How Many Samples is a Good Initial Point Worth in Low-rank Matrix Recovery?

Gavin Zhang
Department of Electrical and Computer Engineering
University of Illinois Urbana Champaign
Illinois, IL61820
jialun2@illinois.edu

Richard Y. Zhang
Department of Electrical and Computer Engineering
University of Illinois Urbana Champaign
Illinois, IL61820
ryz@illinois.edu

Abstract

Given a sufficiently large amount of labeled data, the non-convex low-rank matrix recovery problem contains no spurious local minima, so a local optimization algorithm is guaranteed to converge to a global minimum starting from any initial guess. However, the actual amount of data needed by this theoretical guarantee is very pessimistic, as it must prevent spurious local minima from existing anywhere, including at adversarial locations. In contrast, prior work based on good initial guesses have more realistic data requirements, because they allow spurious local minima to exist outside of a neighborhood of the solution. In this paper, we quantify the relationship between the quality of the initial guess and the corresponding reduction in data requirements. Using the restricted isometry constant as a surrogate for sample complexity, we compute a sharp "threshold" number of samples needed to prevent each specific point on the optimization landscape from becoming a spurious local minimum. Optimizing the threshold over regions of the landscape, we see that for initial points around the ground truth, a linear improvement in the quality of the initial guess amounts to a constant factor improvement in the sample complexity.

1 Introduction

A perennial challenge in non-convex optimization is the possible existence of *bad* or *spurious* critical points and local minima, which can cause a local optimization algorithm like gradient descent to slow down or get stuck. Several recent lines of work showed that the effects of non-convexity can be tamed through a large amount of diverse and high quality training data [17, 1, 9, 3, 18, 12]. Concretely, these authors showed that, for classes of problems based on random sampling, spurious critical points and local minima become progressively less likely to exist with the addition of each new sample. After a *sufficiently large number of samples*, all spurious local minima are eliminated, so any local optimization algorithm is guaranteed to converge to the globally optimal solution starting from an arbitrary, possibly random initial guess.

This notion of a *global* guarantee—one that is valid starting from any initial point—is considerably stronger than what is needed for empirical success to be observed [8]. For example, the existence of a spurious local minimum may not pose an issue if gradient descent does not converge towards it.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

However, a theoretical guarantee is no longer possible, as starting the algorithm from the spurious local minimum would result in failure [22]. As a consequence, these global guarantees tend to be pessimistic, because the number of samples must be sufficiently large to eliminate spurious local minima everywhere, even at adversarial locations. By contrast, the weaker notion of a *local* guarantee [11, 10, 15, 19, 5, 7, 20, 13]—one that is valid only for a specified set of initial points—is naturally less conservative, as it allows spurious local minima to exist outside of the specified set.

In this paper, we provide a unifying view between the notions of the global and local guarantees by quantifying the relationship between the sample complexity and the quality of the initial point. We restrict our attention to the *matrix sensing* problem, which seeks to recover a rank-r positive semidefinite matrix $M^* = ZZ^T \in \mathbb{R}^{n\times n}$ with $Z \in \mathbb{R}^{n\times r}$ from m sub-Gaussian linear measurements of the form

$$b \equiv \mathcal{A}(ZZ^T) \equiv [\langle A_1, M^*\rangle \quad \cdots \quad \langle A_m, M^*\rangle]^T \tag{1}$$

by solving the following non-convex optimization problem:

$$\min_{X\in\mathbb{R}^{n\times r}} f_{\mathcal{A}}(X) \equiv \left\|\mathcal{A}\left(XX^T - ZZ^T\right)\right\|^2 = \sum_{i=1}^{m}\left(\left\langle A_i, XX^T\right\rangle - b_i\right)^2. \tag{2}$$

We characterize a sharp "threshold" on the number of samples m needed to prevent each specific point on the optimization landscape from becoming a spurious local minimum. While the threshold is difficult to solve, we derive a lower-bound in closed-form based on spurious *critical points*, and show that it constitutes a *sharp* lower-bound on the original threshold of interest. The lower-bound reveals a simple geometric relationship: a point X is more likely to be a local minimum if the column spaces of X and Z are close to orthogonal. Optimizing the closed-form lower-bound over regions of the landscape, we show that for initial points close to the ground truth, a constant factor improvement of the initial point amounts to a constant factor reduction in the number of samples needed to guarantee recovery.

2 Related Work

Local Guarantees. The earliest work on exact guarantees for non-convex optimization focused on generating a good initial guess within a local region of attraction. For instance, in [21, 24], the authors showed that when $\mathcal{A}$ satisfies $(\delta, 6r)$-RIP with a constant $\delta \leq 1/10$, and there exists a initial point sufficiently close to the ground truth, then gradient descent starting from this initial point has a linear convergence rate. The typical strategy to find such the initial point is *spectral initialization* [11, 10, 21, 19, 5, 14, 6]: using the singular value decomposition on a surrogate matrix to find low-rank factors that are close to the ground truth.

In this paper, we focus on the trade-off between the quality of an initial point and the number of samples needed to prevent the existence of spurious local minima, while sidestepping the question of how it is found. We note, however, that the number of samples needed to find an ϵ-good initial guess (e.g. via spectral initialization) forms an interesting secondary trade-off. It remains a future work to study the interactions between these two points.

Global Guarantees. Recent work focused on establishing a global guarantee that is independent of the initial guess [17, 1, 9, 3, 18, 12]. For our purposes, Bhojanapalli et al. [2] showed that RIP with $\delta_{2r} < 1/5$ eliminates all spurious local minima, while Zhang et al. [23] refined this to $\delta_{2r} < 1/2$ for the rank-1 case, and showed that this is both and necessary and sufficient. This paper is inspired by proof techniques in the latter paper; an important contribution of our paper is generalizing their rank-1 techniques to accommodate for matrices of arbitrary rank.

3 Our Approach: Threshold RIP Constant

Previous work that studied the global optimization landscape of problem (2) typically relied on the restricted isometry property (RIP) of $\mathcal{A}$. It is now well-known that if the measurement operator $\mathcal{A}$ satisfies the restricted isometry property with a sufficiently small constant $\delta < 1/5$ then problem (2) contains no spurious local minima; see Bhojanapalli et al. [2].

Definition 1 (δ-RIP)**.** Let $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ be a linear measurement operator. We say that $\mathcal{A}$ satisfies the *δ-restricted isometry property* (or simply δ-RIP) if satisfies the following inequality

$$(1-\delta)\|M\|_F^2 \leq \|\mathcal{A}(M)\|^2 \leq (1+\delta)\|M\|_F^2 \qquad \forall M \in \mathcal{M}_{2r}$$

where $\mathcal{M}_{2r} = \{X \in \mathbb{R}^{n\times n} : \mathrm{rank}(X) \leq 2r\}$ denotes the set of rank-$2r$ matrices. The RIP constant of $\mathcal{A}$ is the smallest value of δ such that the inequality above holds.

Let $\delta \in [0, 1)$ denote the RIP constant of $\mathcal{A}$. It is helpful to view δ as a surrogate for the number of measurements $m \geq 0$, with a large value of δ corresponding a smaller value of m and vice versa. For a wide range of sub-Gaussian measurement ensembles, if $m \geq C_0 nr/\delta^2$ where C_0 is an absolute constant, then $\mathcal{A}$ satisfies δ-RIP with high probability [4, 16].

Take $X \in \mathbb{R}^{n\times r}$ to be a *spurious* point such that $XX^T \neq ZZ^T$. Our approach in this paper is to define a *threshold* number of measurements that would be needed to prevent X from becoming a local minimum for problem (1). Viewing the RIP constant δ as a surrogate for the number of measurements m, we follow a construction of Zhang et al. [23], and instead define a threshold $\delta_{\mathrm{soc}}(X)$ on the RIP constant δ that would prevent X from becoming a local minimum for problem (1). Such a construction must necessarily take into account all choices of $\mathcal{A}$ satisfying δ-RIP, including those that adversarially target X, bending the optimization landscape into forming a region of convergence around the point. On the other hand, such adversarial choices of $\mathcal{A}$ must necessarily be defeated for a sufficiently small threshold on δ, as we already know that spurious local minima cannot exist for $\delta < 1/5$. The statement below makes this idea precise, and also extends it to a set of spurious points.

Definition 2 (Threshold for second-order condition)**.** Fix $Z \in \mathbb{R}^{n\times r}$. For $X \in \mathbb{R}^{n\times r}$, if $XX^T = ZZ^T$, then define $\delta_{\mathrm{soc}}(X) = 1$. Otherwise, if $XX^T \neq ZZ^T$, then define

$$\delta_{\mathrm{soc}}(X) \equiv \min_{\mathcal{A}} \{\delta : \nabla f_{\mathcal{A}}(X) = 0, \quad \nabla^2 f_{\mathcal{A}}(X) \succeq 0, \quad \mathcal{A} \text{ satisfies } \delta\text{-RIP}\} \tag{3}$$

where the minimum is taken over all linear measurements $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$. For $\mathcal{W} \subseteq \mathbb{R}^{n\times r}$, define $\delta_{\mathrm{soc}}(\mathcal{W}) = \inf_{X\in\mathcal{W}} \delta_{\mathrm{soc}}(X)$.

If $\delta < \delta_{\mathrm{soc}}(X)$, then X cannot be a spurious local minimum by construction, or it would contradict the definition of $\delta_{\mathrm{soc}}(X)$ as the minimum value. By the same logic, if $\delta < \delta_{\mathrm{soc}}(\mathcal{W})$, then no choice of $X \in \mathcal{W}$ can be a spurious local minimum. In particular, it follows that $\delta_{\mathrm{soc}}(\mathbb{R}^{n\times r})$ is the usual *global* RIP threshold: if $\mathcal{A}$ satisfies δ-RIP with $\delta < \delta_{\mathrm{soc}}(\mathbb{R}^{n\times r})$, then $f_{\mathcal{A}}(X)$ is guaranteed to admit no spurious local minima. Starting a local optimization algorithm from any initial point guarantees exact recovery of an X satisfying $XX^T = ZZ^T$.

Now, suppose we are given an initial point X_0. It is natural to measure the *quality* of X_0 by its relative error, as in $\varepsilon = \|XX^T - ZZ^T\|_F / \|ZZ^T\|_F$. If we define an ε-neighborhood of all points with the same relative error

$$\mathcal{B}_\varepsilon = \{X \in \mathbb{R}^{n\times r}, \|XX^T - ZZ^T\|_F \leq \varepsilon \|ZZ^T\|_F\} \tag{4}$$

then it follows that $\delta_{\mathrm{soc}}(\mathcal{B}_\varepsilon)$ is an analogous *local* RIP threshold: if $\mathcal{A}$ satisfies δ-RIP with $\delta < \delta_{\mathrm{soc}}(\mathcal{B}_\varepsilon)$, then $f_{\mathcal{A}}(X)$ is guaranteed to admit no spurious local minima over all $X \in \mathcal{B}_\varepsilon$. Starting a local optimization algorithm from the initial point X_0 guarantees either exact recovery of an X satisfying $XX^T = ZZ^T$, or termination at a strictly worse point X with $\|XX^T - ZZ^T\|_F > \|X_0X_0^T - ZZ^T\|_F$. Imposing further restrictions on the algorithm prevents the latter scenario from occurring (local strong convexity with gradient descent [19], strict decrements in the levels set [10, 23, 8]), and so exact recovery is guaranteed.

The numerical difference between the global threshold $\delta_{\mathrm{soc}}(\mathbb{R}^{n\times r})$ and the local threshold $\delta_{\mathrm{soc}}(\mathcal{B}_\varepsilon)$ is precisely the number of samples that an ε-quality initial point X_0 is worth, up to some conversion factor. But two major difficulties remain in this line of reasoning. First, evaluating $\delta_{\mathrm{soc}}(X)$ for some $X \in \mathbb{R}^{n\times r}$ requires solving a minimization problem over the set of δ-RIP operators. Second, evaluating $\delta_{\mathrm{soc}}(\mathcal{B}_\varepsilon)$ in turn requires minimizing $\delta_{\mathrm{soc}}(X)$ over all choices of X within an ε-neighborhood. Regarding the first point, Zhang et al. [23] showed that $\delta_{\mathrm{soc}}(X)$ is the optimal value to a *convex* optimization problem, and can therefore be evaluated to arbitrary precising using a numerical algorithm. In the rank-1 case, they solved this convex optimization in closed-form, and use it to optimize over all $X \in \mathcal{B}_\varepsilon$. Their closed-form solution spanned 9 journal pages, and evoked a number of properties specific to the rank-1 case (for example, $xy^T + yx^T = 0$ implies $x = 0$ and $y = 0$, but $XY^T + YX^T = 0$ may hold for $X \neq 0$ and $Y \neq 0$). The authors noted that a similar closed-form solution for the general rank-r case appeared exceedingly difficult. While overall proof technique is sharp and descriptive, its applicability appears to be entirely limited to the rank-1 case.

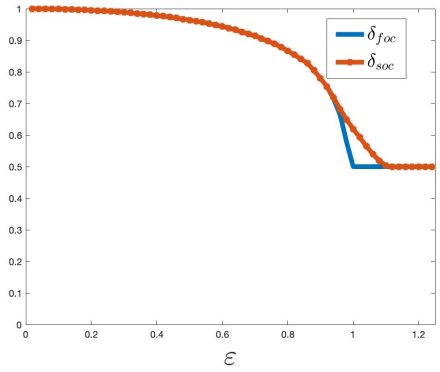


Figure 1: This paper is motivated by two key insights. First, it is relatively straightforward to solve $\delta_{\text{foc}}(X)$ in closed-form (Theorem 8). Second, the resulting lower-bound $\delta_{\text{soc}}(X) \geq \max\{\delta_{\text{foc}}(X), \delta^*\}$ ($\delta^* = 1/2$ for rank 1 and $\delta^* = 1/5$ for rank > 1) is remarkably tight. This means that $\max\{\delta_{\text{foc}}(\mathcal{B}_\varepsilon), \delta^*\}$ is a tight lower bound for $\delta_{\text{foc}}(\mathcal{B}_\varepsilon)$.

4 Main results

In this paper, we bypass the difficulty of deriving a closed-form solution for $\delta_{\text{soc}}(X)$ altogether by adopting a *sharp* lower-bound. This is based on two key insights. First, a spurious local minimum must also be a spurious critical point, so the analogous threshold over critical points would give an obvious lower-bound $\delta_{\text{foc}}(X) \leq \delta_{\text{soc}}(X)$.

Definition 3 (Threshold for first-order condition)**.** Fix $Z \in \mathbb{R}^{n\times r}$. For $X \in \mathbb{R}^{n\times r}$, if $XX^T = ZZ^T$, then define $\delta_{\text{foc}}(X) = 1$. Otherwise, if $XX^T \neq ZZ^T$, then define

$$\delta_{\text{foc}}(X) \equiv \min_{\mathcal{A}} \{\delta : \nabla f_{\mathcal{A}}(X) = 0, \quad \mathcal{A} \text{ satisfies } \delta\text{-RIP}\}, \tag{5}$$

where the minimum is taken over all linear measurements $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$. For $\mathcal{W} \subseteq \mathbb{R}^{n\times r}$, define $\delta_{\text{foc}}(\mathcal{W}) = \inf_{X\in\mathcal{W}} \delta_{\text{foc}}(X)$.

Whereas the main obstacle in Zhang et al. [23] is the considerable difficulty in deriving a closed-form solution for $\delta_{\text{soc}}(X)$, we show in this paper that it is relatively straightforward to solve $\delta_{\text{foc}}(X)$ in closed-form, to result in a simple, geometric solution.

Theorem 4. *Fix $Z \in \mathbb{R}^{n\times r}$. Given $\mathcal{A}$ satisfying δ-RIP and $X \in \mathbb{R}^{n\times r}$ such that $XX^T \neq ZZ^T$, we have $\delta_{\text{foc}}(X) = \cos\theta$, where*

$$\sin\theta = \|Z^T(I - XX^\dagger)Z\|_F \;/\; \|XX^T - ZZ^T\|_F. \tag{6}$$

and $X^\dagger$ denotes the pseudo-inverse of X. It follows that if $\delta < \cos\theta$, then X is not a spurious critical point of $f_{\mathcal{A}}(X)$. If $\delta \geq \cos\theta$, then there exists some $\mathcal{A}^\star$ satisfying $\cos\theta$-RIP such that $\nabla f_{\mathcal{A}}(X) = 0$.

The complete proof of Theorem 8 is given in Appendix A and a sketch is given in section 5. There is a nice geometric interpretation: the exact value of $\delta_{\text{foc}}(X)$ depends largely on the *incidence angle* between the column space of X and the column space of Z. When the angle between XX^T and ZZ^T becomes small, the projection of XX^T onto ZZ^T becomes large. As a result, $\sin\theta$ becomes small and $\cos\theta$ becomes large. Therefore, Theorem 8 says that in regions where XX^T and ZZ^T are more aligned, fewer samples are required to prevent X from becoming a spurious critical point. In regions where XX^T and ZZ^T are more orthogonal, a larger sample complexity is needed. Indeed, these are precisely the adversarial locations for which a large number of samples are required to prevent spurious local minima from appearing.

The lower-bound $\delta_{\text{foc}}(X) \leq \delta_{\text{soc}}(X)$ appears conservative, because critical points should be much more ubiquitous than local minima over a non-convex landscape. In particular, observe that $\delta_{\text{foc}}(X) = \cos\theta \to 0$ as $X \to 0$, which makes sense because $X = 0$ is a saddle point for all choices of $\mathcal{A}$. In other words, for any region $\mathcal{W}$ that contains 0, the lower-bound becomes trivial, as in $\delta_{\text{foc}}(\mathcal{W}) = 0 < \delta_{\text{soc}}(\mathcal{W})$. Our second insight here is that we must simultaneously have $\delta_{\text{soc}}(X) \geq 1/5$ due to

the global threshold of Bhojanapalli et al. [2] (or $\delta_{\text{soc}}(x) \geq 1/2$ in the rank-1 case due to Zhang et al. [23]). Extending this idea over sets yields the following lower-bound

$$\delta_{\text{soc}}(\mathcal{W}) \geq \max\{\delta_{\text{foc}}(\mathcal{W}), \delta^*\} \qquad \text{for all } \mathcal{W} \subseteq \mathbb{R}^{n\times r}, \tag{7}$$

where $\delta^* = 1/2$ for $r = 1$ and $\delta^* = 1/5 > 1$. This bound is *remarkably* tight, as shown in Figure 1 for $\mathcal{W} = \mathcal{B}_\varepsilon$ over a range of ε. Explicitly solving the optimization $\delta_{\text{foc}}(\mathcal{B}_\varepsilon) = \inf_{X\in\mathcal{B}_\varepsilon} \delta_{\text{foc}}(X)$ using Theorem 8 and substituting into (7) yields the following.[1]

Theorem 5. *Let $\mathcal{A}$ satisfy δ-RIP. Then we have $\delta_{\text{foc}}(\mathcal{B}_\varepsilon) > \sqrt{1 - C\varepsilon}$ for all $\epsilon \leq 1/C$, where $C = \|ZZ^T\|_F/\sigma^2_{\min}(Z)$. Hence, if*

$$\delta < \max\left\{\sqrt{[1 - C\varepsilon]_+}, \delta^*\right\} \tag{8}$$

where $\delta^* = 1/2$ if $r = 1$ and $\delta^* = 1/5$ if $r > 1$, then $f_{\mathcal{A}}(X)$ has no spurious critical point within an ε-neighborhood of the solution:

$$\nabla f_{\mathcal{A}}(X) = 0, \quad \|XX^T - ZZ^T\|_F \leq \varepsilon\|ZZ^T\|_F \quad \Longleftrightarrow \quad XX^T = ZZ^T. \tag{9}$$

The complete proof of this theorem is in Appendix B. Theorem 5 says that the number of samples needed to eliminate spurious critical points within an ε-neighborhood of the solution decreases dramatically as ε becomes small. Given that $m \geq C_0 nr/\delta^2$ sub-Gaussian measurements are needed to satisfy δ-RIP, we can translate Theorem 5 into the following sample complexity bound.

Corollary 6. *Let $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ be a sub-Gaussian measurement ensemble. If*

$$m \geq \min\left\{\frac{1}{[1 - C\varepsilon]_+}, 25\right\} C_0 nr$$

then with high probability there are no spurious local minima within $\mathcal{B}_\varepsilon$.

The proof of Corollary 6 follows immediately from Theorem 5 combined with the direct relationship between the RIP-property and the sample complexity for sub-Gaussian measurement ensembles. We see that the relationship between the quality of the initial point and the number of samples saved is essentially *linear*. Improving the quality of the initial point by a linear factor corresponds to a linear decrease in sample complexity. Moreover, the rate of improvement depends on the constant C. This shows that in the non-convex setup of matrix sensing, there is a significant difference between a good initial point and a mediocre initial point. In the case that $C = \|ZZ^T\|_F/\sigma^2_{\min}(Z)$ is large, this difference is even more pronounced.

5 Proof of Main Results

5.1 Notation and Definitions

We use $\|\cdot\|$ for the vector 2-norm and use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix. For two square matrices A and B, $A \succeq B$ means $B - A$ is positive semidefinite. The trace of a square matrix A is denoted by $\text{tr}(A)$. The *vectorization* $\text{vec}\,(A)$ is the length-mn vector obtained by stacking the columns of A. Let $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ be a linear measurement operator, and let $Z \in \mathbb{R}^{n\times r}$ be a fixed ground truth matrix. We define $\mathbf{A} = [\text{vec}\,(A_1), \ldots, \text{vec}\,(A_m)]$ as the matrix representation of $\mathcal{A}$, and note that $\text{vec}\,[\mathcal{A}(X)] = \mathbf{A}\,\text{vec}\,(X)$. We define the error vector $\mathbf{e}$ and its Jacobian $\mathbf{X}$ to satisfy

$$\mathbf{e} = \text{vec}\,(XX^T - ZZ^T) \tag{10a}$$

$$\mathbf{X}\,\text{vec}\,(Y) = \text{vec}\,(XY^T + YX^T) \qquad \text{for all } Y \in \mathbb{R}^{n\times r}. \tag{10b}$$

5.2 Proof Sketch of Theorem 4

A complete proof of Theorem 4 relies on a few technical lemmas, so we defer the complete proof to Appendix A. The key insight is that $\delta_{\text{foc}}(X)$ is the solution to a *convex* optimization problem, which we can solve in closed-form. At first sight, evaluating $\delta_{\text{foc}}(X)$ seems very difficult as it involves solving an optimization problem over the set of δ-RIP operators, as defined in equation 5 . However,

[1] We denote $[x]_+ = \max\{0, x\}$.

a minor modification of Theorem 8 in Zhang et al. [23] shows that $\delta_{\text{foc}}(X)$ can be reformulated as a convex optimization problem of the form

$$\eta(X) \quad \equiv \quad \max_{\eta,\mathbf{H}} \left\{ \eta \quad : \quad \mathbf{X}^T\mathbf{H}\mathbf{e} = 0, \quad \eta I \preceq \mathbf{H} \preceq I \right\}. \tag{11}$$

where $\eta(X)$ is related to $\delta_{\text{foc}}(X)$ by

$$\delta_{\text{foc}}(X) = \frac{1-\eta(X)}{1+\eta(X)}. \tag{12}$$

We will show that problem (17) actually has a simple closed-form solution. First, we write its Lagrangian dual as

$$\begin{array}{ll} \underset{y,U_1,U_2}{\text{minimize}} & \text{tr}(U_2) \\ \text{subject to} & (\mathbf{X}y)\mathbf{e}^T + \mathbf{e}(\mathbf{X}y)^T = U_1 - U_2 \\ & \text{tr}(U_1) = 1, \quad U_1, U_2 \succeq 0. \end{array} \tag{13}$$

Notice that strong duality holds because Slater's condition is trivially satisfied by the dual: $y = 0$ and $U_1 = U_2 = 2I/n(n+1)$ is a strictly feasible point. It turns out that the dual problem can be rewritten as an optimization problem over the eigenvalues of the matrix $(\mathbf{X}y)\mathbf{e}^T + \mathbf{e}(\mathbf{X}y)^T$. The proof of this in in Appendix A.

For any $\alpha \in \mathbb{R}$ we denote $[\alpha]_+ = \max\{0, +\alpha\}$ and $[\alpha]_- = \max\{0, -\alpha\}$. The dual problem can be written as

$$\min_y \frac{\text{tr}[M(y)]_-}{\text{tr}[M(y)]_+} = \min_y \frac{\sum_i \lambda_i[M(y)]_-}{\sum_i \lambda_i[M(y)]_+}, \quad \text{where} \quad M(y) = (\mathbf{X}y)\mathbf{e}^T + \mathbf{e}(\mathbf{X}y)^T,$$

and $\lambda_i[M(y)]$ denotes the eigenvalues of the rank-2 matrix $M(y)$. It is easy to verify that the only two non-zero eigenvalues of $(\mathbf{X}y)\mathbf{e}^T + \mathbf{e}(\mathbf{X}y)^T$ are

$$\|\mathbf{X}y\|\|\mathbf{e}\| \left(\cos\theta_y \pm 1\right), \quad \text{where } \cos\theta_y = \frac{\mathbf{e}^T\mathbf{X}y}{\|\mathbf{e}\|\|\mathbf{X}y\|}.$$

It follows that

$$\eta(X) = \min_y \frac{1-\cos\theta_y}{1+\cos\theta_y}$$

and therefore

$$\delta_{\text{foc}}(X) = \max_y \cos\theta_y = \max_y \frac{\mathbf{e}^T\mathbf{X}y}{\|\mathbf{e}\|\|\mathbf{X}y\|}.$$

Let y^* be the optimizer of the optimization problem above, then θ_{y^*} is simply the incidence angle between the column space of X and the error vector $\mathbf{e}$. Thus we have $y^\star = \arg\min_y \|\mathbf{e} - \mathbf{X}y\|$. Using Lemma 12 in Appendix A, we show that solving for y^* yields a closed-form expression for θ_{y^*} in the form

$$\sin\theta_{y^*} = \frac{\|Z^T(I - XX^\dagger)Z\|_F}{\|XX^T - ZZ^T\|_F}.$$

Hence we have $\delta_{\text{foc}}(X) = \cos\theta$, with $\theta = \theta_{y^*}$ given by the equation above.

5.3 Proof of Theorem 5

The proof of Theorem 5 is based on the following lemma. Its proof is very technical and can be can be found in Appendix B.

Lemma 7. *Let $Z \neq 0$ and suppose that $\|XX^T - ZZ^T\|_F \leq \epsilon\|ZZ\|_F^2$. Then*

$$\sin^2\theta = \frac{\|Z^T(I - XX^\dagger)Z\|_F^2}{\|XX^T - ZZ^T\|_F^2} \leq \frac{\epsilon}{2\sigma_{\min}^2(Z)/\|ZZ^T\|_F - \epsilon}.$$

To prove Theorem 5, we simply set $C_1 = \sigma_{\min}^2(Z)/\|ZZ^T\|_F$ and write

$$\cos\theta = \sqrt{1-\sin^2\theta} \geq \sqrt{1-\frac{\epsilon}{2C_1-\epsilon}}.$$

It is easy to see that $\frac{\epsilon}{2C_1-\epsilon}$ is dominated by the linear function ε/C_1 so long as $\epsilon \leq C_1$. This follows directly from the fact that $\frac{\epsilon}{2C_1-\epsilon}$ is convex between 0 and C_1. Thus we have

$$\cos\theta \geq \sqrt{1-\frac{\epsilon}{C_1}}$$

Since this lower bound holds for all X in $\mathcal{B}_\varepsilon$, it follows that $\delta_{\mathrm{foc}}(\mathcal{B}_\varepsilon) \geq \sqrt{1-\varepsilon/C_1}$.

6 Numerical Results

In this section we give a geometric interpretation for Theorem 8, which we already alluded to in section 4: the sample complexity to eliminate spurious critical points is small in regions where the column spaces of X and Z are more aligned and large in regions where they are orthogonal. We also numerically verify that $\delta_{\mathrm{foc}}(X)$ is a tight lower bound for $\delta_{\mathrm{soc}}(X)$ for a wide range of ε, providing numerical evidence that the bound in Theorem 5 is tight.

Our main results and geometric insights hold for *any rank*, but for ease of visualization we focus on the rank-1 case where x and z are now just vectors. To measure the alignment between the column space of x and that of z in the rank-1 case , we define the length ratio and the incidence angle as

$$\rho = \frac{\|x\|}{\|z\|}, \quad \cos\phi = \frac{x^T z}{\|x\|\|z\|}.$$

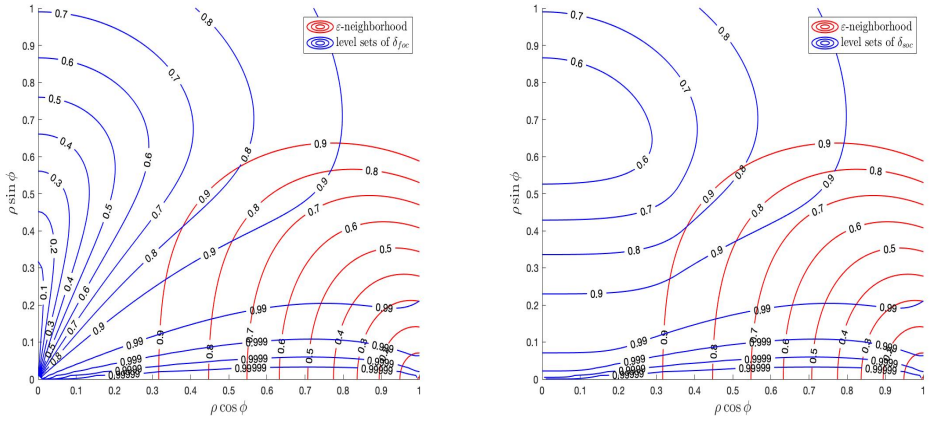
Our goal is to plot how sample complexity depends on this alignment. Visualizing the dependence of sample complexity on ρ and $\cos\phi$ is particularly easy in rank-1 because these two parameters completely determine the values of both $\delta_{\mathrm{foc}}(x)$ and $\delta_{\mathrm{soc}}(x)$. See [23] section 8.1 for a proof of this fact. This allows us to plot the level curves of $\delta_{\mathrm{foc}}(x)$ and $\delta_{\mathrm{soc}}(x)$ over the parameter space ρ and ϕ in Figure 2. This is shown by the blue curves. Since we are particularly interested in sample complexity near the ground truth, we also plot the level sets of the function $\|xx^T - zz^T\|_F/\|zz^T\|_F$ using red curves. The horizontal axis is the value of $\rho\cos\phi$ and the vertical axis is the value of $\rho\sin\phi$.

We can immediately see that in regions in the optimization landscape where x is more aligned with z, i.e., when $\sin\phi$ is small, the values of both threshold functions tend to be high and a relatively small number of samples suffices to prevent x from becoming a spurious critical point. However, when x and z becomes closer to being orthogonal, i.e., when $\cos\phi$ is close to 0, then $\delta_{\mathrm{foc}}(x)$ becomes arbitrarily small, and $\delta_{\mathrm{soc}}(x)$ also becomes smaller, albeit to a lesser extent. As a result, preventing x from becoming a spurious critical point (or spurious local minima) in these regions require many more samples. This intuition also permeates to the high-rank case, even though visualization becomes difficult, and a slightly more general definition of length ratio and alignment is required. Similar to the rank-1 case, in regions where XX^T and ZZ^T are more aligned, the sample complexity required to eliminate spurious critical points is small and in regions where XX^T and ZZ^T are close to orthogonal, a small sample complexity is required.

Regarding the tightness of using $\delta_{\mathrm{foc}}(X)$ as a lower bound for $\delta_{\mathrm{soc}}(X)$, note that if we look at the level sets of $\|xx^T - zz^T\|_F/\|zz^T\|_F$, we see that in regions close to the ground truth, both $\delta_{\mathrm{soc}}(x)$ and $\delta_{\mathrm{foc}}(x)$ are very close to 1. This is in perfect agreement with our results in Theorem 5, where we showed that a small ε results in a large $\delta_{\mathrm{foc}}(\mathcal{B}_\varepsilon)$. Moreover, the shapes of the level curves of δ_{soc} and δ_{foc} that flow through the regions near the ground truth are almost identical. This indicates that for a large region near the ground truth, the second-order condition, i.e., the hessian being positive semidefinite, is inactive. This is the underlying mechanism that causes δ_{foc} to be a tight lower bound for δ_{foc}.

7 Conclusions

Recent work by Bhojanapalli et al. [2] has shown that the non-convex optimization landscape of matrix sensing contains no spurious local minima when there are sufficiently large amount of samples.



(a) (b)

Figure 2: (a) the level sets of δ_{foc} and $\|xx^T - zz^T\|_F / \|zz^T\|_F$ (b) the level sets of δ_{soc} and $\|xx^T - zz^T\|_F / \|zz^T\|_F$

However, these theoretical bounds on the sample complexity are very conservative compared to the number of samples needed in real applications like power state estimation. In our paper, we provide one explanation for this phenomenon: in real life, we often have access to good initial points, which can reduce the number of samples we need. The main results of our paper give a mathematical characterization of this phenomenon. We define a function $\delta_{\text{soc}}(X)$ that gives a *precise* threshold on the number of samples needed to prevent X from becoming a spurious local minima. Although δ_{soc} is difficult to compute exactly, we obtain a closed-form, sharp lower bound using convex optimization. As a result, we are able to characterize the *tradeoff* between the quality of the initial point and the sample complexity. In particular, we show that a linear improvement in the quality of the initial point corresponds to a linear decrease in sample complexity.

On a more general level, our work uses new techniques to paint a full picture for the non-convex landscape of matrix sensing: the problem becomes more "non-convex" (requiring more samples to eliminate spurious local minima) as we get further and further away from the global min. Once we are sufficiently far away, it becomes necessary to rely on global guarantees instead. Thus, our work brings new insight into how a non-convex problem can gradually become more tractable either through more samples or a better initial point and provides a tradeoff between these two mechanisms. For future work, it would be interesting to see if similar techniques can be extended to other non-convex models such as neural networks.

Acknowledgements

Partial financial support was provided by the National Science Foundation under award ECCS-1808859.

Broader Impact

Many modern applications in engineering and computer science, and in machine learning in particular often have to deal with non-convex optimization. However, many aspects of non-convex optimization are still not well understood. Our paper provides more insight into the optimization landscape of a particular problem: low-rank matrix factorization. In addition, the methods we develop can potentially be used to understand many other non-convex problems. This is a step towards a more thorough analysis of current algorithms for non-convex optimization and also a step towards developing better and more efficient algorithms with theoretical guarantees.

References

[1] Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582, 2016. 1, 2

[2] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016. 2, 3, 4, 7

[3] Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016. 1, 2

[4] E Candes and Y Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. to appear. *IEEE Trans. Info. Theo*, 2009. 3

[5] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015. 1, 2

[6] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015. 2

[7] Yuxin Chen and Emmanuel Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747, 2015. 1

[8] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019. 1, 3

[9] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016. 1, 2

[10] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013. 1, 2, 3

[11] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010. 1, 2

[12] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2019. 1, 2

[13] Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and computational harmonic analysis*, 47(3):893–934, 2019. 1

[14] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, pages 1–182, 2019. 2

[15] Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014. 1

[16] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010. 3

[17] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery using nonconvex optimization. In *International Conference on Machine Learning*, pages 2351–2360, 2015. 1, 2

[18] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018. 1, 2

[19] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016. 1, 2, 3

[20] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973, 2016. 1

[21] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015. 2

[22] Richard Zhang, Cédric Josz, Somayeh Sojoudi, and Javad Lavaei. How much restricted isometry is needed in nonconvex matrix recovery? In *Advances in neural information processing systems*, pages 5586–5597, 2018. 1

[23] Richard Y Zhang, Somayeh Sojoudi, and Javad Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *Journal of Machine Learning Research*, 20(114):1–34, 2019. 2, 3, 3, 4, 4, 5.2, 6, 7

[24] Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015. 2

Appendix A.1

In Appendix A we fill out the missing details in the proof sketch of Section 5.2 and provide a complete proof of Theorem 4, which we restate below.

Theorem 8. *(Same as theorem 4). Fix $Z \in \mathbb{R}^{n\times r}$. Given $\mathcal{A}$ satisfying δ-RIP and $X \in \mathbb{R}^{n\times r}$ such that $XX^T \neq ZZ^T$, we have $\delta_{\text{foc}}(X) = \cos\theta$, where*

$$\sin\theta = \|Z^T(I - XX^\dagger)Z\|_F \;/\; \|XX^T - ZZ^T\|_F. \tag{14}$$

and $X^\dagger$ denotes the pseudo-inverse of X. It follows that if $\delta < \cos\theta$, then X is not a spurious critical point of $f_{\mathcal{A}}(X)$. If $\delta \geq \cos\theta$, then there exists some $\mathcal{A}^\star$ satisfying $\cos\theta$-RIP such that $\nabla f_{\mathcal{A}}(X) = 0$.

Before we prove the theorem above, we first prove two technical lemmas. The first lemma gives an explicit solution to the eigenvalues of a rank-2 matrix and the second lemma characterizes the solution to an SDP that will be a part of the proof of theorem 4.

Lemma 9. *Given $a, b \in \mathbb{R}^n$, the matrix $M = ab^T + ba^T$ has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ where:*

$$\lambda_i = \begin{cases} +\|a\|\|b\|(1+\cos\theta) & i = 1 \\ -\|a\|\|b\|(1-\cos\theta) & i = n \\ 0 & \text{otherwise} \end{cases}$$

and $\theta \equiv \arccos\left(\frac{a^Tb}{\|a\|\|b\|}\right)$ is the angle between a and b.

Lemma 10. *Given a matrix $M \neq 0$ we can split the matrix M into a positive and negative part satisfying*

$$M = M_+ - M_- \quad \textit{where} \quad M_+, M_- \succeq 0, \quad M_+M_- = 0.$$

Then the following problem has solution

$$\min_{\substack{\alpha\in\mathbb{R}\\ U,V\succeq 0}} \{\text{tr}(V) : \text{tr}(U) = 1, \alpha M = U - V\} = \min\left\{\frac{\text{tr}\,(M_-)}{\text{tr}\,(M_+)}, \frac{\text{tr}\,(M_+)}{\text{tr}\,(M_-)}\right\}.$$

Proof. (Lemma 9). Without loss of generality, assume that $\|a\| = \|b\| = 1$. (Otherwise, we can rescale $\hat{a} = a/\|a\|, \hat{b} = b/\|b\|$ and write $M = \|a\|\|b\|(\hat{a}\hat{b}^T + \hat{b}\hat{a}^T)$. Now decompose b into a tangent and normal component with respect to a, as in

$$b = a\underbrace{a^Tb}_{\cos\theta} + \underbrace{\left(I - aa^T\right)b}_{c\sin\theta} = a\cos\theta + c\sin\theta$$

where c is a unit normal vector with $\|c\| = 1$ and $a^Tc = 0$. Thus $ab^T + ba^T$ can be written as

$$ab^T + ba^T = \begin{bmatrix} a & c \end{bmatrix} \begin{bmatrix} 2\cos\theta & \sin\theta \\ \sin\theta & 0 \end{bmatrix} \begin{bmatrix} a & c \end{bmatrix}^T.$$

This shows that M is spectrally similar to a 2×2 matrix with eigenvalues $\cos\theta \pm 1$. □

Proof. (Lemma 10). In this proof we will consider two cases: $\text{tr}(M_-) \leq \text{tr}(M_+)$ and $\text{tr}(M_-) \geq \text{tr}(M_+)$. We'll see that in the first case, the optimal value is $\text{tr}(M_-)/\text{tr}(M_+)$ and in the second case, the optimal value is $\text{tr}(M_+)/\text{tr}(M_1)$.

First, assume that $\text{tr}(M_-) \leq \text{tr}(M_+)$. Let p^* be the optimal value. Then we have

$$\begin{aligned} p^\star &= \max_\beta \min_{\substack{\alpha\in\mathbb{R}\\ U,V\succeq 0}} \{\text{tr}(V) + \beta\cdot[1 - \text{tr}(U)] : \alpha M = U - V\} \qquad (15)\\ &= \max_\beta \min_{\alpha\in\mathbb{R}} \left\{\beta + \min_{U,V\succeq 0}\{\text{tr}(V) - \beta\cdot\text{tr}(U) : \alpha M = U - V\}\right\} \\ &= \max_\beta \min_{\alpha\in\mathbb{R}} \left\{\beta + \min_U \left[\text{tr}\,(U - \alpha M) - \beta\cdot\text{tr}\,(U)\right] : U - \alpha M \succeq 0, U \succeq 0\right\} \\ &= \max_\beta \min_{\alpha\in\mathbb{R}} \left\{\beta + \min_U [-\alpha\text{tr}(M) + (1-\beta)\text{tr}(U)] : U - \alpha M \succeq 0, U \succeq 0\right\} \\ &= \max_{\beta\leq 1} \min_{\alpha\in\mathbb{R}} \left\{\beta + \min_U [-\alpha\text{tr}(M) + (1-\beta)\text{tr}(U)] : U - \alpha M \succeq 0, U \succeq 0\right\}. \qquad (16) \end{aligned}$$

Note that the first line converts the equality constraint into a Lagrangian. The second line simply rearranges the terms. The third line plugs in $V = U - \alpha M$. The fourth line again rearranges the terms. The last line follows from the observation that if $\beta > 1$, then the inner minimization over U will go to negative infinity since the trace of U can be arbitrarily large.

First, consider the case $\alpha \geq 0$. Then we have $\alpha M = \alpha M_+ - \alpha M_-$. Since $1 - \beta \geq 0$, the minimization over U is achieved at $U = \alpha M_+$. Plugging this value into the optimization problem, then (19) becomes

$$\max_{\beta \leq 1} \min_{\alpha \geq 0} \{\beta + \alpha[\mathrm{tr}(M_-) - \beta \mathrm{tr}(M_+)]\}$$

If $\mathrm{tr}(M_-) - \beta \mathrm{tr}(M_+) < 0$, then the optimal value of the inner minimization will go to negative infinity. On the other hand, if $\mathrm{tr}(M_-) - \beta \mathrm{tr}(M_+) \geq 0$ then the minimum inside is achieved at $\alpha = 0$. Thus the problem above is equivalent to

$$\max_{\beta \leq 1} \{\beta : \mathrm{tr}(M_-) - \beta \mathrm{tr}(M_+) \geq 0\}.$$

Since $\mathrm{tr}(M_-) \leq \mathrm{tr}(M_+)$, the optimal value of the problem above is achieved at $\mathrm{tr}(M_-)/\mathrm{tr}(M_+) \leq 1$. Now suppose that $\alpha \leq 0$. Then the optimal value for U is achieved at $U = -\alpha M_-$. Plugging this value in and (19) becomes

$$\max_{\beta \leq 1} \min_{\alpha \leq 0} \{\beta + \alpha[\beta \mathrm{tr}(M_-) - \mathrm{tr}(M_+)]\}.$$

Similar to before, we must have $\beta \mathrm{tr}(M_-) - \mathrm{tr}(M_+) \leq 0$, so $\beta \leq \mathrm{tr}(M_+)/\mathrm{tr}(M_-)$. Since $\mathrm{tr}(M_-) \leq \mathrm{tr}(M_+)$, the optimal value in this case is just $\beta = 1$. Combining the results for $\alpha \geq 0$ and $\alpha \leq 0$, we find that when $\mathrm{tr}(M_-) \leq \mathrm{tr}(M_+)$, the optimal value is

$$p^* = \min \left\{ 1, \frac{\mathrm{tr}(M_-)}{\mathrm{tr}(M_+)} \right\} = \frac{\mathrm{tr}(M_-)}{\mathrm{tr}(M_+)}.$$

Repeating the same arguments for when $\mathrm{tr}(M_-) \geq \mathrm{tr}(M_+)$, we see that in this case the optimal value becomes

$$p^* = \min \left\{ \frac{\mathrm{tr}(M_+)}{\mathrm{tr}(M_-)}, 1 \right\} = \frac{\mathrm{tr}(M_+)}{\mathrm{tr}(M_-)}.$$

Finally, combining these two cases, i.e., $\mathrm{tr}(M_-) \geq \mathrm{tr}(M_+)$ and $\mathrm{tr}(M_-) \leq \mathrm{tr}(M_+)$, we obtain

$$p^* = \min \left\{ \frac{\mathrm{tr}\,(M_-)}{\mathrm{tr}\,(M_+)}, \frac{\mathrm{tr}\,(M_+)}{\mathrm{tr}\,(M_-)} \right\},$$

which completes the proof.

□

Appendix A.2

Now we are ready to prove Theorem 4. Recall that the first order threshold function is defined as the solution to the following optimization problem:

$$\delta_{\mathrm{foc}}(X) \equiv \min_{\mathcal{A}} \{\delta : \nabla f_{\mathcal{A}}(X) = 0, \quad \mathcal{A} \text{ satisfies } \delta\text{-RIP}\}$$

Using Theorem 8 from [23], the optimization problem above can be formulated as

$$\eta(X) \quad \equiv \quad \max_{\eta, \mathbf{H}} \left\{ \eta \quad : \quad \mathbf{X}^T \mathbf{H} \mathbf{e} = 0, \quad \eta I \preceq \mathbf{H} \preceq I \right\}. \tag{17}$$

where $\eta = (1 - \delta_{\mathrm{foc}})/(1 + \delta_{\mathrm{foc}})$. Our goal is to solve this optimization problem in closed form. In Section 5.2, we wrote the dual of problem (17) as

$$\begin{aligned} \min_{y, U_1, U_2} \quad & \mathrm{tr}(U_2) \\ \text{subject to} \quad & (\mathbf{X}y)\mathbf{e}^T + \mathbf{e}(\mathbf{X}y)^T = U_1 - U_2 \\ & \mathrm{tr}(U_1) = 1, \quad U_1, U_2 \succeq 0. \end{aligned} \tag{18}$$

and stated that this dual problem can be rewritten as an optimization problem over the eigenvalues of a rank-2 matrix. This is given in the lemma below. To simplify notation, here we define a positive/negative splitting: for any $\alpha \in \mathbb{R}_+$ we denote $[\alpha]_+ = \max\{0, +\alpha\}$ and $[\alpha]_- = \max\{0, -\alpha\}$. This idea can be extended to matrices by applying splitting to the eigenvalues.

Lemma 11. *Given data* $\mathbf{e}$ *and* $\mathbf{X} \neq 0$*, define*

$$\begin{aligned} \eta = \min_{y, U_1, U_2} \quad & \operatorname{tr}(U_2) \\ \text{subject to} \quad & (\mathbf{X}y)\mathbf{e}^T + \mathbf{e}(\mathbf{X}y)^T = U_1 - U_2 \\ & \operatorname{tr}(U_1) = 1, \quad U_1, U_2 \succeq 0. \end{aligned} \tag{19}$$

Define $M(y)$ *to be the rank-2 matrix* $(\mathbf{X}y)\mathbf{e}^T + \mathbf{e}(\mathbf{X}y)^T$ *and let* $\lambda_i[M(y)]$ *denote its eigenvalues. Then* η *can be evaluated as*

$$\eta = \min_{y \neq 0} \frac{\operatorname{tr}[M(y)]_-}{\operatorname{tr}[M(y)]_+} = \min_{y \neq 0} \frac{\sum_i \lambda_i[M(y)]_-}{\sum_i \lambda_i[M(y)]_+} = \min_{y \neq 0} \frac{1 - \cos\theta_y}{1 + \cos\theta_y},$$

where $\cos\theta_y = \mathbf{e}^T\mathbf{X}y / \|\mathbf{e}\|\|\mathbf{X}y\|$.

The proof of Lemma 11 relies mainly on the two lemmas we proved in the preceding section.

Proof. (Lemma 11). Let $y = \alpha\hat{y}$, where $\|\hat{y}\| = 1$ and $\alpha \in \mathbb{R}^n$. Thus the optimization problem (19) becomes

$$\begin{aligned} \eta = \min_{\alpha, \hat{y}, U_1, U_2} \quad & \operatorname{tr}(U_2) \\ \text{subject to} \quad & \alpha \cdot [(\mathbf{X}\hat{y})\mathbf{e}^T + \mathbf{e}(\mathbf{X}\hat{y})^T] = U_1 - U_2 \\ & \operatorname{tr}(U_1) = 1, \quad \|\hat{y}\| = 1, \quad U_1, U_2 \succeq 0. \end{aligned}$$

To solve this problem, first we keep $\hat{y}$ fixed, and optimize over α, U_1, U_2. This gives us the problem

$$\begin{aligned} \min_{\alpha, U_1, U_2} \quad & \operatorname{tr}(U_2) \\ \text{subject to} \quad & \alpha \cdot [(\mathbf{X}\hat{y})\mathbf{e}^T + \mathbf{e}(\mathbf{X}\hat{y})^T] = U_1 - U_2 \\ & \operatorname{tr}(U_1) = 1, \quad U_1, U_2 \succeq 0. \end{aligned}$$

Notice that if we set $M(\hat{y}) = (\mathbf{X}\hat{y})\mathbf{e}^T + \mathbf{e}(\mathbf{X}\hat{y})^T$, then the problem above is in exactly the same form as the one in lemma 10. Therefore, its optimal value is

$$\min\left\{ \frac{\operatorname{tr}(M(\hat{y})_-)}{\operatorname{tr}(M(\hat{y})_+)}, \frac{\operatorname{tr}(M(\hat{y})_+)}{\operatorname{tr}(M(\hat{y})_-)} \right\}.$$

Finally, to obtain η, we still need to optimize over $\hat{y}$, i.e.,

$$\eta = \min_{\|\hat{y}\|=1} \min\left\{ \frac{\operatorname{tr}(M(\hat{y})_-)}{\operatorname{tr}(M(\hat{y})_+)}, \frac{\operatorname{tr}(M(\hat{y})_+)}{\operatorname{tr}(M(\hat{y})_-)} \right\}.$$

Since both the numerator and the denominator are linear in y, we can ignore the constraint $\|\hat{y}\| = 1$ and simply optimize over y, which gives us

$$\eta = \min_{y \neq 0} \min\left\{ \frac{\operatorname{tr}(M(y)_-)}{\operatorname{tr}(M(y)_+)}, \frac{\operatorname{tr}(M(y)_+)}{\operatorname{tr}(M(y)_-)} \right\}.$$

With lemma 9, we see that the only two eigenvalues of $M(y)$ are

$$\|\mathbf{X}y\|\|y\|(\cos\theta_y + 1), \qquad \|\mathbf{X}y\|\|y\|(\cos\theta_y - 1),$$

where $\cos\theta_y = \mathbf{e}^T\mathbf{X}y / \|\mathbf{e}\|\|\mathbf{X}y\|$. It follows that $\operatorname{tr}(M_-) = \|\mathbf{X}y\|\|y\|(1 - \cos\theta_y)$ and $\operatorname{tr}(M_+) = \|\mathbf{X}y\|\|y\|(\cos\theta_y + 1)$. Thus

$$\eta = \min_{y \neq 0} \min\left\{ \frac{1 - \cos\theta_y}{1 + \cos\theta_y}, \frac{1 + \cos\theta_y}{1 - \cos\theta_y} \right\}.$$

Notice that in the optimization problem above, if the minimum is achieved at some y^*, it must also be achieved at $-y^*$, due to symmetry. Therefore, it suffices to optimize over only the first term $\frac{1-\cos\theta_y}{1+\cos\theta_y}$, so we get

$$\eta = \min_{y \neq 0} \frac{1 - \cos\theta_y}{1 + \cos\theta_y}.$$

This completes the proof. □

Notice that Lemma 11 reduces problem 17 to only depend on the values of $\cos\theta_y$. Now, to complete the proof of Theorem 4, we just need one additional lemma that gives a closed form solution for $\cos\theta_y$, which we state below.

Lemma 12. *Let X, Z be $n \times r$ matrices of any rank, and define $\mathbf{e}$ and $\mathbf{X} \neq 0$ as in equations 10(a) and 10(b). Then, the incidence angle θ between $\mathbf{e}$ and $\mathrm{range}(\mathbf{X})$, defined as in*

$$\cos\theta = \max_{y\neq 0}\left\{\frac{\mathbf{e}^T\mathbf{X}y}{\|\mathbf{e}\|\|\mathbf{X}y\|}\right\} = \frac{\|\mathbf{X}\mathbf{X}^\dagger\mathbf{e}\|}{\|\mathbf{e}\|},$$

has closed-form expression

$$\sin\theta = \frac{\|Z^T(I - XX^\dagger)Z\|_F}{\|XX^T - ZZ^T\|_F}$$

where $X^\dagger$ denotes the Moore–Penrose pseudoinverse of X.

Proof. (Lemma 12). Define $y^\star = \arg\min_y \|\mathbf{e} - \mathbf{X}y\|$ and decompose $\mathbf{e} = \mathbf{X}y^\star + w$. The optimality condition for $y^\star$ reads $\mathbf{X}^T(\mathbf{e} - \mathbf{X}y^\star) = \mathbf{X}^T w = 0$, so we substitute $\mathbf{e}^T\mathbf{X} = (y^*)^T\mathbf{X}^T\mathbf{X}$ to yield

$$\|\mathbf{e}\|\cos\theta = \|\mathbf{e}\|\max_{y\neq 0}\left\{\frac{\mathbf{e}^T\mathbf{X}y}{\|\mathbf{e}\|\|\mathbf{X}y\|}\right\} = \max_{y\neq 0}\left\{\frac{(y^\star)^T\mathbf{X}^T\mathbf{X}y}{\|\mathbf{X}y\|}\right\} = \|\mathbf{X}y^\star\|,$$

and therefore $\|\mathbf{e}\|\sin\theta = \|w\| = \min_y \|\mathbf{e} - \mathbf{X}y\|$, because we have $\mathbf{e} = \mathbf{X}y^* + w$ with $w^T\mathbf{X}y^* = 0$. Now, define $Q = \mathrm{orth}(X) \in \mathbb{R}^{n\times q}$ where $q = \mathrm{rank}(X) \leq r$, and define $P \in \mathbb{R}^{n\times(n-q)}$ as the orthogonal complement of Q. Decompose $X = Q\hat{X}$, and $Z = Q\hat{Z}_1 + P\hat{Z}_2$, and note that

$$\begin{aligned}
\|w\| &= \min_y \|\mathbf{e} - \mathbf{X}y\| \\
&= \min_Y \|(XX^T - ZZ^T) - (XY^T + YX^T)\|_F \\
&= \min_{[\hat{Y}_1;\hat{Y}_2]\in\mathbb{R}^{n\times r}} \left\| \begin{bmatrix} \hat{X}\hat{X}^T - \hat{Z}_1\hat{Z}_1^T & -\hat{Z}_1\hat{Z}_2^T \\ -\hat{Z}_2\hat{Z}_1^T & -\hat{Z}_2\hat{Z}_2^T \end{bmatrix} - \begin{bmatrix} \hat{X}\hat{Y}_1^T + \hat{Y}_1\hat{X}^T & \hat{X}\hat{Y}_2^T \\ \hat{Y}_2\hat{X}^T & 0 \end{bmatrix} \right\|_F \\
&= \|\hat{Z}_2\hat{Z}_2^T\|_F
\end{aligned}$$

From the second line to the third, we apply a change of basis onto $[Q\ P]$, which preserves the Frobenius norm. To derive the last line, notice that the $q \times r$ matrix $\hat{X}$ has full row rank, so that $\hat{X}\hat{X}^T \succ 0$ and $\hat{X}\hat{X}^\dagger = I_q$. We want to show that there exists $\hat{Y}_1$ such that

$$\hat{X}\hat{Y}_1^T + \hat{Y}_1\hat{X}^T = \hat{X}\hat{X}^T - \hat{Z}_1\hat{Z}_1^T.$$

Since the right hand side is symmetric, we can write it as $L + L^T$, where L is some lower-triangular matrix. Thus it suffices to show that there exists $\hat{Y}_1$ such that $\hat{X}\hat{Y}_1^T = L$, which follows from that fact that $\hat{X}$ has full row-rank. Similarly, there exists some $\hat{Y}_2$ such that $\hat{X}\hat{Y}_2 = -\hat{Z}_2\hat{Z}_1^T$. Thus, all terms except the last one cancels out and we are left with $\min_y \|\mathbf{e} - \mathbf{X}y\| = \|\hat{Z}_2\hat{Z}_2^T\|_F$.

Finally, note that $Q\hat{Z}_1 = XX^\dagger Z$ and $P\hat{Z}_2 = (I - XX^\dagger)Z$ and that

$$\begin{aligned}
\|\hat{Z}_2\hat{Z}_2^T\|_F^2 &= \|P\hat{Z}_2\hat{Z}_2^T P^T\|_F^2 \\
&= \|(I - XX^\dagger)ZZ^T(I - XX^\dagger)\|_F^2 \\
&= \mathrm{tr}[(I - XX^\dagger)ZZ^T(I - XX^\dagger)ZZ^T(I - XX^\dagger)] \\
&= \mathrm{tr}[Z^T(I - XX^\dagger)ZZ^T(I - XX^\dagger)Z] \\
&= \|Z^T(I - XX^\dagger)Z\|_F^2.
\end{aligned}$$

Substituting the definition of $\mathbf{e}$ completes the proof. □

Now theorem 4 will be a direct consequence of lemma 11 and lemma 12. We give a proof below.

Proof. (Theorem 4). Note that δ_{foc} is related to η by the equation

$$\eta = \frac{1-\delta_{\text{foc}}}{1+\delta_{\text{foc}}}.$$

Applying lemma 11, we immediately get

$$\delta_{\text{foc}}(X) = \max_{y\neq 0} \cos\theta_y = \max_{y\neq 0} \frac{\mathbf{e}^T \mathbf{X} y}{\|\mathbf{e}\|\|\mathbf{X}y\|}.$$

From lemma 12, we see that this optimization problem over y has a simple closed form solution of the form

$$\delta_{\text{foc}}(X) = \cos\theta, \quad \text{where } \sin\theta = \frac{\|Z^T(I-XX^\dagger)Z\|_F}{\|XX^T - ZZ^T\|_F}.$$

This completes the proof. □

Appendix B

In this section we provide a complete proof of Theorem 5, which includes all the intermediate calculations that was skipped in Section 5.3. We begin by proving a bound on $\sin\theta$.

Lemma 13 (Same as Lemma 7)**.** *Let $Z \neq 0$ and suppose that $\|XX^T - ZZ^T\|_F \leq \epsilon \|ZZ\|_F^2$. Then*

$$\sin^2\theta = \frac{\|Z^T(I-XX^\dagger)Z\|_F^2}{\|XX^T-ZZ^T\|_F^2} \leq \frac{\epsilon}{2(\sigma_{\min}^2(Z)/\|ZZ^T\|_F) - \epsilon}.$$

Proof. The problem is homogeneous to scaling $X \leftarrow \alpha X$ and $Z \leftarrow \alpha Z$ for the same α; Since $Z \neq 0$, we may rescale X and Z until $\|ZZ\|_F^2 = 1$. Additionally, we can assume that

$$X = \begin{bmatrix} X_1 \\ 0 \end{bmatrix} \qquad Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \qquad \text{where } X_1, Z_1 \in \mathbb{R}^{r\times r}, Z_2 \in \mathbb{R}^{(n-r)\times r}$$

due to the rotational invariance of the problem. (Concretely, we compute the QR decomposition $QR = [X, Z]$ with $Q \in \mathbb{R}^{n\times 2r}$ noting that $X = QQ^TX$ and $Z = QQ^TZ$. We then make a change of basis $X \leftarrow Q^TX$ and $Z \leftarrow Q^TZ$). Then, observe that

$$\|Z^T(I-XX^\dagger)Z\|_F = \left\| \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}^T \left(I - \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \right) \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \right\|_F = \|Z_2^T Z_2\|_F = \|Z_2 Z_2^T\|_F \tag{20}$$

and that $\|Z_2Z_2^T\|_F^2 \leq \epsilon^2$ because

$$\begin{aligned} \|XX^T - ZZ^T\|_F^2 &= \left\| \begin{bmatrix} Z_1Z_1^T - X_1X_1^T & Z_1Z_2^T \\ Z_2Z_{1in}^T & Z_2Z_2^T \end{bmatrix} \right\|_F^2 \\ &= \|Z_1Z_1^T - X_1X_1^T\|_F^2 + 2\|Z_1Z_2^T\|_F^2 + \|Z_2Z_2^T\|_F^2 \leq \epsilon^2. \end{aligned} \tag{21}$$

In order to derive a non-vacuous bound, we will need to lower-bound the term $\|Z_1Z_2^T\|_F^2$ as follows

$$\|Z_1Z_2^T\|_F^2 = \text{tr}(Z_1^TZ_1Z_2^TZ_2) \geq \lambda_{\min}(Z_1^TZ_1)\text{tr}(Z_2^TZ_2) = \sigma_{\min}^2(Z_1)\|Z_2\|_F^2. \tag{22}$$

To lower-bound $\sigma_{\min}^2(Z_1)$, observe that

$$A + B \succeq \mu I \iff A \succeq \mu I - B \succeq (\mu - \|B\|_2)I,$$

and therefore

$$\begin{aligned} \sigma_{\min}^2(Z_1) = \lambda_{\min}(Z_1^TZ_1) &\geq \lambda_{\min}(Z_1^TZ_1 + Z_2^TZ_2) - \lambda_{\max}(Z_2^TZ_2) \\ &= \sigma_{\min}^2(Z) - \|Z_2Z_2^T\|_2 \geq \sigma_{\min}^2(Z) - \|Z_2Z_2^T\|_F \\ &\geq \sigma_{\min}^2(Z) - \epsilon. \end{aligned} \tag{23}$$

Finally, we substitute (20) and (21) and perform a sequence of reductions:

$$\begin{aligned}
\frac{\|Z^T(I - XX^\dagger)Z\|_F^2}{\|XX^T - ZZ^T\|_F^2} &= \frac{\|Z_2 Z_2^T\|_F^2}{\|Z_1 Z_1^T - X_1 X_1^T\|_F^2 + 2\|Z_1 Z_2^T\|_F^2 + \|Z_2 Z_2^T\|_F^2} \\
&\overset{\text{(a)}}{\leq} \frac{\|Z_2 Z_2^T\|_F^2}{2\|Z_1 Z_2^T\|_F^2 + \|Z_2 Z_2^T\|_F^2} \overset{\text{(b)}}{\leq} \frac{\|Z_2 Z_2^T\|_F^2}{2\sigma_{\min}^2(Z_1)\|Z_2\|_F^2 + \|Z_2 Z_2^T\|_F^2} \\
&\overset{\text{(c)}}{\leq} \frac{\|Z_2 Z_2^T\|_F \|Z_2\|_F^2}{2\sigma_{\min}^2(Z_1)\|Z_2\|_F^2 + \|Z_2 Z_2^T\|_F \|Z_2\|_F^2} = \frac{\|Z_2 Z_2^T\|_F}{2\sigma_{\min}^2(Z_1) + \|Z_2 Z_2^T\|_F} \\
&\overset{\text{(d)}}{\leq} \frac{\epsilon}{2(\sigma_{\min}^2(Z) - \epsilon) + \epsilon} \leq \frac{\epsilon}{2\sigma_{\min}^2(Z) - \epsilon}.
\end{aligned}$$

Step (a) sets $X_1 = Z_1$ to minimize the denominator; step (b) bounds $\|Z_1 Z_2^T\|_F^2$ using (22); step (c) bounds $\|Z_2 Z_2^T\|_F \leq \|Z_2\|_F^2$ noting that a function like $x/(1+x)$ is increasing with x; step (d) substitutes $\|Z_2 Z_2\|_F \leq \epsilon$ and $\sigma_{\min}^2(Z_1) \geq \sigma_{\min}^2(Z) - \epsilon$. □