boldsymbol{X}, \boldsymbol{Y}\right)\right\|_{\text{F}}. \tag{24}$$

Remark 2.4. In fact, this lemma continues to hold if $\boldsymbol{Z}_{\mathsf{cvx}}$ is replaced by any $\boldsymbol{Z}$ obeying $g(\boldsymbol{Z}) \leq g(\boldsymbol{X}\boldsymbol{Y}^\top)$, where $g(\cdot)$ is the objective function defined in (3) and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are low-rank factors obeying conditions of Lemma 2.3. This is important in providing statistical guarantees for iterative methods like SVT [8], FPC [72], SOFT–IMPUTE [73], FISTA [5], etc. To be more specific, suppose that $(\boldsymbol{X}, \boldsymbol{Y})$ results in an approximate optimizer of (3), namely, $g(\boldsymbol{X}\boldsymbol{Y}^\top) = g(\boldsymbol{Z}_{\mathsf{cvx}}) + \varepsilon$ for some sufficiently small $\varepsilon > 0$. Then for any $\boldsymbol{Z}$ obeying $g(\boldsymbol{Z}) \leq g(\boldsymbol{X}\boldsymbol{Y}^\top) = g(\boldsymbol{Z}_{\mathsf{cvx}}) + \varepsilon$, one has

$$\left\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{Z}\right\|_{\text{F}} \lesssim \frac{\kappa}{c_{\text{inj}}}\frac{1}{\sqrt{\sigma_{\min}}}\left\|\nabla f\left(\boldsymbol{X}, \boldsymbol{Y}\right)\right\|_{\text{F}}. \tag{25}$$

As a result, as long as the above-mentioned algorithms converge in terms of the objective value, they must return a solution obeying (25), which is exceedingly close to $\boldsymbol{X}\boldsymbol{Y}^\top$ if $\|\nabla f(\boldsymbol{X}, \boldsymbol{Y})\|_{\text{F}}$ is small.

It is clear from Lemma 2.3 that, as the size of the gradient $\nabla f(\boldsymbol{X}, \boldsymbol{Y})$ gets smaller, the nonconvex estimate $\boldsymbol{X}\boldsymbol{Y}^\top$ becomes an increasingly tighter approximation of any convex optimizer of (3), which is consistent with Lemma 2.2. In contrast to Lemma 2.2, due to the lack of strong convexity, a nonconvex estimate with a near-zero gradient does not imply the uniqueness of the optimizer of the convex program (3); rather, it indicates that any minimizer of (3) lies within a sufficiently small neighborhood surrounding $\boldsymbol{X}\boldsymbol{Y}^\top$ (cf. (24)).

2.4. Construction of an approximate nonconvex optimizer. So far, Lemmas 2.2–2.3 are both deterministic results based on Condition 1. As we will soon see, under Assumption 1, we can derive simpler conditions that—with high probability—guarantee Condition 1. We start with Condition 1(a).

LEMMA 2.5. *Suppose* $n^2 p \geq Cn\log^2 n$ *for some sufficiently large constant* $C > 0$. *Then with probability at least* $1 - O(n^{-10})$, *one has* $\|\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\| \lesssim \sigma\sqrt{np}$. *As a result, Condition* 1 *holds (i.e.,* $\|\mathcal{P}_\Omega(\boldsymbol{E})\| < \lambda/8$*) as long as* $\lambda = C_\lambda\sigma\sqrt{np}$ *for some sufficiently large constant* $C_\lambda > 0$.

Proof. This follows from [31, Lemma 11] with a slight and direct modification to accommodate the asymmetric noise here. For brevity, we omit the proof. ∎

Turning attention to Condition 1(b) and Condition 2, we have the following lemma, the proof of which is deferred to https://arxiv.org/pdf/1902.07698.pdf.

LEMMA 2.6. *Under the assumptions of Theorem* 1.4*, with probability exceeding* $1 - O(n^{-10})$ *we have*

$$\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star) - p(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star)\| < \lambda/8 \qquad \text{(Condition 1(b))},$$

$$\frac{1}{p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{H}\right)\right\|_{\mathrm{F}}^2 \geq \frac{1}{32\kappa}\left\|\boldsymbol{H}\right\|_{\mathrm{F}}^2 \quad \textit{for all } \boldsymbol{H} \in T \qquad \text{(Condition 2 with } c_{\mathrm{inj}} = (32\kappa)^{-1})$$

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

$$\begin{aligned}&\max\left\{\left\|\boldsymbol{X} - \boldsymbol{X}^\star\right\|_{2,\infty}, \left\|\boldsymbol{Y} - \boldsymbol{Y}^\star\right\|_{2,\infty}\right\}\\ &\qquad \leq C_\infty \kappa \left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\max\left\{\left\|\boldsymbol{X}^\star\right\|_{2,\infty}, \left\|\boldsymbol{Y}^\star\right\|_{2,\infty}\right\}. \end{aligned} \tag{26}$$

Here, T *denotes the tangent space of* $\boldsymbol{X}\boldsymbol{Y}^\top$*, and* $C_\infty > 0$ *is some absolute constant.*

This lemma is a uniform result, namely, the bounds hold irrespective of the statistical dependency between $(\boldsymbol{X}, \boldsymbol{Y})$ and Ω. As a consequence, to demonstrate the proximity between the convex and the nonconvex solutions (cf. (24)), it remains to identify a point $(\boldsymbol{X}, \boldsymbol{Y})$ with vanishingly small gradient (cf. (23)) that is sufficiently close to the truth (cf. (26)).

As we already alluded to previously, a simple gradient descent algorithm aimed at solving the nonconvex problem (5) might help us produce an approximate nonconvex optimizer. This procedure is summarized in Algorithm 1. Our hope is this: when initialized at the ground truth and run for sufficiently many iterations, the gradient descent (GD) trajectory produced by Algorithm 1 will contain at least one approximate stationary point of (5) with the desired properties (23) and (26). We shall note that Algorithm 1 is *not practical* since it starts from the ground truth $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$; this is an auxiliary step mainly to simplify the theoretical analysis. While we can certainly make it practical by adopting spectral initialization as in [19, 71], it requires more lengthy proofs without further improving our statistical guarantees.

Algorithm 1 Construction of an approximate primal solution.

Initialization: $\boldsymbol{X}^0 = \boldsymbol{X}^\star$; $\boldsymbol{Y}^0 = \boldsymbol{Y}^\star$.

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

(27a)
$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta\nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) = \boldsymbol{X}^t - \frac{\eta}{p}\Big(\mathcal{P}_\Omega\left(\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}\right)\boldsymbol{Y}^t + \lambda\boldsymbol{X}^t\Big);$$

(27b)
$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \eta\nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) = \boldsymbol{Y}^t - \frac{\eta}{p}\Big(\left[\mathcal{P}_\Omega\left(\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}\right)\right]^\top\boldsymbol{X}^t + \lambda\boldsymbol{Y}^t\Big).$$

Here, $\eta > 0$ is the step size.

2.5. Properties of the nonconvex iterates. In this subsection, we will build upon the literature on nonconvex low-rank matrix completion to justify that the estimates returned by Algorithm 1 satisfy the requirement stated in (26). Our theory will be largely established upon the leave-one-out strategy introduced by Ma et al. [71], which is an effective analysis technique to control the $\ell_{2,\infty}$ error of the estimates. This strategy has recently been extended by Chen, Liu, and Li [19] to the more general rectangular case with an improved sample complexity bound.

Before continuing, we introduce several useful notations. Notice that the matrix product of $\boldsymbol{X}^\star$ and $\boldsymbol{Y}^{\star\top}$ is invariant under global orthonormal transformation, namely, for any orthonormal matrix $\boldsymbol{R} \in \mathbb{R}^{r\times r}$ one has $\boldsymbol{X}^\star\boldsymbol{R}(\boldsymbol{Y}^\star\boldsymbol{R})^\top = \boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}$. Viewed in this light, we shall consider distance metrics modulo global rotation. In particular, the theory relies heavily on a specific global rotation matrix defined as follows

$$\boldsymbol{H}^t \triangleq \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left(\left\|\boldsymbol{X}^t\boldsymbol{R} - \boldsymbol{X}^\star\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Y}^t\boldsymbol{R} - \boldsymbol{Y}^\star\right\|_{\mathrm{F}}^2 \right)^{1/2}, \tag{28}$$

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

We are now ready to present the performance guarantees for Algorithm 1.

LEMMA 2.7 (quality of the nonconvex estimates). *Instate the notation and hypotheses of Theorem* 1.4. *With probability at least* $1 - O\left(n^{-3}\right)$, *the iterates* $\{(\boldsymbol{X}^t, \boldsymbol{Y}^t)\}_{0\le t\le t_0}$ *of Algorithm* 1 *satisfy*

$$\max\left\{\left\|\boldsymbol{X}^t\boldsymbol{H}^t - \boldsymbol{X}^\star\right\|_{\mathrm{F}}, \left\|\boldsymbol{Y}^t\boldsymbol{H}^t - \boldsymbol{Y}^\star\right\|_{\mathrm{F}}\right\} \le C_{\mathrm{F}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}, \tag{29a}$$

$$\max\left\{\left\|\boldsymbol{X}^t\boldsymbol{H}^t - \boldsymbol{X}^\star\right\|, \left\|\boldsymbol{Y}^t\boldsymbol{H}^t - \boldsymbol{Y}^\star\right\|\right\} \le C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\left\|\boldsymbol{X}^\star\right\|, \tag{29b}$$

$$\begin{aligned}&\max\left\{\left\|\boldsymbol{X}^t\boldsymbol{H}^t - \boldsymbol{X}^\star\right\|_{2,\infty}, \left\|\boldsymbol{Y}^t\boldsymbol{H}^t - \boldsymbol{Y}^\star\right\|_{2,\infty}\right\}\\ &\qquad \le C_\infty\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\max\left\{\left\|\boldsymbol{X}^\star\right\|_{2,\infty}, \left\|\boldsymbol{Y}^\star\right\|_{2,\infty}\right\},\end{aligned} \tag{29c}$$

$$\min_{0\le t<t_0}\left\|\nabla f\left(\boldsymbol{X}^t, \boldsymbol{Y}^t\right)\right\|_{\mathrm{F}} \le \frac{1}{n^5}\frac{\lambda}{p}\sqrt{\sigma_{\min}}, \tag{30}$$

where $C_{\mathrm{F}}, C_{\mathrm{op}}, C_\infty > 0$ *are some absolute constants, provided that* $\eta \asymp 1/(n\kappa^3\sigma_{\max})$ *and that* $t_0 = n^{18}$.

This lemma, which we establish in https://arxiv.org/pdf/1902.07698.pdf, reveals that for a polynomially large number of iterations, all iterates of the GD sequence—when initialized at the ground truth—remain fairly close to the true low-rank factors. This holds in terms of the estimation errors measured by the Frobenius norm, the spectral norm, and the $\ell_{2,\infty}$ norm. In particular, the proximity in terms of the $\ell_{2,\infty}$ norm error plays a pivotal role in implementing our analysis strategy (particularly Lemmas 2.3–2.6) described previously. In addition, this lemma (cf. (30)) guarantees the existence of a small-gradient point within this sequence $\{(\boldsymbol{X}^t, \boldsymbol{Y}^t)\}_{0\le t\le t_0}$, a somewhat straightforward property of GD tailored to smooth problems [75]. This in turn enables us to invoke Lemma 2.3.

As immediate consequences of Lemma 2.7, with high probability we have

$$\left\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top}-\boldsymbol{M}^\star\right\|_{\mathrm{F}} \le 3\kappa C_{\mathrm{F}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\|_{\mathrm{F}}, \tag{31a}$$

$$\left\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top}-\boldsymbol{M}^\star\right\|_{\infty} \le 3C_{\infty}\sqrt{\kappa^3\mu r}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\|_{\infty}, \tag{31b}$$

$$\left\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top}-\boldsymbol{M}^\star\right\| \le 3C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\| \tag{31c}$$

for all $0\le t\le t_0$. The proof is deferred to https://arxiv.org/pdf/1902.07698.pdf.

2.6. Proof of Theorem 1.4. Let $t_* \triangleq \arg\min_{0\le t<t_0}\|\nabla f(\boldsymbol{X}^t,\boldsymbol{Y}^t)\|_{\mathrm{F}}$, and take $(\boldsymbol{X}_{\mathsf{ncvx}},\boldsymbol{Y}_{\mathsf{ncvx}}) = (\boldsymbol{X}^{t_*}\boldsymbol{H}^{t_*},\boldsymbol{Y}^{t_*}\boldsymbol{H}^{t_*})$ (cf. (28)). It is straightforward to verify that $(\boldsymbol{X}_{\mathsf{ncvx}},\boldsymbol{Y}_{\mathsf{ncvx}})$ obeys (i) the small-gradient condition (23), and (ii) the proximity condition (26). We are now positioned to invoke Lemma 2.3: for any optimizer $\boldsymbol{Z}_{\mathsf{cvx}}$ of (3), one has

$$\begin{aligned}\left\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top\right\|_{\mathrm{F}} &\lesssim \frac{\kappa}{c_{\mathrm{inj}}}\frac{1}{\sqrt{\sigma_{\min}}}\|\nabla f(\boldsymbol{X}_{\mathsf{ncvx}},\boldsymbol{Y}_{\mathsf{ncvx}})\|_{\mathrm{F}} \lesssim \frac{\kappa^2}{n^5}\frac{\lambda}{p}\\ &= \frac{\kappa}{n^5}\frac{\lambda}{p\,\sigma_{\min}}(\kappa\sigma_{\min}) = \frac{\kappa}{n^5}\frac{\lambda}{p\,\sigma_{\min}}\|\boldsymbol{M}^\star\|\\ &\lesssim \frac{1}{n^4}\frac{\lambda}{p\,\sigma_{\min}}\|\boldsymbol{M}^\star\|. \end{aligned}\tag{32}$$

The last line arises since $n\gg\kappa$—a consequence of the sample complexity condition $np\gtrsim\kappa^4\mu^2r^2\log^3 n$ (and hence $n\ge np\gtrsim\kappa^4\mu^2r^2\log^3 n\gg\kappa^4$). This taken collectively with the property (31) implies that

$$\begin{aligned}\left\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{M}^\star\right\|_{\mathrm{F}} &\le \left\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top\right\|_{\mathrm{F}} + \left\|\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top-\boldsymbol{M}^\star\right\|_{\mathrm{F}}\\ &\lesssim \frac{1}{n^4}\frac{\lambda}{p\,\sigma_{\min}}\|\boldsymbol{M}^\star\| + \kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\|_{\mathrm{F}}\\ &\asymp \kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\|_{\mathrm{F}}. \end{aligned}$$

In other words, since $\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top$ and $\boldsymbol{Z}_{\mathsf{ncvx}}$ are exceedingly close, the error $\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{M}^\star$ is mainly accredited to $\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top-\boldsymbol{M}^\star$. Similar arguments lead to

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{M}^\star\right\| \lesssim \left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\|,$$

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{M}^\star\right\|_{\infty} \lesssim \sqrt{\kappa^3\mu r}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\|_{\infty}.$$

We are left with proving the properties of $\boldsymbol{Z}_{\mathsf{cvx},r}$. Since $\boldsymbol{Z}_{\mathsf{cvx},r}$ is defined to be the best rank-r approximation of $\boldsymbol{Z}_{\mathsf{cvx}}$, one can invoke (32) to derive

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{Z}_{\mathsf{cvx},r}\right\|_{\mathrm{F}} \le \left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^{\top}\right\|_{\mathrm{F}} \lesssim \frac{1}{n^4}\frac{\lambda}{p\,\sigma_{\min}}\|\boldsymbol{M}^{\star}\|$$

from which (12) follows. Repeating the above calculations implies that (11) holds if $\boldsymbol{Z}_{\mathsf{cvx}}$ is replaced by $\boldsymbol{Z}_{\mathsf{cvx},r}$, thus concluding the proof.

3. Prior art. Nuclear norm minimization, pioneered by the seminal works [14, 15, 45, 80], has been a popular and principled approach to low-rank matrix recovery. In the noiseless setting, i.e., $\boldsymbol{E} = \boldsymbol{0}$, it amounts to solving the following constrained convex program

$$\text{minimize}_{\boldsymbol{Z}\in\mathbb{R}^{n\times n}} \|\boldsymbol{Z}\|_{*} \qquad \text{subject to} \quad \mathcal{P}_{\Omega}(\boldsymbol{Z}) = \mathcal{P}_{\Omega}(\boldsymbol{M}^{\star}), \tag{33}$$

which enjoys great theoretical success. Informally, this approach enables exact recovery of a rank-r matrix $\boldsymbol{M}^{\star} \in \mathbb{R}^{n\times n}$ as soon as the sample size is about the order of nr—the intrinsic degrees of freedom of a rank-r matrix [20, 51, 79]. In particular, Gross [51] blazed a trail by developing an ingenious golfing scheme for dual construction—an analysis technique that has found applications far beyond matrix completion. When it comes to the noisy case, Candès and Plan [13] first studied the stability of convex programming when the noise is bounded and possibly adversarial, followed by [74] and [64] using two modified convex programs. As we have already discussed, none of these papers provide optimal statistical guarantees under our model when $r = O(1)$. Other related papers such as [10, 62] include similar estimation error bounds and suffer from similar suboptimality issues.

Turning to nonconvex optimization, we note that this approach has recently received much attention for various low-rank matrix factorization problems, owing to its superior computational advantage compared to convex programming (e.g., [12, 22, 56, 60, 91, 97]). The convergence guarantees for matrix completion have been established for various algorithms such as GD on manifold [60, 61], alternating minimization [53, 56], GD [19, 71, 87, 94], and projected GD [31], provided that a suitable initialization (like spectral initialization) is available [23, 56, 60, 71, 87]. Our work is mostly related to [19, 71], which studied (vanilla) GD for nonconvex matrix completion. This algorithm was first analyzed by [71] via a leave-one-out argument—a technique that proves useful in analyzing various statistical algorithms [1, 26, 27, 35, 38, 67, 88, 101]. In the absence of noise and omitting logarithmic factors, [71] showed that $O(nr^3)$ samples are sufficient for vanilla GD to yield ε accuracy in $O(\log\frac{1}{\varepsilon})$ iterations (without the need of extra regularization procedures); the sample complexity was further improved to $O(nr^2)$ by [19]. Apart from GD, other nonconvex methods (e.g., [16, 35, 48, 52, 53, 55, 56, 57, 65, 81, 82, 92, 95, 96, 99]) and landscape / geometry properties have been investigated [18, 49, 50, 77, 83]; these are, however, beyond the scope of the current paper.

Another line of works asserted that a large family of semidefinite programs (SDPs) admits low-rank solutions [4], which in turn motivates the Burer–Monteiro approach [6, 7]. When applied to matrix completion, however, the generic theoretical guarantees therein lead to conservative results. Take the noiseless case (33) for instance: these results revealed the existence of a solution of rank at most $O(\sqrt{n^2p})$, which, however, is often much larger than the true rank (e.g., when $r \asymp 1$ and $p \asymp \mathrm{poly}\log(n)/n$, one has $\sqrt{n^2p} \gg \sqrt{n} \gg r$). Moreover, this line of works does not imply that all solutions to the SDP of interest are (approximately) low rank.

Finally, the connection between convex and nonconvex optimization has also been explored in line spectral estimation [66], although the context therein is drastically different from ours.

4. Discussion. This paper provides an improved statistical analysis for the natural convex program (3), without the need of enforcing additional spikiness constraint. Our theoretical analysis uncovers an intriguing connection between convex relaxation and nonconvex optimization, which we believe is applicable to many other problems beyond matrix completion. Having said that, our current theory leaves open a variety of important directions for future exploration. Here we sample a few interesting ones.

- *Improving dependency on r and κ.* While our theory is optimal when r and κ are both constants, it becomes increasingly looser as either r or κ grows. For instance, in the noiseless setting, it has been shown that the sample complexity for convex relaxation scales as $O(nr)$—linear in r and independent of κ—which is better than the current results. It is worth noting that existing theory for nonconvex matrix factorization typically falls short of providing optimal scaling in r and κ [19, 31, 60, 71, 87]. Thus, tightening the dependency of sample complexity on r and κ might call for new analysis tools.
- *Approximate low rank structure.* So far our theory is built upon the assumption that the ground-truth matrix $\boldsymbol{M}^\star$ is exactly low-rank, which falls short of accommodating the more realistic scenario where $\boldsymbol{M}^\star$ is only approximately low rank. For the approximate low-rank case, it is not yet clear whether the nonconvex factorization approach can still serve as a tight proxy. In addition, the landscape of nonconvex optimization for the approximately low-rank case [18] might shed light on how to handle this case.
- *Extension to deterministic noise.* Our current theory—in particular, the leave-one-out analysis for the nonconvex approach—relies heavily on the randomness assumption (i.e., i.i.d. sub-Gaussian) of the noise. In order to justify the broad applicability of convex relaxation, it would be interesting to see whether one can generalize the theory to cover deterministic noise with bounded magnitudes.
- *Extension to structured matrix completion.* Many applications involve low-rank matrices that exhibit additional structures, enabling a further reduction of the sample complexity [9, 24, 47]. For instance, if a matrix is Hankel and low rank, then the sample complexity can be $O(n)$ times smaller than the generic low-rank case. The existing stability guarantee of Hankel matrix completion, however, is overly pessimistic compared to practical performance [24]. The analysis framework herein might be amenable to the study of Hankel matrix completion and help close the theory-practice gap.
- *Extension to robust principal component analysis and blind deconvolution.* Moving beyond matrix completion, there are other problems that are concerned with recovering low-rank matrices. Notable examples include robust principal component analysis [11, 17, 30], blind deconvolution [2, 68], and blind demixing [59, 69]. The stability analyses of the convex relaxation approaches for these problems [2, 69, 102] often adopt a similar approach as [13], and consequently are sub-optimal. The insights from the present paper might promise tighter statistical guarantees for such problems.

Finally, we remark that the intimate link between convex and nonconvex optimization enables statistically optimal inference and uncertainty quantification for noisy matrix completion (e.g., construction of optimal confidence intervals for each

missing entry). The interested readers are referred to our companion paper [28] for in-depth discussions.

Acknowledgments. Y. Chen thanks Emmanuel Candès for motivating discussions about noisy matrix completion.

REFERENCES

[1] E. Abbe, J. Fan, K. Wang, and Y. Zhong, *Entrywise eigenvector analysis of random matrices with low expected rank*, Ann. Statist., 48 (2020), pp. 1452–1474.
[2] A. Ahmed, B. Recht, and J. Romberg, *Blind deconvolution using convex programming*, IEEE Trans. Inform. Theory, 60 (2014), pp. 1711–1732.
[3] J. Bai and S. Ng, *Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions*, Econometrica, 74 (2006), pp. 1133–1150.
[4] A. I. Barvinok, *Problems of distance geometry and convex properties of quadratic maps*, Discrete Comput. Geom., 13 (1995), pp. 189–202.
[5] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.
[6] N. Boumal, V. Voroninski, and A. Bandeira, *The non-convex Burer-Monteiro approach works on smooth semidefinite programs*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2016, pp. 2757–2765.
[7] S. Burer and R. D. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Math. Program., 95 (2003), pp. 329–357.
[8] J.-F. Cai, E. J. Candès, and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM J. Optim., 20 (2010), pp. 1956–1982.
[9] J.-F. Cai, T. Wang, and K. Wei, *Fast and provable algorithms for spectrally sparse signal reconstruction via low-rank Hankel matrix completion*, Appl. Comput. Harmon. Anal., 46 (2019), pp. 94–121.
[10] T. T. Cai and W.-X. Zhou, *Matrix completion via max-norm constrained optimization*, Electron. J. Stat., 10 (2016), pp. 1493–1525.
[11] E. Candès, X. Li, Y. Ma, and J. Wright, *Robust principal component analysis?*, J. ACM, 58 (2011), 11.
[12] E. Candès, X. Li, and M. Soltanolkotabi, *Phase retrieval via Wirtinger flow: Theory and algorithms*, IEEE Trans. Inform. Theory, 61 (2015), pp. 1985–2007.
[13] E. Candès and Y. Plan, *Matrix completion with noise*, Proc. IEEE, 98 (2010), pp. 925 –936.
[14] E. Candès and B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math., 9 (2009), pp. 717–772.
[15] E. Candès and T. Tao, *The power of convex relaxation: Near-optimal matrix completion*, IEEE Trans. Inform. Theory, 56 (2010), pp. 2053 –2080.
[16] Y. Cao and Y. Xie, *Poisson matrix recovery and completion*, IEEE Trans. Signal Process., 64 (2016), pp. 1609–1620.
[17] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, *Rank-sparsity incoherence for matrix decomposition*, SIAM J. Optim., 21 (2011), pp. 572–596.
[18] J. Chen and X. Li, *Model-free nonconvex matrix completion: Local minima analysis and applications in memory-efficient Kernel PCA*, J. Mach. Learn. Res., 20 (2010), 39.
[19] J. Chen, D. Liu, and X. Li, *Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization*, IEEE Trans. Inform. Theory, 66 (2020), pp. 5806–5841.
[20] Y. Chen, *Incoherence-optimal matrix completion*, IEEE Trans. Inform. Theory, 61 (2015), pp. 2909–2923.
[21] Y. Chen and E. Candès, *The projected power method: An efficient algorithm for joint alignment from pairwise differences*, Comm. Pure Appl. Math., 71 (2018), pp. 1648–1714.
[22] Y. Chen and E. J. Candès, *Solving random quadratic systems of equations is nearly as easy as solving linear systems*, Comm. Pure Appl. Math., 70 (2017), pp. 822–883, https://doi.org/10.1002/cpa.21638.
[23] Y. Chen, C. Cheng, and J. Fan, *Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices*, Ann. Statist., to appear.
[24] Y. Chen and Y. Chi, *Robust spectral compressed sensing via structured matrix completion*, IEEE Trans. Inform. Theory, 60 (2014), pp. 6576–6601.
[25] Y. Chen and Y. Chi, *Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization*,

IEEE Signal Process. Mag., 35 (2018), pp. 14–31, https://doi.org/10.1109/MSP.2018.2821706.

[26] Y. Chen, Y. Chi, J. Fan, and C. Ma, *Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval*, Math. Program., 176 (2019), pp. 5–37.

[27] Y. Chen, J. Fan, C. Ma, and K. Wang, *Spectral method and regularized MLE are both optimal for top-K ranking*, Ann. Statist., 47 (2019), pp. 2204–2235.

[28] Y. Chen, J. Fan, C. Ma, and Y. Yan, *Inference and uncertainty quantification for noisy matrix completion*, Proc. Natl. Acad. Sci. USA, 116 (2019), pp. 22931–22937.

[29] Y. Chen, L. J. Guibas, and Q. Huang, *Near-optimal joint optimal matching via convex relaxation*, in International Conference on Machine Learning (ICML), ACM, New York, 2014, pp. 100–108.

[30] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis, *Low-rank matrix recovery from errors and erasures*, IEEE Trans. Inform. Theory, 59 (2013), pp. 4324–4337.

[31] Y. Chen and M. J. Wainwright, *Fast Low-Rank Estimation by Projected Gradient Descent: General Statistical and Algorithmic Guarantees*, preprint, https://arxiv.org/abs/1509.03025, 2015.

[32] Y. Cheng and R. Ge, *Non-convex matrix completion against a semi-random adversary*, Proc. Mach. Learn. Res., 75 (2018), pp. 1362–1394.

[33] Y. Chi, Y. M. Lu, and Y. Chen, *Nonconvex optimization meets low-rank matrix factorization: An overview*, IEEE Trans. Signal Process., 67 (2019), pp. 5239–5269.

[34] M. A. Davenport and J. Romberg, *An overview of low-rank matrix recovery from incomplete observations*, IEEE J. Sel. Topics Signal Process., 10 (2016), pp. 608–622.

[35] L. Ding and Y. Chen, *The Leave-One-Out Approach for Matrix Completion: Primal and Dual Analysis*, preprint, https://arxiv.org/abs/1803.07554, 2018.

[36] B. Efron, *Correlation and large-scale simultaneous significance testing*, J. Amer. Statist. Assoc., 102 (2007), pp. 93–103.

[37] B. Efron, *Correlated z-values and the accuracy of large-scale statistical estimates*, J. Amer. Statist. Assoc., 105 (2010), pp. 1042–1055.

[38] N. El Karoui, *On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators*, Probab. Theory Related Fields, 170 (2018), pp. 95–175.

[39] J. Fan, X. Han, and W. Gu, *Estimating false discovery proportion under arbitrary covariance dependence*, J. Amer. Statist. Assoc., 107 (2012), pp. 1019–1035.

[40] J. Fan, Y. Ke, Q. Sun, and W.-X. Zhou, *Farmtest: Factor-adjusted robust multiple testing with approximate false discovery control*, J. Amer. Statist. Assoc., 114 (2019), pp. 1880–1893.

[41] J. Fan, Y. Ke, and K. Wang, *Factor-adjusted regularized model selection*, J. Econometrics, 216 (2020), pp. 71–85.

[42] J. Fan, Y. Liao, and M. Mincheva, *Large covariance estimation by thresholding principal orthogonal complements*, J. R. Stat. Soc. Ser. B Stat. Methodol., 75 (2013), pp. 603–680.

[43] J. Fan, W. Wang, and Y. Zhong, *Robust covariance estimation for approximate factor models*, J. Econometrics, 208 (2019), pp. 5–22.

[44] J. Fan, L. Xue, and J. Yao, *Sufficient forecasting using factor models*, J. Econometrics, 201 (2017), pp. 292–306.

[45] M. Fazel, *Matrix Rank Minimization with Applications*, PhD thesis, Stanford University, Stanford, CA, 2002.

[46] M. Fazel, H. Hindi, and S. Boyd, *Rank minimization and applications in system theory*, in American Control Conference, Vol. 4, IEEE, Piscataway, NJ, 2004, pp. 3273–3278.

[47] M. Fazel, H. Hindi, and S. P. Boyd, *Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices*, in American Control Conference, IEEE, Piscataway, NJ, 2003, pp. 2156–2162.

[48] M. Fornasier, H. Rauhut, and R. Ward, *Low-rank matrix recovery via iteratively reweighted least squares minimization*, SIAM J. Optim., 21 (2011), pp. 1614–1640.

[49] R. Ge, C. Jin, and Y. Zheng, *No spurious local minima in nonconvex low rank problems: A unified geometric analysis*, International Conference on Machine Learning, ACM, New York, 2017, pp. 1233–1242.

[50] R. Ge, J. D. Lee, and T. Ma, *Matrix completion has no spurious local minimum*, in Advances in Neural Information Processing Systems, 2016, Curran Associates, Red Hook, NY, pp. 2973–2981.

[51] D. Gross, *Recovering low-rank matrices from few coefficients in any basis*, IEEE Trans. Inform. Theory, 57 (2011), pp. 1548–1566.

[52] S. Gunasekar, A. Acharya, N. Gaur, and J. Ghosh, *Noisy matrix completion using alternating minimization*, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Berlin, 2013, pp. 194–209.
[53] M. Hardt, *Understanding alternating minimization for matrix completion*, in Foundations of Computer Science (FOCS), IEEE, Piscataway, NJ, 2014, pp. 651–660.
[54] Q.-X. Huang and L. Guibas, *Consistent shape maps via semidefinite programming*, Comput. Graph. Forum, 32 (2013), pp. 177–186.
[55] P. Jain, R. Meka, and I. S. Dhillon, *Guaranteed rank minimization via singular value projection*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2010, pp. 937–945.
[56] P. Jain, P. Netrapalli, and S. Sanghavi, *Low-rank matrix completion using alternating minimization*, in ACM Symposium on Theory of Computing, ACM, New York, 2013, pp. 665–674.
[57] C. Jin, S. M. Kakade, and P. Netrapalli, *Provable efficient online matrix completion via non-convex stochastic gradient descent*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2016, pp. 4520–4528.
[58] I. T. Jolliffe, *A note on the use of principal components in regression*, J. R. Stat. Soc. Ser. C Appl. Stat., 31 (1982), pp. 300–303.
[59] P. Jung, F. Krahmer, and D. Stöger, *Blind demixing and deconvolution at near-optimal rate*, IEEE Trans. Inform. Theory, 64 (2017), pp. 704–727.
[60] R. H. Keshavan, A. Montanari, and S. Oh, *Matrix completion from a few entries*, IEEE Trans. Inform. Theory, 56 (2010), 2980 –2998.
[61] R. H. Keshavan, A. Montanari, and S. Oh, *Matrix completion from noisy entries*, J. Mach. Learn. Res., 11 (2010), pp. 2057–2078.
[62] O. Klopp, *Noisy low-rank matrix completion with general sampling distribution*, Bernoulli, 20 (2014), pp. 282–303.
[63] A. Kneip and P. Sarda, *Factor models and variable selection in high-dimensional regression analysis*, Ann. Statist., 39 (2011), pp. 2410–2447.
[64] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, *Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion*, Ann. Statist., 39 (2011), pp. 2302–2329, https://doi.org/10.1214/11-AOS894.
[65] M.-J. Lai, Y. Xu, and W. Yin, *Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization*, SIAM J. Numer. Anal., 51 (2013), pp. 927–957.
[66] Q. Li and G. Tang, *Approximate support recovery of atomic line spectral estimation: A tale of resolution and precision*, Appl. Comput. Harmon. Anal., 48 (2020), pp. 891–948.
[67] Y. Li, C. Ma, Y. Chen, and Y. Chi, *Nonconvex matrix factorization from rank-one measurements*, Proc. Mach. Learn. Res., 89 (2019), pp. 1496–1505.
[68] S. Ling and T. Strohmer, *Self-calibration and biconvex compressive sensing*, Inverse Problems, 31 (2015), 115002.
[69] S. Ling and T. Strohmer, *Blind deconvolution meets blind demixing: Algorithms and performance bounds*, IEEE Trans. Inform. Theory, 63 (2017), pp. 4497–4520.
[70] Z. Liu and L. Vandenberghe, *Interior-point method for nuclear norm approximation with application to system identification*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 1235–1256.
[71] C. Ma, K. Wang, Y. Chi, and Y. Chen, *Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution*, Found. Comput. Math., 20 (2020), pp. 451–632.
[72] S. Ma, D. Goldfarb, and L. Chen, *Fixed point and Bregman iterative methods for matrix rank minimization*, Math. Program., 128 (2011), pp. 321–353.
[73] R. Mazumder, T. Hastie, and R. Tibshirani, *Spectral regularization algorithms for learning large incomplete matrices*, J. Mach. Learn. Res., 11 (2010), pp. 2287–2322.
[74] S. Negahban and M. J. Wainwright, *Restricted strong convexity and weighted matrix completion: Optimal bounds with noise*, J. Mach. Learn. Res., 13 (2012), pp. 1665–1697.
[75] Y. Nesterov, *How to make the gradients small*, Optima, 88 (2012), pp. 10–11.
[76] N. Parikh and S. Boyd, *Proximal algorithms*, Found. Trends Optim., 1 (2014), pp. 127–239.
[77] D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi, *Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach*, Proc. Mach. Learn. Res., 54 (2017), pp. 65–74.
[78] D. Paul, E. Bair, T. Hastie, and R. Tibshirani, *"Preconditioning" for feature selection and regression in high-dimensional problems*, Ann. Statist., 36 (2008), pp. 1595–1618.
[79] B. Recht, *A simpler approach to matrix completion*, J. Mach. Learn. Res., 12 (2011), pp. 3413–3430.

[80] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Rev., 52 (2010), pp. 471–501.
[81] J. D. Rennie and N. Srebro, *Fast maximum margin matrix factorization for collaborative prediction*, in International conference on Machine Learning, ACM, New York, 2005, pp. 713–719.
[82] A. Rohde and A. B. Tsybakov, *Estimation of high-dimensional low-rank matrices*, Ann. Statist., 39 (2011), pp. 887–930.
[83] A. Shapiro, Y. Xie, and R. Zhang, *Matrix completion with deterministic pattern: A geometric perspective*, IEEE Trans. Signal Process., 67 (2019), pp. 1088–1103.
[84] A. Singer, *Angular synchronization by eigenvectors and semidefinite programming*, Appl. Comput. Harmon. Anal., 30 (2011), pp. 20–36.
[85] A. M.-C. So and Y. Ye, *Theory of semidefinite programming for sensor network localization*, Math. Program., 109 (2007), pp. 367–384.
[86] N. Srebro and A. Shraibman, *Rank, trace-norm and max-norm*, in International Conference on Computational Learning Theory, Springer, Berlin, 2005, pp. 545–560.
[87] R. Sun and Z.-Q. Luo, *Guaranteed matrix completion via non-convex factorization*, IEEE Trans. Inform. Theory, 62 (2016), pp. 6535–6579.
[88] P. Sur, Y. Chen, and E. J. Candès, *The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square*, Probab. Theory Related Fields, 175 (2019), pp. 487–558.
[89] K.-C. Toh and S. Yun, *An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems*, Pac. J. Optim., 6 (2010), pp. 615–640.
[90] C. Tomasi and T. Kanade, *Shape and motion from image streams under orthography: A factorization method*, Int. J. Comput. Vis., 9 (1992), pp. 137–154.
[91] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, *Low-rank solutions of linear matrix equations via Procrustes flow*, in International Conference on Machine Learning, ACM, New York, 2016, pp. 964–973.
[92] B. Vandereycken, *Low-rank matrix completion by Riemannian optimization*, SIAM J. Optim., 23 (2013), pp. 1214–1236.
[93] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, in Compressed Sensing, Theory and Applications, Cambridge University Press, Cambridge, 2012, pp. 210–268.
[94] L. Wang, X. Zhang, and Q. Gu, *A unified computational and statistical framework for nonconvex low-rank matrix estimation*, Proc. Mach. Learn. Res., 54 (2017), pp. 981–990.
[95] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung, *Guarantees of Riemannian optimization for low rank matrix recovery*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 1198–1222.
[96] Z. Wen, W. Yin, and Y. Zhang, *Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm*, Math. Program. Comput., 4 (2012), pp. 333–361.
[97] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, *A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms*, J. Mach. Learn. Res., 18 (2017), pp. 5164–5198.
[98] T. Zhang, J. M. Pauly, and I. R. Levesque, *Accelerating parameter mapping with a locally low rank constraint*, Magn. Resonance Med., 73 (2015), pp. 655–661.
[99] T. Zhao, Z. Wang, and H. Liu, *A nonconvex optimization framework for low rank matrix estimation*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2015, pp. 559–567.
[100] Q. Zheng and J. Lafferty, *Convergence Analysis for Rectangular Matrix Completion using Burer-Monteiro Factorization and Gradient Descent*, preprint, https://arxiv.org/abs/1605.07051, 2016.
[101] Y. Zhong and N. Boumal, *Near-optimal bounds for phase synchronization*, SIAM J. Optim., 28 (2018), pp. 989–1016.
[102] Z. Zhou, X. Li, J. Wright, E. Candès, and Y. Ma, *Stable principal component pursuit*, in International Symposium on Information Theory, IEEE, Piscataway, NJ, 2010, pp. 1518–1522.