How Much Restricted Isometry is Needed In Nonconvex Matrix Recovery?

Richard Y. Zhang Cédric Josz Somayeh Sojoudi Javad Lavaei

May 20, 2018

Abstract

When the linear measurements of an instance of low-rank matrix recovery satisfy a restricted isometry property (RIP)—i.e. they are approximately norm-preserving—the problem is known to contain *no spurious local minima*, so exact recovery is guaranteed. In this paper, we show that moderate RIP is not enough to eliminate spurious local minima, so existing results can only hold for near-perfect RIP. In fact, counterexamples are ubiquitous: we prove that every x is the spurious local minimum of a rank-1 instance of matrix recovery that satisfies RIP. One specific counterexample has RIP constant $\delta = 1/2$, but causes randomly initialized stochastic gradient descent (SGD) to fail 12% of the time. SGD is frequently able to avoid and escape spurious local minima, but this empirical result shows that it can occasionally be defeated by their existence. Hence, while exact recovery guarantees will likely require a proof of *no spurious local minima*, arguments based solely on norm preservation will only be applicable to a narrow set of nearly-isotropic instances.

1 Introduction

Recently, several important nonconvex problems in machine learning have been shown to contain *no spurious local minima* [17, 4, 19, 8, 18, 29, 25]. These problems are easily solved using local search algorithms despite their nonconvexity, because every local minimum is also a global minimum, and every saddle-point has sufficiently negative curvature to allow escape. Formally, the usual first- and second-order necessary conditions for local optimality (i.e. zero gradient and a positive semidefinite Hessian) are also *sufficient* for global optimality; satisfying them to ϵ-accuracy will yield a point within an ϵ-neighborhood of a globally optimal solution.

Many of the best-understood nonconvex problems with no spurious local minima are variants of the *low-rank matrix recovery* problem. The simplest version (known as *matrix sensing*) seeks to recover an $n \times n$ positive semidefinite matrix Z of low rank $r \ll n$, given measurement matrices $A_1, \dots, A_m$ and noiseless data $b_i = \langle A_i, Z \rangle$. The usual, nonconvex approach is to solve the following

$$\underset{x \in \mathbb{R}^{n \times r}}{\text{minimize}} \ \|\mathcal{A}(xx^T) - b\|^2 \quad \text{where} \quad \mathcal{A}(X) = \begin{bmatrix} \langle A_1, X \rangle & \cdots & \langle A_m, X \rangle \end{bmatrix}^T \tag{1}$$

to second-order optimality, using a local search algorithm like (stochastic) gradient descent [17, 21] and trust region Newton's method [15, 7], starting from a random initial point. Exact recovery of the ground truth Z is guaranteed under the assumption that $\mathcal{A}$ satisfies the *restricted isometry property* [13, 12, 26, 10].

Definition 1 (Restricted Isometry Property)**.** The linear map $\mathcal{A} : \mathbb{R}^{n \times n} \to \mathbb{R}^m$ is said to satisfy (r, δ_r)-RIP with constant $0 \leq \delta_r < 1$ if there exists $p > 0$ such that for all rank-r matrices X:

$$(1 - \delta_r)\|X\|_F^2 \leq \frac{1}{p}\|\mathcal{A}(X)\|^2 \leq (1 + \delta_r)\|X\|_F^2. \tag{2}$$

Most existing proofs of "no spurious local minima" are based on an argument of norm preservation: if $\mathcal{A}$ satisfies $(2r, \delta_{2r})$-RIP, then the least-squares residual $\mathcal{A}(xx^T) - b$ can be viewed a *dimension-reduced embedding* of the displacement vector $xx^T - Z$, as in

$$\|\mathcal{A}(xx^T) - b\|^2 = \|\mathcal{A}(xx^T - Z)\|^2 \approx \|xx^T - Z\|_F^2 \text{ up to scaling.} \tag{3}$$

If the high-dimensional problem of minimizing $\|xx^T - Z\|_F^2$ over x contains no spurious local minima (this is easily verified [19]), then its dimension-reduced embedding (1) should satisfy a similar statement.

Theorem 2 (Exact recovery [18, Theorem 8])**.** *Let $\mathcal{A}$ satisfy $(2r, \delta_{2r})$-RIP with $\delta_{2r} < 1/5$. Then, (1) has no spurious local minima: every local minimum x satisfies $xx^T = Z$, and every saddle point has an escape (the Hessian has a negative eigenvalue). Hence, any algorithm that converges to a local minimum is guaranteed to recover Z exactly.*

Exact recovery guarantees based on the norm preservation argument described above were first established for matrix sensing by [4] using RIP, and for matrix completion (i.e. matrix sensing with sparse measurements) by [19] using concetration inequalities and an incoherence assumption. The above bound was actually taken from [18], which had slightly refined the constants from the original result [4]. The same proof technique has been generalized to the noisy and approximate cases [4, 19, 18], and also to tensors [17], and nonsymmetric matrices [18], all resulting in similar guarantees. Note that some of these problems require additional assumptions for norm preservation to hold at their stationary points; these are typically enforced using a regularizer [19, 18].

1.1 How much restricted isometry?

The guarantee in Theorem 2 is conservative; nonconvex matrix completion frequently achieves good performance in practice on datasets that do not satisfy its assumptions [6, 5, 27, 1]. Indeed, assuming only that $\delta_{2r} < 1$, solving (1) to global optimality is enough to guarantee exact recovery [26, Theorem 3.2]. In turn, global optimality is often attained by stochastic optimization algorithms like stochastic gradient descent (SGD). This disconnect between theory and practice motivates the following question.

Can Theorem 2 be substantially improved—is it possible to guarantee the inexistence of spurious local minima with $(2r, \delta_{2r})$-RIP and any choice of $\delta_{2r} < 1$?

At a basic level, the question gauges the generality and usefulness of RIP as a base assumption for proving the inexistence of spurious local minima. RIP with any $\delta_{2r} < 1$ is reasonably general on its own, and most measurement ensembles—even correlated and "bad" measurement ensembles—will eventually come to satisfy the condition as the number of measurements m grows large, certainly once $m = \Theta(n^2)$. On the other hand, RIP with $\delta_{2r} < 1/5$ is far more restrictive, usually requiring the use of nearly-isotropic ensembles like the Gaussian and the sparse binary; see [26, 10].

At a higher level, the question also gauges the wisdom of making exact recovery guarantees through "no spurious local minima" results. For example, in cases where spurious local minima actually exist, exact recovery may hinge on SGD's ability to avoid and escape spurious local minima. Indeed, there is growing empirical evidence that SGD outmaneuvers the "optimization landscape" of nonconvex functions [6, 5, 22, 27, 1], and is able to achieve some global properties [20, 32, 31]. It is unclear whether the success of SGD for matrix recovery should be attributed to the inexistence of spurious local minima, or to some global property of SGD.

1.2 Our results

In this paper, we give a strong negative answer to the question above. Consider the counterexample below, which satisfies (r, δ_{2r})-RIP with $\delta_{2r} = 1/2$, but nevertheless contains a spurious local minimum that causes SGD to fail in 12% of trials.

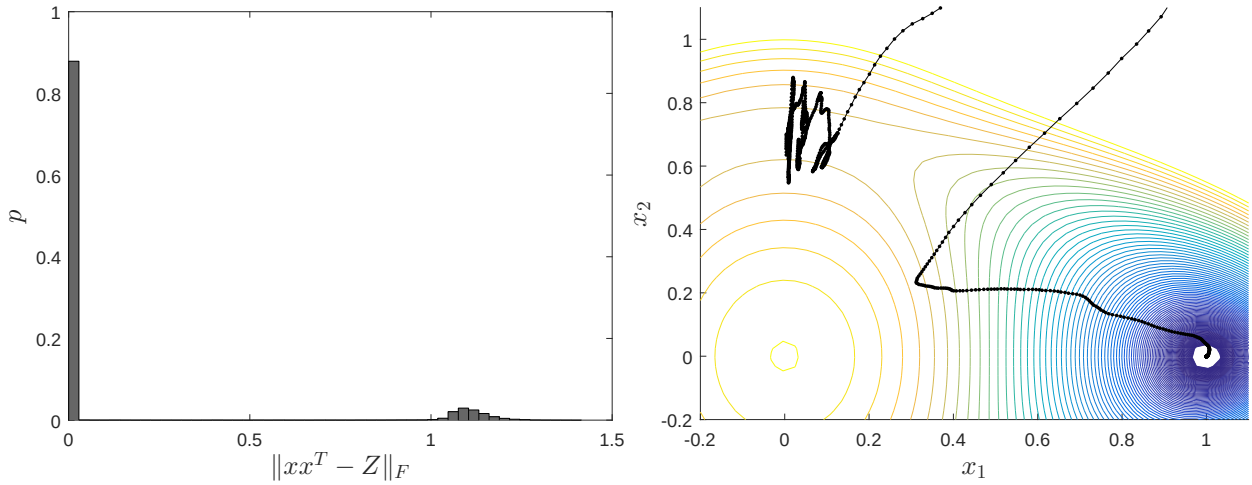

Figure 1: Solving Example 3 using stochastic gradient descent randomly initialized with the standard Gaussian. **(Left)** Histogram over 100,000 trials of final error $\|xx^T - Z\|_F$ after 10^3 steps with learning rate $\alpha = 10^{-3}$ and momentum $\beta = 0.9$. **(Right)** Two typical stochastic gradient descent trajectories, showing convergence to the spurious local minimum at $(0, 1/\sqrt{2})$, and to the ground truth at $(1, 0)$.

Example 3. Consider the following $(2, 1/2)$-RIP instance of (1) with matrices

$$Z = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1/\sqrt{2} \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & \sqrt{3/2} \\ \sqrt{3/2} & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & 0 \\ 0 & \sqrt{3/2} \end{bmatrix}.$$

Note that the associated operator $\mathcal{A}$ is is bijective and satisfies $\|X\|_F^2 \leq \|\mathcal{A}(X)\|^2 \leq 3\|X\|_F^2$ for all X. Nevertheless, the point $x = (0, 1/\sqrt{2})$ satisfies second-order optimality,

$$f(x) \equiv \|\mathcal{A}(xx^T - Z)\|^2 = \frac{3}{2}, \qquad \nabla f(x) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \nabla^2 f(x) = \begin{bmatrix} 0 & 0 \\ 0 & 8 \end{bmatrix},$$

and randomly initialized SGD can indeed become stranded around this point, as shown in Figure 1. Repeating these trials 100,000 times yields 87,947 successful trials, for a failure rate of $12.1 \pm 0.3\%$ to three standard deviations.

Accordingly, Theorem 2 is sharp up to a small factor, and no substantial improvements will be possible. Also, the associated empirical result shows that SGD may fail to avoid and escape spurious local minima. Hence, exact recovery guarantees based on randomly initialized SGD will continue to require a proof of "no spurious local minima".

In fact, there exists an infinite number of counterexamples like Example 3. In Section 3, we prove that, in the rank-1 case, almost every choice of x, z can be used to generate an instance of (1) with a spurious local minimum.

Theorem 4 (Informal)**.** *Let $x, z \in \mathbb{R}^n$ be nonzero and not colinear. Then, there exists an instance of (1) satisfying (n, δ_n)-RIP with $\delta_n < 1$ that has $Z = zz^T$ as the ground truth and x as a spurious local minimum. Moreover, δ_n is explicitly given as*

$$\delta_n \leq \sqrt{1 - \frac{\sin^4 \phi}{1 + \rho^4}}, \qquad \textit{where } \rho = \frac{\|x\|}{\|z\|}, \quad \phi = \arccos\left(\frac{x^T z}{\|x\|\|z\|}\right)$$

if $\rho \geq |\sin \phi|$ holds. Equality is attained if $\phi = \pm\pi/2$.

Hence, attempting to prove "no spurious local minima" using a norm preserving argument [4, 19, 18] will limit us to very small distortion factors δ, which are attainable only by a narrow set of nearly-isotropic measurement ensembles. By constrast, moderate values of δ are insufficiently powerful to prevent spurious local minima from existing.

How do spurious local minima affect the practical performance of SGD? As part of proof for Theorem 4, we formulate a convex optimization problem in Section 2 that takes any arbitrary $x \in \mathbb{R}^{n\times r}$ and rank-r matrix $Z \succeq 0$, and generates an instance of (1) satisfying RIP with Z as ground truth and x as a spurious local minima. Numerically solving this problem, we obtain instances of (1) that can be used to test the behavior of SGD in the presence of spurious local minima.

In Section 5, we apply SGD to two instances of (1) designed to contain spurious local minima. The first, "bad" instance is ill-conditioned, with multiple nonisolated global minima, and a large RIP constant $\delta = 0.975$. Nevertheless, randomly initialized SGD recovers the ground truth with a 100% success rate, as if the problem contained no spurious local minima. The second, "good" instance is well-conditioned, rank-1, and has a moderate RIP constant $\delta = 1/2$ (it is a high-dimensional version of Example 3). However, SGD fails in 59 of 1,000 trials, for a failure rate of $5.90 \pm 2.24\%$ to three standard deviations. Examining the failure cases, we observe that SGD indeed becomes trapped around a spurious local minimum, similar to Figure 1 in Example 3.

1.3 Related work

Special properties of SGD. Practitioners have long known that stochastic gradient descent (SGD) enjoys properties inherently suitable for the sort of nonconvex optimization problems that appear in machine learning [22, 6]. SGD is known to enjoy strong generalization properties to unseen data [20, 32, 31], but its specific behavior is yet not well understood. Our empirical findings in Section 5 partially confirms a common suspicion that SGD is able to avoid and escape spurious local minima even when they do exist. However, as predicted by theory [17], we also find that SGD can become stuck at a local minimum.

Spurious local minima in nonconvex optimization. Recent global optimality guarantees for nonconvex optimization commonly prove a result on "no spurious local minima" using some notion of norm-preservation closely related to RIP [17, 4, 19, 18, 25]. Our results lend support to this approach: if spurious local minima exist, then SGD-based recovery can fail. At the same time, we find the approach to be very pessimistic, because moderate RIP constants δ from typical measurement ensembles are not powerful enough to prevent spurious local minima from existing.

Comparison with convex recovery. Classical theory for the low-rank matrix recovery problem is based on convex relaxation: replacing xx^T in (1) by a convex term $X \succeq 0$, and augmenting the objective with a trace penalty $\lambda \cdot \mathrm{tr}(X)$ to induce a low-rank solution [11, 26, 14, 10]. Convex recovery is usually much more expensive than the nonconvex approach, because it requires optimizing over an $n \times n$ matrix variable instead of an $n \times r$ vector-like variable. However, the convex approach enjoys stronger statistical guarantees. Assuming RIP, the Matrix Danzig Selector and the Matrix Lasso require $\delta_{4r} < \sqrt{2} - 1 \approx 0.4142$ and $\delta_{4r} < (3\sqrt{2} - 1)/17 \approx 0.1907$ to work under noise [10]. Nonconvex noisy recovery require much more conservative values [4, 18], and our findings suggest that there is little margin for improvement. Moreover, the convex approach is statistically consistent [3]. In the noiseless case, exact guarantee is assured once $m \geq \frac{1}{2}n(n+1)$, because $\mathcal{A}$ becomes bijective. By comparison, our results show that the nonconvex approach can consistently fail even with $m \geq \frac{1}{2}n(n+1)$ noiseless measurements.

Notation

We use x to refer to any candidate point, and $Z = zz^T$ to refer to a rank-r factorization of the ground truth Z. For clarity, we use lower-case x, z even when these are $n \times r$ matrices. The sets $\mathbb{R}^{n\times n} \supset \mathbb{S}^n$

are the space of $n \times n$ real matrices and real symmetric matrices, and $\langle X, Y \rangle \equiv \mathrm{tr}(X^T Y)$ and $\|X\|_F^2 \equiv \langle X, X \rangle$ are the Frobenius inner product and norm. We write $X \succeq 0$ (resp. $X \succ 0$) if X is positive semidefinite (resp. positive definite). Given a matrix M, its spectral norm is $\|M\|$, and its eigenvalues are $\lambda_1(M), \ldots, \lambda_n(M)$. If $M = M^T$, then $\lambda_1(M) \geq \cdots \geq \lambda_n(M)$ and $\lambda_{\max}(M) \equiv \lambda_1(M)$, $\lambda_{\min}(M) \equiv \lambda_n(M)$. If M is invertible, then its condition number is $\mathrm{cond}(M) = \|M\|\|M^{-1}\|$; if not, then $\mathrm{cond}(M) = \infty$. The vectorization operator $\mathrm{vec} : \mathbb{R}^{n \times n} \to \mathbb{R}^{n^2}$ preserves inner products $\langle X, Y \rangle = \mathrm{vec}\,(X)^T \mathrm{vec}\,(Y)$ and Euclidean norms $\|X\|_F = \|\mathrm{vec}\,(X)\|$. We overload the same operator $\mathrm{vec} : \mathbb{S}^n \to \mathbb{R}^{n(n+1)/2}$ to map an $n \times n$ symmetric matrix to its $\frac{1}{2}n(n+1)$ degrees of freedom; the exact case should be clear from context. In each case, the matricization operator $\mathrm{mat}(\cdot)$ is the inverse of $\mathrm{vec}\,(\cdot)$.

2 Key idea: Spurious local minima via convex optimization

Given arbitrary $x \in \mathbb{R}^{n \times r}$ and rank-r positive semidefinite matrix $Z \in \mathbb{S}^n$, consider the problem of finding an instance of (1) with Z as the ground truth and x as a spurious local minimum. While not entirely obvious, this problem is actually convex, because the first- and second-order optimality conditions associated with (1) are *linear matrix inequality* (LMI) constraints [9] with respect to the *kernel* operator $\mathcal{H} \equiv \mathcal{A}^T \mathcal{A}$. The problem of finding an instance of (1) that also satisfies RIP is indeed nonconvex. However, we can use the *condition number* of $\mathcal{H}$ as a convex surrogate for the RIP constant δ of $\mathcal{A}$: if the former is finite, then the latter is guaranteed to be less than 1. The resulting LMI optimization can be numerically solved using an interior-point method, like those implemented in SeDuMi [28], SDPT3 [30], and MOSEK [2], to high accuracy.

We begin by fixing some definitions. Given a choice of $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ and the ground truth $Z = zz^T$, we define the nonconvex objective

$$f : \mathbb{R}^{n \times r} \to \mathbb{R} \qquad \text{such that} \qquad f(x) = \|\mathcal{A}(xx^T - zz^T)\|_F^2. \tag{4}$$

Taking partial derivatives yields the first and second-order necessary optimality conditions

$$\langle \nabla f(x), u \rangle = 2\langle \mathcal{A}(xx^T - zz^T), \mathcal{A}(xu^T + ux^T) \rangle = 0 \qquad \forall u \in \mathbb{R}^{n \times r}, \tag{5}$$

$$\langle \nabla^2 f(x) u, u \rangle = 2\langle \mathcal{A}(xx^T - zz^T), uu^T \rangle + \|\mathcal{A}(xu^T + ux^T)\|_F^2 \geq 0 \qquad \forall u \in \mathbb{R}^{n \times r}. \tag{6}$$

For convenience, we refer to a point x satisfying (5) and (6) as a *local minimum*, but note that such a point may still be a saddle point in general (certifying local optimally is NP-hard in general [23]). Any globally optimal choice of x satisfying $xx^T = zz^T$ is easily verified to be a local minimum. We call any x satisfying (5) and (6) while $xx^T \neq zz^T$ a *spurious* local minimum.

Given arbitrary choices of $x, z \in \mathbb{R}^{n \times r}$, let us formulate the problem of picking $\mathcal{A}$ satisfying (5) and (6) as an LMI feasibility. First, we define $\mathbf{A} = [\mathrm{vec}\,(A_1), \ldots, \mathrm{vec}\,(A_m)]^T$ as the matrix representation of the operator $\mathcal{A}$, i.e. the matrix satisfying $\mathbf{A} \cdot \mathrm{vec}\,(X) = \mathcal{A}(X)$ for all X. Then, (5) and (6) may be rewritten as $\mathscr{L}(\mathbf{A}^T \mathbf{A}) = 0$ and $\mathscr{M}(\mathbf{A}^T \mathbf{A}) \succeq 0$, where the linear operators $\mathscr{L}$ and $\mathscr{M}$ are defined

$$\mathscr{L} : \mathbb{S}^{\frac{1}{2}n(n+1)} \to \mathbb{R}^{n \times r} \qquad \text{such that} \qquad \mathscr{L}(\mathbf{H}) \equiv 2\mathbf{X}^T \mathbf{H} e, \tag{7}$$

$$\mathscr{M} : \mathbb{S}^{\frac{1}{2}n(n+1)} \to \mathbb{S}^{nr \times nr} \qquad \text{such that} \qquad \mathscr{M}(\mathbf{H}) \equiv 2\mathrm{mat}(\mathbf{H}e)^T + \mathbf{X}^T \mathbf{H} \mathbf{X}, \tag{8}$$

with respect to the error vector $e = \mathrm{vec}\,(xx^T - zz^T)$ and the $\frac{1}{2}n(n+1) \times nr$ matrix $\mathbf{X}$ that implements the symmetric product operator $\mathbf{X} \cdot \mathrm{vec}\,(u) = \mathrm{vec}\,(xu^T + ux^T)$. To compute a choice of $\mathbf{A}$ satisfying $\mathscr{L}(\mathbf{A}^T \mathbf{A}) = 0$ and $\mathscr{M}(\mathbf{A}^T \mathbf{A}) \succeq 0$, we solve the following LMI feasibility problem

$$\underset{\mathbf{H}}{\text{maximize}} \quad 0 \qquad \text{subject to} \qquad \mathscr{L}(\mathbf{H}) = 0, \quad \mathscr{M}(\mathbf{H}) \succeq 0, \quad \mathbf{H} \succeq 0, \tag{9}$$

and factor a feasible $\mathbf{H}$ back into $\mathbf{A}^T\mathbf{A}$, e.g. using Cholesky factorization or an eigendecomposition. Once a matrix representation $\mathbf{A}$ is found, we may recover the matrices $A_1, \dots, A_m$ implementing the operator $\mathcal{A}$ by matricizing each row of $\mathbf{A}$.

Now, the problem of picking $\mathcal{A}$ with the smallest condition number may be formulated as the following LMI optimization

$$\underset{\mathbf{H},\eta}{\text{maximize}} \quad \eta \qquad \text{subject to} \qquad \eta I \preceq \mathbf{H} \preceq I, \quad \mathscr{L}(\mathbf{H}) = 0, \quad \mathscr{M}(\mathbf{H}) \succeq 0, \quad \mathbf{H} \succeq 0, \tag{10}$$

with solution $\mathbf{H}^\star, \eta^\star$. Then, $1/\eta^\star$ is the best condition number achievable, and any $\mathcal{A}$ recovered from $\mathbf{H}^\star$ will satisfy

$$\left(1 - \frac{1-\eta^\star}{1+\eta^\star}\right)\|X\|^2 \leq \frac{2}{1+\eta^\star}\|\mathcal{A}(X)\|_F^2 \leq \left(1 + \frac{1-\eta^\star}{1+\eta^\star}\right)\|X\|^2$$

for all X, that is, with *any rank*. As such, $\mathcal{A}$ is (n, δ_n)-RIP with $\delta_n = (1-\eta^\star)/(1+\eta^\star)$, and hence also (p, δ_p)-RIP with $\delta_p \leq \delta_n$ for all $p \in \{1, \dots, n\}$; see e.g. [26, 10]. If the optimal value $\eta^\star$ is strictly positive, then the recovered $\mathcal{A}$ yields an RIP instance of (1) with zz^T as the ground truth and x as a spurious local minimum, as desired.

It is worth emphasizing that a small condition number—a large $\eta^\star$ in (10)—will always yield a small RIP constant δ_n, which then bounds all other RIP constants via $\delta_n \geq \delta_p$ for all $p \in \{1, \dots, n\}$. However, the converse direction is far less useful, as the value of $\delta_n = 1$ does not preclude δ_p with $p < n$ from being small.

3 Closed-form solutions

It turns out that the LMI problem (10) in the rank-1 case is sufficiently simple that it can be solved in closed-form. (All proofs are given in the Appendix.) Let $x, z \in \mathbb{R}^n$ be arbitrary nonzero vectors, and define

$$\rho \equiv \frac{\|x\|}{\|z\|}, \qquad\qquad \phi \equiv \arccos\left(\frac{x^T z}{\|x\|\|z\|}\right), \tag{11}$$

as their associated length ratio and incidence angle. We begin by examining the prevalence of spurious critical points.

Theorem 5 (First-order optimality)**.** *The best-conditioned $\mathbf{H}^\star \succeq 0$ such that $\mathscr{L}(\mathbf{H}^\star) = 0$ satisfies*

$$\operatorname{cond}(\mathbf{H}^\star) \leq \frac{1+\delta}{1-\delta} \qquad \textit{where} \qquad \delta = \sqrt{1 - \frac{\sin^4\phi}{1+\rho^4}}, \tag{12}$$

with equality if $\phi = \pm\pi/2$. Hence, if $\phi \neq 0$, then x is a first-order critical point for an $(2r, \delta_{2r})$ instance of (1) with $\delta_{2r} = \delta < 1$ given in (12).

We see that first-order critical points—local extrema and saddle points—may exist for *all* RIP instances of (1), even those with very small RIP constants δ_{2r}. In the practical context of nonconvex matrix recovery, this result highlights the importance of solving (1) using a local search algorithm that can efficiently escape saddle points, such as noisy (stochastic) gradient descent [21], trust-region Newton's method [15, 7], and Nesterov's cubic regularized Newton [24]. Even though Theorem 2 guarantees no spurious local minima if $\delta_{2r} < 1/5$, a poor algorithm may nevertheless lead to failure; line-search Newton's method can converge to a saddle point, while basic gradient descent can require exponential time to escape one [16].

Next, we examine the prevalence of spurious second-order critical points. Note that practical algorithms can only find first- and second-order critical points in polynomial time, so it is standard to view spurious second-order critical points as spurious local minima.

Theorem 6 (Second-order optimality)**.** *The best-conditioned* $\mathbf{H}^\star \succeq 0$ *such that* $\mathscr{L}(\mathbf{H}^\star) = 0$ *and* $\mathscr{M}(\mathbf{H}^\star) \succeq 0$ *satisfies*

$$\operatorname{cond}(\mathbf{H}^\star) \leq \left(\frac{1+\delta}{1-\delta}\right)\left(1 + \frac{(1+\rho^4)^{3/2}}{\rho^2 \sin^4\phi}\right).$$

Hence, if $\phi \neq 0$*, then* x *is a second-order critical point for an* $(2r, \delta_{2r})$ *instance of (1) with* $\delta_{2r} < 1$*.*

Therefore, spurious local minima are ubiquitous, and almost every x is the spurious local minimum of an instance satisfying RIP. However, the associated RIP constants δ_{2r} cannot be too small, because spurious local minima must cease to exist once $\delta_{2r} < 1/5$ according to Theorem 2. Our final result refines this bound, and gives the key estimate.

Theorem 7. *If* $|\sin\phi| \leq \rho$*, then the choice of* $\mathbf{H}$ *in Theorem 5 also satisfies* $\mathscr{M}(\mathbf{H}) \succeq 0$*. Hence, if* $0 < |\sin\phi| \leq \rho$*, then* x *is a second-order critical point for an* $(2r, \delta_{2r})$ *instance of (1) with* $\delta_{2r} = \delta < 1$ *given in (12).*

According to the theorem, every x satisfying $x^T z = 0$ and $\|x\| = \|z\|$ is the spurious local minimum of an instance of (1) satisfying RIP with constant $\delta = 1/\sqrt{2} \approx 0.7$ and condition number $(\sqrt{2}-1)/(\sqrt{2}+1) \approx 5.8284$. Example 3 shows that the smallest δ achievable is actually $1/2$, with condition number 3. However, this requires $|\sin\phi| > \rho$, and is obtained by solving (10) numerically, without insights or guarantees.

4 Experiment 1: Minimum δ over all x, z

What is the most isotropic instance of (1) that still contains a spurious local minimum? In Section 2, we gave a convex formulation that generates an $(2r, \delta)$-RIP instance, given an arbitrary x as a spurious local minimum and an arbitrary $Z \succeq 0$ as the ground truth. The smallest RIP constant $\delta^\star \leq \delta$ (corresponding to the most isotropic instance) depends on the choice of x and Z: our theorems in Section 3 guarantee $\delta^\star \leq 1/\sqrt{2}$, while Example 3 is a constructive proof that $\delta^\star \leq 1/2$. On the other hand, we must have $\delta^\star \geq 1/5$, because no spurious local minimum can exist once $\delta < 1/5$.

The bound $\delta^\star \leq 1/2$ may be tight. To find a smaller value of δ, we use Monte Carlo over all choices of x, z. Fixing n and r, we randomly select $x, z \in \mathbb{R}^{n\times r}$ i.i.d. from the standard Gaussian, solve (10) using MOSEK [2], and record the resulting value of δ. This trial is repeated for 3 hours for every fixed $n \in \{1, 2, \ldots, 10\}$ and $r \in \{1, 2\}$. In every case, we find $\delta \geq 1/2$. This suggests that the bound $\delta^\star \leq 1/2$ may be tight, or that it is the best achievable via our convex formulation (10).

Value of $\delta = 1/2$ is attained in rank-1 case whenever $x^T z = 0$, and $\|x\| = \|z\|/\sqrt{2}$. By randomly selecting $x, z \in \mathbb{R}^n$ i.i.d from the standard Gaussian, projecting $x \leftarrow x - x^T z/\|z\|^2$ and scaling $x \leftarrow \|z\|x/(\sqrt{2}\|x\|)$, we consistently generate instances that attain $\delta = 1/2$ in over 1,000 trials.

5 Experiment 2: SGD escapes spurious local minima

How is the performance of SGD affected by the the presence of spurious local minima? Given that spurious local minima cease to exist with $\delta < 1/5$, we might conjecture that the performance of SGD is a decreasing function of δ. Indeed, this conjecture is generally supported by evidence from the nearly-isotropic measurement ensembles [6, 5, 27, 1], all of which show improving performance with increasing number of measurements m.

This section empirically measures SGD (with momentum, fixed learning rates, and batchsizes of one) on two instances of (1) with different values of δ, both engineered to contain spurious local minima by numerically solving (10). We consider a "bad" instance, with $\delta = 0.975$ and rank $r = 2$, and a "good"

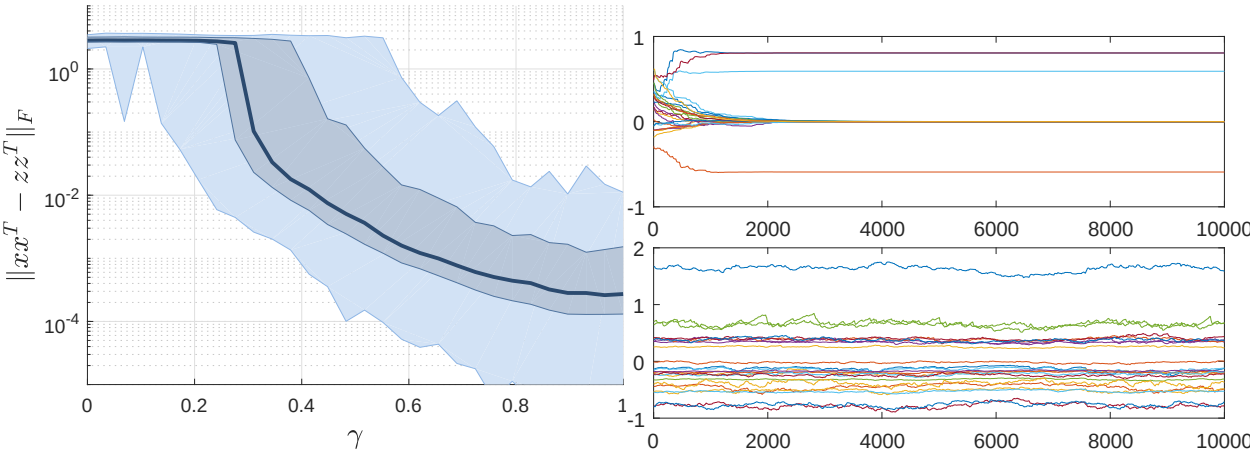

Figure 2: "Bad" instance ($n = 12$, $r = 2$) with RIP constant $\delta = 0.973$ and spurious local min at x_{loc} satisfying $\|xx^T\|_F/\|zz^T\|_F \approx 4$. Here, γ controls initial SGD $x = \gamma w + (1-\gamma)x_{loc}$ where w is random Gaussian. **(Left)** Error distribution after 10,000 SGD steps (rate 10^{-4}, momentum 0.9) over 1,000 trials. Line: median. Inner bands: 5%-95% quantile. Outer bands: min/max. **(Right top)** Random initialization with $\gamma = 1$; **(Right bottom)** Initialization at local min with $\gamma = 0$.

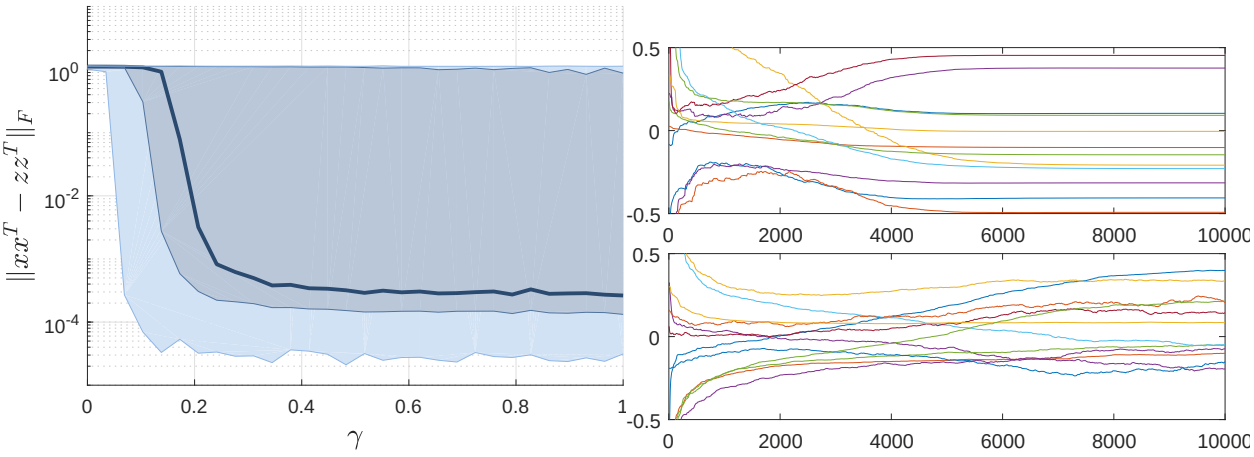

Figure 3: "Good" instance ($n = 12$, $r = 1$) with RIP constant $\delta = 1/2$ and spurious local min at x_{loc} satisfying $\|xx^T\|_F/\|zz^T\|_F = 1/2$ and $x^T z = 0$. Here, γ controls initial SGD $x = \gamma w + (1-\gamma)x_{loc}$ where w is random Gaussian. **(Left)** Error distribution after 10,000 SGD steps (rate 10^{-3}, momentum 0.9) over 1,000 trials. Line: median. Inner bands: 5%-95% quantile. Outer bands: min/max. **(Right top)** Random initialization $\gamma = 1$ with success; **(Right bottom)** Random initialization $\gamma = 1$ with failure.

instance, with $\delta = 1/2$ and rank $r = 1$. The condition number of the "bad" instance is 25 times higher than the "good" instance, so classical theory suggests the former to be a factor of 5-25 times harder to solve than the former. Moreover, the "good" instance is locally strongly convex at its isolated global minima while the "bad" instance is only locally weakly convex, so first-order methods like SGD should locally converge at a linear rate for the former, and sublinearly for the latter.

SGD consistently succeeds on "bad" instance with $\delta = 0.975$ and $r = 2$. We generate the "bad" instance by fixing $n = 12$, $r = 2$, selecting $x, z \in \mathbb{R}^{n\times r}$ i.i.d. from the standard Gaussian, rescale z so that $\|zz^T\|_F = 1$ and rescale x so that $\|xx^T\|_F/\|zz^T\|_F \approx 4$, and solving (10); the results are shown in Figure 2. The results at $\gamma \approx 0$ validate x_{loc} as a true local minimum: if initialized here, then SGD remains stuck here with $> 100\%$ error. The results at $\gamma \approx 1$ shows randomly initialized SGD either escaping our engineered spurious local minimum, or avoiding it altogether. All 1,000 trials at $\gamma = 1$ recover the ground truth to $< 1\%$ accuracy, with 95% quantile at $\approx 0.6\%$.

SGD consistently fails on "good" instance with $\delta = 1/2$ and $r = 1$. We generate the "good" instance with $n = 12$ and $r = 1$ using the procedure in the previous Section; the results are shown in Figure 3. As expected, the results at $\gamma \approx 0$ validate x_{loc} as a true local minimum. However, even with $\gamma = 1$ yielding a random initialization, 59 of the 1,000 trials still result in an error of $> 50\%$, thereby yielding a failure rate of $5.90 \pm 2.24\%$ up to three standard deviations. Examine the failed trials closer, we do indeed find SGD hovering around our engineered spurious local minimum.

Repeating the experiment over other instances of (1) obtained by solving (10) with randomly selected x, z, we generally obtain graphs that look like Figure 2. In other words, SGD usually escapes spurious local minima even when they are engineered to exist. These observations continue to hold true with even massive condition numbers on the order of 10^4, with corresponding RIP constant $\delta = 1 - 10^{-4}$. On the other hand, we do occasionally sample well-conditioned instances that behave closer to the "good" instance describe above, causing SGD to consistently fail.

6 Conclusions

The nonconvex formulation of low-rank matrix recovery is highly effective, despite the apparent risk of getting stuck at a spurious local minimum. Recent results have shown that if the linear measurements of the low-rank matrix satisfy a restricted isometry property (RIP), then the problem contains *no spurious local minima*, so exact recovery is guaranteed. Most of these existing results are based on an argument of norm preservation: relating $\|\mathcal{A}(xx^T - Z)\| \approx \|xx^T - Z\|_F$ and arguing that a lack of spurious local minima in the latter implies a similar statement in the former.

Our key message in this paper is that moderate RIP is not enough to eliminate spurious local minima. To prove this, we formulate a convex optimization problem in Section 2 that generates counterexamples that satisfy RIP but contain spurious local minima. Solving this convex formulation in closed-form in Section 3 shows that counterexamples are ubiquitous: almost any rank-1 $Z \succeq 0$ and any $x \in \mathbb{R}^n$ can respectively be the ground truth and spurious local minimum to an instance of matrix recovery satisfying RIP. We gave one specific counterexample with RIP constant $\delta = 1/2$ in the introduction that causes randomly initialized stochastic gradient descent (SGD) to fail 12% of the time.

Moreover, stochastic gradient descent (SGD) is often but not always able to avoid and escape spurious local minima. In Section 5, randomly initialized SGD solved one example with a 100% success rate over 1,000 trials, despite the presence of spurious local minima. However, it failed with a consistent rate of $\approx 6\%$ on another other example with an RIP constant of just $1/2$. Hence, as long as spurious local minima exist, we cannot expect to guarantee exact recovery with SGD (without a much deeper understanding of the algorithm).

Overall, exact recovery guarantees will generally require a proof of no spurious local minima. However,

arguments based solely on norm preservation are conservative, because most measurements are not isotropic enough to eliminate spurious local minima.

References

[1] Alekh Agarwal, Olivier Chapelle, Miroslav Dudík, and John Langford. A reliable effective terascale linear learning system. *The Journal of Machine Learning Research*, 15(1):1111–1133, 2014.

[2] Erling D Andersen and Knud D Andersen. The MOSEK interior point optimizer for linear programming: an implementation of the homogeneous algorithm. In *High performance optimization*, pages 197–232. Springer, 2000.

[3] Francis R Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9(Jun):1019–1048, 2008.

[4] 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.

[5] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[6] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008.

[7] Nicolas Boumal, P-A Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, page drx080, 2018.

[8] 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.

[9] Stephen Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear matrix inequalities in system and control theory*, volume 15. Siam, 1994.

[10] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[11] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[12] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[13] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[14] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[15] Coralia Cartis, Nicholas IM Gould, and Ph L Toint. Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, 28(1):93–108, 2012.

[16] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1067–1077, 2017.

[17] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points–online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[18] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242, 2017.

[19] 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.

[20] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

[21] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732, 2017.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[23] Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

[24] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[25] Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74, 2017.

[26] 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.

[27] Benjamin Recht and Christopher Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[28] Jos F Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optim. Method. Softw.*, 11(1-4):625–653, 1999.

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

[30] Kim-Chuan Toh, Michael J Todd, and Reha H Tütüncü. Sdpt3–a matlab software package for semidefinite programming, version 1.3. *Optimization methods and software*, 11(1-4):545–581, 1999.

[31] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4151–4161, 2017.

[32] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

A Proof of Main Results

A.1 Proof of Theorem 5

The problem of finding the best-conditioned $\mathbf{H}$ satisfying $\mathscr{L}(\mathbf{H}) = 0$ is the following primal-dual LMI pair

$$\begin{aligned} &\underset{\mathbf{H},\eta}{\text{maximize}}\ \eta && \underset{y,U_1,U_2}{\text{minimize}}\ \mathrm{tr}(U_2) \\ &\text{subject to } \mathscr{L}(\mathbf{H}) = 0, && \text{subject to } \mathscr{L}^T(y) = U_1 - U_2, \\ &\qquad \eta I \preceq \mathbf{H} \preceq I. && \qquad \mathrm{tr}(U_1) = 1, \quad U_1, U_2 \succeq 0, \end{aligned} \tag{13}$$

where $\mathscr{L}^T$ is the adjoint operator to $\mathscr{L}$ in (7). Slater's condition is trivially satisfied by the dual: $y = 0$ and $U_1 = U_2 = \nu^{-1} I$ with $\nu = \frac{1}{2}n(n+1)$ is a strictly feasible point. Hence, strong duality holds, meaning that the two objectives coincide $\mathrm{tr}(U_2^\star) = \eta^\star$ at optimality, so we implicitly solve the primal by solving the dual.

The mechanics of the dual problem become more obvious if we first optimize over U_1 and U_2 and a fixed value of y. If we compute the eigendecomposition $\mathscr{L}^T(y) = V \Lambda V^T$, we find that $U_1 = V(\Lambda)_+ V^T$ takes on the positive eigenvalues of $\mathscr{L}^T(y)$, while $U_2 = V(-\Lambda)_+ V^T$ takes on the negative eigenvalues. The constraint $\mathrm{tr}(U_1) = 1$ rescales y until $\mathrm{tr}(\Lambda)_+ = 1$. This rescales the objective to be $\mathrm{tr}(U_2) = \mathrm{tr}(-\Lambda)_+ / \mathrm{tr}(\Lambda)_+$, and we have

$$\underset{y}{\text{minimize}}\ \frac{\sum_{i=1}^n (-\lambda_i(\mathscr{L}^T(y))_+}{\sum_{i=1}^n (\lambda_i(\mathscr{L}^T(y))_+} \qquad \text{where} \qquad (\alpha)_+ \equiv \arg\min_{\beta \geq 0} |\alpha - \beta| = \begin{cases} \alpha & \alpha \geq 0 \\ 0 & \alpha < 0 \end{cases}. \tag{14}$$

The goal of this latter problem is to find a vector y that maximizes the absolute sum of the positive eigenavlues of $\mathscr{L}^T(y)$, while minimizing the absolute sum of the negative eigenvalues. In Lemma 8, we prove that $\mathscr{L}^T(y)$ has exactly one positive eigenvalue and one negative eigenvalue, and their values in the rank-1 case are closely related to the angle ϕ between x and z. Substituting this into (14) yields an unconstrained minimization

$$\underset{y}{\text{minimize}}\ \frac{1 - \cos\theta}{1 + \cos} \qquad \text{where} \qquad \cos\theta = \frac{e^T \mathbf{X} y}{\|e\| \|\mathbf{X} y\|}.$$

In turn, plugging the lower-bound on θ from Lemma 9 yields the upper-bound $\cos^2\theta = 1 - \sin^2\theta \leq 1 - \sin^4\phi/(1+\rho^4)$, which we have defined as δ^2 in the statement of Theorem 5. Hence, we have $\eta^\star \geq (1-\delta)/(1+\delta)$ as desired. When $|\sin\phi| = 1$, the lower-bound in Lemma 9 is attained, so $\eta^\star = (1-\delta)/(1+\delta)$ as desired.

A.2 Proof of Theorem 6

We show that $\mathbf{H}_\tau \equiv \tau P_{e\perp} + \mathbf{H}_0$ with some $\tau \geq 0$ is a feasible point for (10) with a small condition number. Here, $P_{e\perp} = I - ee^T/\|e\|^2$ is the projection onto the kernel of e, and $\mathbf{H}_0$ is the best-conditioned $\mathbf{H} \succeq 0$ satisfying $\mathscr{L}(\mathbf{H}) = 0$ from Theorem 5, rescaled such that $(1-\delta)I \preceq \mathbf{H}_0 \preceq (1+\delta)I$. Observe that $\mathrm{cond}(\mathbf{H}) \leq (1+\delta+\tau)/(1-\delta)$.

Let us find the smallest $\tau \geq 0$ to guarantee that $\mathscr{L}(\mathbf{H}_\tau) = 0$ and $\mathscr{M}(\mathbf{H}_\tau) \succeq 0$. Note that $\mathscr{L}(\mathbf{H}_\tau) = 0$ is satisfied by construction, because $\mathscr{L}(\mathbf{H}_\tau) = \tau\mathscr{L}(P_{e\perp}) + \mathscr{L}(\mathbf{H}_0)$, and $\mathscr{L}(\mathbf{H}_0) = 0$ by hypothesis while $\mathscr{L}(P_{e\perp}) = 2\mathbf{X}^T(P_{e\perp})e = 0$. Hence, our only difficulty is finding the smallest $\tau \geq 0$ such that

$$\mathscr{M}(\mathbf{H}_\tau) = 2\mathrm{mat}(\mathbf{H}_0 e) + \mathbf{X}^T(\tau P_{e\perp} + \mathbf{H}_0)\mathbf{X} \succeq 0.$$

In Lemma 11, we prove the following two inequalities

$$\|\mathrm{mat}(\mathbf{H}_0 e)\| \leq (1+\delta)\sqrt{1+\rho^4}\|z\|^2, \qquad \lambda_{\min}(\mathbf{X}^T P_{e\perp}\mathbf{X}) \geq \frac{2\|x\|^2 \sin^4\phi}{1+\rho^4}.$$

Hence, $\mathscr{M}(\mathbf{H}_\tau) \succeq 0$ is guaranteed if we set

$$\tau = \frac{(1+\rho^4)^{3/2}(1+\delta)}{\sin^4\phi}\frac{\|z\|^2}{\|x\|^2} \geq \frac{\|\mathrm{mat}(\mathbf{H}_0 e)\|}{\lambda_{\min}(\mathbf{X}^T P_{e\perp}\mathbf{X})}$$

so that $2\mathrm{mat}(\mathbf{H}_0 e) + \tau \mathbf{X}^T P_{e\perp}\mathbf{X} \succeq 0$. Substituting this τ yields the desired condition number bound.

A.3 Proof of Theorem 7

The problem of finding the best-conditioned $\mathbf{H}$ satisfying $\mathscr{L}(\mathbf{H}) = 0$ and $\mathscr{M}(\mathbf{H}) \succeq 0$ is the following primal-dual LMI pair

$$\begin{array}{ll} \text{maximize } \eta & \text{minimize } \mathrm{tr}(U_2) \\ \text{subject to } \mathscr{L}(\mathbf{H}) = 0, & \text{subject to } \mathscr{L}^T(y) - \mathscr{M}^T(U_3) \\ \quad \mathscr{M}(\mathbf{H}) \succeq 0. & \quad = U_1 - U_2, \\ \quad \eta I \preceq \mathbf{H} \preceq I. & \quad \mathrm{tr}(U_1) = 1, \quad U_1, U_2, U_3 \succeq 0, \end{array} \tag{15}$$

where $\mathscr{M}^T$ is the adjoint operator to $\mathscr{M}$ in (8). Slater's condition is trivially satisfied by picking any $y = 0$ and $U_3 = \epsilon I$, and then splitting $\epsilon \mathscr{M}^T(I)$ over U_1 and U_2, so strong duality holds, and we implicitly solve the primal by solving the dual.

Repeating the proof of Theorem 5 yields the following reformulation of the dual problem

$$\begin{aligned} &\text{minimize } \mathrm{tr}\mathbf{X}U_3\mathbf{X}^T + \|e\|\|\mathbf{X}y - u_3\|(1-\cos\theta) \\ &\text{subject to } \|e\|\|\mathbf{X}y - u_3\|(1+\cos\theta) = 1, \\ &\qquad \cos\theta = \frac{e^T(\mathbf{X}y - u_3)}{\|e\|\|\mathbf{X}y - u_3\|}. \end{aligned}$$

Observe that for $U_3 = 0$, the problem reduces to (13). If we can prove that $U_3 = 0$ is optimal for (15), then the same Theorem 5 that we had proved for (13) will also hold for (15).

To proceed, we make a polar decomposition $y = \alpha\hat{y}$ and $U_3 = \alpha\hat{U}_3$ where $\|\mathbf{X}\hat{y} - \hat{u}_3\| = 1$, and substitute the lower-bound $\mathrm{tr}\mathbf{X}\hat{U}_3\mathbf{X}^T \geq 2\|x\|^2\mathrm{tr}(\hat{U}_3)$ to obtain the following problem, written only in terms of the directions

$$\begin{aligned} &\text{minimize } \frac{2\|x\|^2\mathrm{tr}(\hat{U}_3)}{\|e\|(1+\cos\theta)} + \frac{1-\cos\theta}{1+\cos\theta} \\ &\text{subject to } \|\mathbf{X}\hat{y} - \hat{u}_3\| = 1, \\ &\qquad \cos\theta = \frac{e^T(\mathbf{X}\hat{y} - \hat{u}_3)}{\|e\|\|\mathbf{X}\hat{y} - \hat{u}_3\|}. \end{aligned}$$

We parameterize this problem in terms of a fixed $\beta = \mathrm{tr}(\hat{U}_3)$, and apply Lemma 10 to optimize over the exact values of $\hat{y}, \hat{u}_3$. This yields an unconstrained problem with respect to $\beta \geq 0$

$$\text{minimize } \frac{2\|x\|^2/\|e\|\beta + 1 - g(\beta)}{1+g(\beta)},$$

where $g(\beta)$ is defined in Lemma 10. If we can show that the gradient of this function is positive at $\beta = 0$, then the optimal value is $\beta^\star = 0$, and this proves $U_3 = 0$. Lemma 10 says that $g(0) = \cos\theta$ and $g'(0) =$

$\sin\theta$, so we require

$$\left(2\frac{\|x\|^2}{\|e\|} - \sin\theta\right)(1+\cos\theta) \geq (1-\cos\theta)\sin\theta$$
$$\iff \qquad \frac{2\|x\|^2}{\|e\|\sin\theta} \geq 1 + \frac{1-\cos\theta}{1+\cos\theta}$$
$$\iff \qquad \frac{\|z\|^2}{\|x\|^2}\sin^2\phi = \frac{\|e\|\sin\theta}{\|x\|^2} \leq 1+\cos\theta$$

Hence, if $|\sin\phi| \leq \rho$, then the condition is guaranteed, $U_3 = 0$ is optimal, and the proof is complete.

B Proof of technical lemmas

Recall that we have defined

$$\mathscr{L} : \mathbb{S}^{n(n+1)/2} \to \mathbb{R}^{n\times r} \qquad \mathscr{L}(\mathbf{H}) = 2\mathbf{X}^T\mathbf{H}e,$$
$$\mathscr{L}^T : \mathbb{R}^{n\times r} \to \mathbb{S}^{n(n+1)/2} \qquad \mathscr{L}^T(y) = ey^T\mathbf{X}^T + \mathbf{X}ye^T$$

and also

$$\mathscr{M} : \mathbb{S}^{n(n+1)/2} \to \mathbb{S}^{nr} \qquad \mathscr{M}(\mathbf{H}) = 2\mathrm{mat}(\mathbf{H}e) + \mathbf{X}^T\mathbf{H}\mathbf{X},$$
$$\mathscr{M}^T : \mathbb{S}^{nr} \to \mathbb{S}^{n(n+1)/2} \qquad \mathscr{M}^T(U) = \mathrm{vec}\,(U)\mathbf{X}^T + \mathbf{X}\mathrm{vec}\,(U)^T + 2\mathbf{X}U\mathbf{X}^T.$$

Moreover, we use $\rho = \|x\|/\|z\|$ and $\phi = \arccos(x^Tz/\|x\|\|z\|)$. The matrix $\mathscr{L}^T(y)$ is rank-2 with the following eigenvalues.

Lemma 8. *The matrix $\mathscr{L}^T(y)$ is rank-2, and its two nonzero eigenvalues are*

$$\|\mathbf{X}y\|\|e\|(\cos\theta \pm 1), \quad \textit{where } \cos\theta = \frac{e^T\mathbf{X}y}{\|e\|\|\mathbf{X}y\|}. \tag{16}$$

Proof. We project $\mathbf{X}y$ onto e and define w as the residual, as in $\mathbf{X}y = \alpha e + w$ with $\alpha = (e^T\mathbf{X}y)/\|e\|^2$. Then we have the similarity relation

$$\mathscr{L}^T(y) = \begin{bmatrix} e & w \end{bmatrix}\begin{bmatrix} 2\alpha & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} e & w \end{bmatrix}^T \sim \|e\| \cdot \begin{bmatrix} 2\alpha\|e\| & \|w\| \\ \|w\| & 0 \end{bmatrix},$$

and the 2×2 matrix has eigenvalues $\|\alpha e\|^2 \pm \sqrt{\|\alpha e\|^2 + \|w\|^2}$. Substituting $\|\mathbf{X}y\|^2 = \|\alpha e\|^2 + \|w\|^2$ completes the proof. □

Also, the angle between e and $\mathrm{range}(\mathbf{X})$ is closely associated with the angle between x and z.

Lemma 9. *The angle θ in (16) is related to the angle ϕ in (11) by the following (with equality if $\phi = \pm\pi/2$)*

$$\sin\theta = \frac{(\|z\|\sin\phi)^2}{\|e\|} \geq \frac{\sin^2\phi}{\sqrt{1+\rho^4}}.$$

Proof. We project z onto $\mathrm{range}(x)$ and define w as the residual, as in $z = x\alpha + w$ where $\alpha = (x^T z)/\|x\|^2$. Then, we have the similarity relation

$$xx^T - zz^T = \begin{bmatrix} x & w \end{bmatrix} \begin{bmatrix} (1-\alpha^2)I_r & -\alpha I_r \\ -\alpha I_r & -I_r \end{bmatrix} \begin{bmatrix} x & w \end{bmatrix}^T \sim \begin{bmatrix} (1-\alpha^2)\|x\|^2 & -\alpha\|x\|\|w\| \\ -\alpha\|x\|\|w\| & -\|w\|^2 \end{bmatrix},$$

and may solve the problem of projecting e onto $\mathrm{range}(\mathbf{X})$ after a change of basis

$$\begin{aligned} \|e\| \sin\theta =& \min_y \|\mathbf{X}y - e\| \\ =& \min_y \|xy^T + yx^T - (xx^T - zz^T)\|_F, \\ =& \min_{\tilde{y}_1, \tilde{y}_2} \left\| \begin{bmatrix} \tilde{y}_1 & \tilde{y}_2 \\ \tilde{y}_2 & 0 \end{bmatrix} - \begin{bmatrix} (1-\alpha^2)\|x\|^2 & -\alpha\|x\|\|w\| \\ -\alpha\|x\|\|w\| & -\|w\|^2 \end{bmatrix} \right\|_F, \\ =& \|w\|^2 = \|z\|^2 \sin^2\phi. \end{aligned}$$

This proves the equality. On the other hand, we have

$$\|e\| = \|xx^T - zz\|_F = \sqrt{\|x\|^4 + \|z\|^4 - 2(x^T z)^2} \leq \sqrt{\|x\|^4 + \|z\|^4}.$$

Substituting yields the desired lower-bound. □

Lemma 10. *Define the following*

$$g(\beta) = \min_{U \succeq 0, y} \left\{ \frac{e^T(\mathbf{X}y - \mathrm{vec}\,(U))}{\|e\|\|\mathbf{X}y - \mathrm{vec}\,(U)\|} : \frac{\mathrm{tr}U}{\|\mathbf{X}y - \mathrm{vec}\,(U)\|} = \beta \right\}.$$

Then we have

$$g(\beta) = \sqrt{1-\beta^2}\sin\theta + \beta\cos\theta,$$

where $\cos(\theta) = g(0)$.

Proof. From the proof of Lemma 9 we have

$$\begin{aligned} xx^T - zz^T \sim& \begin{bmatrix} (1-\alpha^2)\|x\|^2 & -\alpha\|x\|\|w\| \\ -\alpha\|x\|\|w\| & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & -\|w\|^2 \end{bmatrix} \\ =& V_1\|e\|\sin\theta + V_2\|e\|\cos\theta, \end{aligned}$$

where $\|V_1\|_F = \|V_2\|_F = 1$ and $\langle V_1, V_2 \rangle = 0$, and $\cos\theta$ is from Lemma 9. Suppose at optimality $y^\star, U^\star$, we have $\mathrm{mat}(\mathbf{X}y^\star) = \alpha V_1$ and $U^\star = \beta V_2$. Then

$$\begin{aligned} \frac{e^T(\mathbf{X}y^\star - \mathrm{vec}\,(U^\star))}{\|e\|\|\mathbf{X}y^\star - \mathrm{vec}\,(U^\star)\|} =& \frac{\langle V_1\sin\theta + V_2\cos\theta,\ V_1\alpha + V_2\beta\rangle}{\sqrt{\alpha^2+\beta^2}} \\ =& \sqrt{1-\beta^2}\sin\theta + \beta\cos\theta \end{aligned}$$

as desired. To prove the hypothesis, we just need to show that $\mathrm{mat}(\mathbf{X}y^\star)$ and $U^\star$ are orthogonal. Note that the same $y^\star, U^\star$ also solve the following (up to scaling)

$$\begin{aligned} & \min_y \|\mathbf{X}y - u - e\| \\ =& \min_y \|xy^T + yx^T - U - (xx^T - zz^T)\|_F, \\ =& \min \left\| \begin{bmatrix} \tilde{y}_1 & \tilde{y}_2 \\ \tilde{y}_2 & 0 \end{bmatrix} - \begin{bmatrix} \tilde{u}_1 & \tilde{u}_2 \\ \tilde{u}_2 & \tilde{u}_3 \end{bmatrix} - \begin{bmatrix} (1-\alpha^2)\|x\|^2 & -\alpha\|x\|\|w\| \\ -\alpha\|x\|\|w\| & -\|w\|^2 \end{bmatrix} \right\|_F. \end{aligned}$$

For every fixed $\tilde{u}$, we will simply get $\tilde{y}_1$ and $\tilde{y}_2$ canceling with $\tilde{u}_1$ and $\tilde{u}_2$. Hence, the effect of U with a fixed trace is maximized by setting $\tilde{u}_1 = \tilde{u}_2 = 0$ and $\tilde{u}_3 > 0$, and $y^\star, U^\star$ are indeed orthogonal. □

Lemma 11. *Let $\hat{\mathbf{H}}$ be the optimal choice in Theorem 5. Then*

$$\|\mathrm{mat}(\hat{\mathbf{H}}e)\| \le (1+\delta)\|e\|, \qquad \lambda_{\min}(\mathbf{X}^T P_{e\perp}\mathbf{X}) \ge \frac{2\|x\|^2 \sin^4\phi}{1+(\|x\|/\|z\|)^4}.$$

Proof. For the first bound, we have

$$v^T \mathcal{D}_{\hat{\mathbf{H}}} v = \langle \hat{\mathbf{H}}(xx^T - zz^T), vv^T \rangle \le \underbrace{\|\hat{H}\|}_{(1+\delta)} \underbrace{\|xx^T - zz^T\|_F}_{\|e\|} \|vv^T\|_F.$$

For the second bound, let θ be the angle between e and $\mathrm{range}(\mathbf{X})$

$$\begin{aligned} v^T(\mathbf{X}^T P_{e\perp}\mathbf{X})v &= \|P_{e\perp}\mathbf{X}v\|^2 \\ &\ge \|\mathbf{X}v\|^2 \sin^2\theta \\ &= \|xv^T + vx^T\|_F^2 \cdot \sin^2\theta \\ &\ge \left(2\|x\|^2 \sin^2\theta\right) \cdot \|v\|^2. \end{aligned}$$

Substituting the lower-bound on θ from Lemma 9 completes the proof. □